UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

DOCTORADO EN ESTADÍSTICA APLICADA

TESIS

ANÁLISIS PREDICTIVO DE LA DESERCIÓN ESTUDIANTIL EN LOS ALUMNOS DE LA UNIVERSIDAD NACIONAL DE MOQUEGUA ENTRE 2009 Y 2019, USANDO MACHINE LEARNING.

PRESENTADA POR:

JOSÉ ORLANDO QUINTANA QUISPE

PARA OPTAR EL GRADO ACADÉMICO DE:

DOCTOR EN ESTADÍSTICA APLICADA

PUNO, PERÚ

2023

UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

DOCTORADO EN ESTADÍSTICA APLICADA

TESIS

ANÁLISIS PREDICTIVO DE LA DESERCIÓN ESTUDIANTIL EN LOS ALUMNOS DE LA UNIVERSIDAD NACIONAL DE MOQUEGUA ENTRE 2009 Y 2019, USANDO MACHINE LEARNING.

PRESENTADA POR:

JOSE ORLANDO QUINTANA QUISPE

PARA OPTAR EL GRADO ACADÉMICO DE:

DOCTOR EN ESTADÍSTICA APLICADA

APROBADA POR EL JURADO SIGUIENTE:

PRESIDENTE

Firmado digitalmente por HUATA PANCA Percy FAU 20145496170 soft
Motivo: Soy el autor del documento
Fecha: 31.03.2023 07:06:35 -05:00

.....
D.Sc. PERCY HUATA PANCA

PRIMER MIEMBRO

Firmado digitalmente por SALAS PILCO Maria Maura FAU 20145496170 soft
Motivo: Soy el autor del documento
Fecha: 31.03.2023 15:51:35 -05:00

.....
Dra. MARIA MAURA SALAS PILCO

SEGUNDO MIEMBRO

Firmado digitalmente por PEREZ QUISPE Samuel Donato FAU 20145496170 soft
Motivo: Soy el autor del documento
Fecha: 31.03.2023 13:59:47 -05:00

.....
Dr. SAMUEL DONATO PEREZ QUISPE

ASESOR DE TESIS

Firmado digitalmente por CARPIO VARGAS EDGAR ELOY
Motivo: Soy el autor del documento
Fecha: 31.03.2023 06:20:40 -05:00

.....
Dr. EDGAR ELOY CARPIO VARGAS

Puno, 12 de enero de 2023

ÁREA: Estadística.
TEMA: Análisis predictivo de la deserción estudiantil en los alumnos de la Universidad Nacional de Moquegua entre 2009 y 2019, usando Machine Learning.
LÍNEA: Técnicas multivariadas supervisadas.

DEDICATORIA

El presente trabajo lo dedico a Dios por la vida, salud, la inspiración y la humildad.

A mis profesores quienes en el tiempo que estuvimos en interacción trasmitieron sus experiencias y conocimiento.

A la memoria de mi padre César quien siempre inculco el respeto y la paciencia para lograr nuestras metas

A mi madre autora este ser imperfecto y lleno de debilidades las cuales el tiempo ha ido puliendo sus imperfecciones.

A mi esposa Ana y a mis hijos Máximo y Diana los motores y motivos de mi vida.

AGRADECIMIENTOS

Agradecer a la Universidad Nacional del Altiplano y sus docentes de la escuela de posgrado quienes permitieron, compartieron y trasmitieron las competencias necesarias para lograr este Doctorado.

Agradecer a la Universidad Nacional de Moquegua mi centro laboral por permitirme usar los datos de las oficinas del Servicio de Asistencia Social en la persona de la Mg. Marlene Cajaña Quispe quien proporciono la información y los datos y el tiempo necesario para recopilarla y poder culminar satisfactoriamente este trabajo doctoral.

A mis jurados de tesis, Dr. Percy Huata Panca, Dra. María Maura Salas Pilco y Dr. Samuel Donato Pérez Quispe.

También, agradecer, de manera muy especial, al Dr. Edgar Eloy Carpio por su experiencia, guía y consejos los cuales fueron precisos y puntuales en el desarrollo de este trabajo.

ÍNDICE GENERAL

CAPÍTULO III

MATERIALES Y MÉTODOS

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

El trabajo analizó y validó el mejor modelo machine learning para predecir la deserción en alumnos de la universidad nacional de Moquegua, entre 2009 y 2019 sede Mariscal Nieto, además se determinaron las características influyentes en la deserción. Los modelos: Regresión logística, Árboles de decisión, Máquinas de vector soporte y Naive Bayes, junto a la metodología CRISP-DM y las métricas como matriz de confusión y validación cruzada K-Folds, fueron usados, y la selección de características importantes se hizo con dos técnicas Random Forest y Featurewiz; se usa el software Python y sus librerías; el tipo de investigación es descriptivo correlacional y el diseño es observacional con obtención de datos de fuente secundaria. La muestra se seleccionó por muestreo probabilístico aleatorio estratificado, resultando la muestra de alumnos que abandonaron la universidad sin concluir los estudios universitarios en 109 y no desertores en 220 en total 329 datos; luego, la validación cruzada indica que: Árbol de decisión logra 76% de éxito seguido de Regresión logística 73%, Máquina vector soporte 71% y Naive Bayes 62%. Las características que influyen en la deserción involucran los datos generales como: ciclo, edad, dirección zonal; en el aspecto económico: ingreso total, carga familiar, hijos en estudios superiores, componentes del hogar, vivienda del estudiante y sostiene el hogar; en el aspecto de vivienda fueron: tipo de construcción, número de dormitorios.

Palabras clave:

Deserción Universitaria, Máquinas de aprendizaje, Estudiante universitario, Modelos predictivos.

ABSTRACT

The research analyzed and validated the best machine learning model to predict dropout in students at the National University of Moquegua, between 2009 to 2019 in Mariscal Nieto campus, in addition to determining the influential characteristics in dropout. The models: Logistic Regression, Decision Trees, Support Vector Machines and Naive Bayes, together with the CRISP-DM methodology and the metrics as a K-Folds confusion and cross-validation matrix were used, and the selection of important characteristics was applied two techniques Random Forest and Featurewiz; the Python software and its libraries are used; The type of research is correlational descriptive, and the design is observational with obtaining data from secondary sources. The sample was selected by stratified random probability sampling, resulting in the sample of students who left the university without completing university studies in 109 and non-dropouts in 220 in total 329 data; then, cross-validation indicates that: Decision tree achieves 76% success followed by Logistic Regression 73%, Support Vector Machine 71%, and Naive Bayes 62%. The characteristics that influence attrition involve general data such as: cycle, age, zonal direction; in the economic aspect: total income, family burden, children in higher education, household components, student housing and household support; In the housing aspect were construction type, number of bedrooms.

Keywords:

University dropout, Learning machines, University student, Predictive models.

INTRODUCCIÓN

La deserción estudiantil universitaria es un problema latente en nuestro país y esta involucra muchos factores entre los cuales podemos mencionar factores sicológicos, económicos, sociológicos, organizacionales y los que son de interacción entre el estudiante y la institución.

Buscamos establecer el mejor modelo en Machine Learning para predecir la deserción estudiantil de los alumnos de la Universidad Nacional de Moquegua en la sede Mariscal Nieto entre 2009 y 2019. El diseño de investigación es no experimental y de tipo de investigación explicativo correlacional, la población está constituida por 2305 alumnos y la muestra 329 los métodos a emplear son: Regresión Logística (RL), Máquina Vector Soporte (MVS), Arboles de Decisión (DT) y Clasificador Bayesiano (NB); el propósito es obtener el mejor modelo para establecer las variables que más influyen en la deserción estudiantil.

Nuestra investigación está centrada en el área de la Estadística Aplicada, línea de investigación descriptiva cuasi-experimental y experimental y los temas modelos predictivos uni y multivariante, en concordancia a los objetivos de la Escuela de Posgrado de la UNA-Puno. Método de investigación no experimental, estudio transversal y correlacional.

El Primer Capítulo está referido a la revisión del marco teórico, relacionado con la metodología CRISP-DM, deserción universitaria, Machine Learning y métodos de clasificación y antecedentes. En el Segundo Capítulo trata sobre el planteamiento del problema, identificación del problema, enunciado, justificación, objetivos generales y especifico y las hipótesis general y específica. En el Tercer Capítulo se describe los materiales y métodos, el lugar de estudio, población, muestra y métodos de acuerdo con los objetivos de investigación. En el Cuarto Capítulo, se exponen los resultados de la investigación, y se hace la comparación con otras investigaciones, así como la comprobación de la hipótesis, conclusiones y recomendaciones.

CAPÍTULO I

REVISIÓN DE LITERATURA

1.1 Marco teórico

Existe una relación que beneficia a las estadísticas aplicadas y el aprendizaje automático que está siendo reconocida. Esta interfaz es cada vez más famosa debido a que existen áreas donde estas se superponen. Las estadísticas son de amplio valor en la investigación de ML. (Ferrero, 2021)

CRISP DM son las siglas de Cross-Industry Standard Process for Data Mining, 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.(1994., 2021)

Crear un sistema de aprendizaje automático implica más que simplemente seleccionar un modelo, entrenarlo y aplicarlo a nuevos datos. Existen marcos que nos ayudan a organizar proyectos de aprendizaje automático.

Uno de esos marcos es CRISP-DM, el proceso estándar entre industrias para la minería de datos. Se inventó hace bastante tiempo, en 1996, pero a pesar de su antigüedad, sigue siendo aplicable a los problemas actuales.

Según CRISP-DM, el proceso de aprendizaje automático tiene seis pasos:

1) Comprensión empresarial
2) Comprensión de datos
3) Preparación de datos
4) Modelado
5) Evaluación

6) Despliegue

Cada fase cubre tareas típicas:

En el paso de comprensión empresarial, intentamos identificar el problema, comprender cómo podemos resolverlo y decidir si el aprendizaje automático será una herramienta útil para resolverlo.

En el paso de preparación de datos, transformamos los datos en forma tabular que podemos usar como entrada para un modelo de aprendizaje automático.

Cuando los datos están preparados, pasamos al paso de modelado, en el que entrenamos un modelo.

Después de que se identifica el mejor modelo, está el paso de evaluación, donde evaluamos el modelo para ver si resuelve el problema comercial original y medir su éxito al hacerlo.

Finalmente, en el paso de implementación, implementamos el modelo en el entorno de producción. Esta metodología permite de manera cíclica perfeccionar el modelo matemático.

1.1.1 Deserción Universitaria

De acuerdo a Tinto (1989) citado de en Iván y Echeverry (2017) Hablar sobre deserción es muy complicado, existen muchos enfoques y una variedad de formas de abandono. No hay una definición que capte de manera amplia este complicado fenómeno universitario. Es necesario que los que investiguen sobre el tema deban elegir aquellas definiciones que coincidan a sus intereses y objetivos. Se debe tener presente que uno de los fines de la Universidad es la educación de las personas. Analizar la retención divorciada de sus causas y consecuencias educativas no es de interés del individuo ni de las instituciones. (Tinto 1982)

Consideramos la siguiente definición: “La deserción se puede definir como el proceso de abandono, voluntario o forzoso de la carrera en la que se matricula un estudiante, por la influencia positiva o negativa de circunstancias internas o externas a él o ella.” (Fiegehen y Díaz 2016)

Según los conceptos anteriores la deserción tiene una amplia variedad de causas

todas ellas desde el punto de vista del que investiga y que se determinan 112 factores usados para predecir la deserción estudiantil universitaria, los cuales se identifican a partir de las cinco dimensiones (personales, académicos, económicos, sociales e institucionales).(Alban Taipe 2019)

Uno se pregunta, si al estudiar los desertores se debe incluir los no desertores, pero se recomienda no incluirlos debido a que podrían generar un ruido y provocaría que el algoritmo predictivo no sea eficiente.(Solis et al., 2018b)

En un interesante estudio de las causas desde el punto de vista del actor, es decir del alumno, se incorporan nuevos factores y que se logra construir desde el análisis de 65 teorías organizacionales, 12 teorías educativas y el raciocinio lógico. Como resultado del estudio se obtuvo 11 factores: Conocimiento limitado sobre el uso de software especializado de la carrera, Embarazo planificado/no planificado, Compromiso del docente con el estudiante, Compromiso económico del hijo primogénito con la familia, Bullying, Machismo/feminismo, Vicios adquiridos por el estudiante, Número de hijos del estudiante, Adaptación de estudiante a las metodologías de formación universitaria, Ranking de la Universidad o Carrera, Perspectivas del estudiante sobre la inserción en el campo laboral. (Alban Taipe 2019)

Entendemos que a medida que se investiga sobre el tema se va comprendiendo mejor el problema y encontrando nuevas causas o factores por lo que es un problema de mucho interés y varía según la región y la interacción en el ámbito o localidad.

1.1.2 Tipos de deserción

Se identifican:

- Deserción definitiva: cuando un estudiante no retoma su formación académica.
- Deserción por factores: depende de la causal que ocasiona la separación del estudiante del sistema de educación superior.
- Deserción por cambio de facultad
- Deserción por cambio de programa académico

Tomando como referencia el trabajo de (Alban Taipe 2019)

1.1.3 Machine Learning

El Machine learning (ML) se está expandiendo, aplicándose a diversas áreas apoyando en la toma de decisiones especialmente en el tratamiento de grandes volúmenes de datos y la toma de decisiones por medio de modelos predictivos y de clasificación mediante algoritmos los cuales se pueden mejorar y estimar la calidad del mismo y que se están usando en gran variedad de aplicaciones de internet y tecnologías de los teléfonos inteligentes que en estos últimos años han desarrollado de manera vertiginosa, sin embargo, ¿cuáles son las métricas más adecuadas para evaluar la calidad de una explicación? Esto se responde por medio de una revisión del estado actual sobre la interpretación del aprendizaje automático enfocándose en el impacto social y en los métodos y métricas desarrollados.(Carvalho, et. al., 2019)

Machine learning es un campo que se centra en el aspecto de aprendizaje de la Inteligencia Artificial (AI) mediante el desarrollo de algoritmos que representan mejor un conjunto de Datos supervisados. A diferencia de la programación clásica, en la que un algoritmo se puede codificar explícitamente utilizando características conocidas, machine learning usa subconjuntos de datos para generar un algoritmo que puede usar combinaciones nuevas o diferentes de características y pesos que se pueden derivar de los primeros principios. En ML, hay cuatro métodos de aprendizaje de uso común, cada uno útil para resolver diferentes tareas: aprendizaje supervisado, no supervisado, semisupervisado y por refuerzo. (Choi, et. al. , 2020)

El aprendizaje supervisado necesita datos etiquetados para la formación. Un ejemplo de entrenamiento se denomina punto de datos o instancia y consta de un par de entrada y salida (x, y). y es la verdad de salida o de tierra para la entrada x. A diferencia del aprendizaje supervisado, en el aprendizaje no supervisado, los datos solo proporcionan entradas.(Wlodarczak, 2019)

La deserción está relacionada con el estudiante universitario, directamente y en asociación con la universidad. Proponemos usar los modelos de clasificación para determinar patrones y luego predecir la posibilidad de la deserción. Comparar la eficiencia de los modelos y usarlos para clasificar a los estudiantes. (Zárate-Valderrama et al., 2021)

1.1.3.1 Métodos de Clasificación

¿Qué es la clasificación? La clasificación consiste básicamente en reconocer los atributos del elemento a clasificar utilizando el conocimiento adquirido durante la etapa de entrenamiento para poder asignarle un valor a la variable objetivo de dicho elemento.(Soto, 2015)

De acuerdo a Zuluaga (2011) citado en Iván y Echeverry (2017) La clasificación es una metodología importante en ML y lo que hace es proporcionar un modelo matemático que sea capaz de asignar la pertenencia de un objeto a una clase que tiene una característica común a todos sus elementos. Al lograr determinar este modelo matemático mediante nuevas observaciones la pertenencia a esa clase puede ser predicha. En estos casos, además de la matriz tradicional de datos, se tiene un vector de clases que asigna la pertenencia a cada observación proporcionada ya sea de un origen geográfico o un atributo cualitativo (bueno/malo), (si/no).

Los sistemas inteligentes tienen la tarea de hacer la clasificación supervisada con mucha frecuencia. Por tanto, la metodología desarrollada por la estadística (Regresión Logística, Análisis Discriminante) o también por la Inteligencia Artificial (Redes Neuronales, Inducción de Reglas, Árboles de Decisión, Redes Bayesianas) permiten llevar a cabo las tareas que comprende la clasificación. Antes de aplicar un método de clasificación se debe tener los datos particionados en dos que serán utilizados con los siguientes fines: entrenamiento (75%) y test(25%) (Parra Rodríguez 2017a)

a) Regresión Logística

El modelo de regresión logística es una técnica de aprendizaje supervisado y establece la relación entre la probabilidad de que ocurra un suceso dado que el individuo presenta los valores $(X = x_1, X = x_2, \dots, X = x_k)$

$$P[Y = 1 \mid x_1, x_2, \dots, x_k] = \frac{1}{e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_k x_k}}$$

El objetivo es hallar los coeficientes $(\beta_1, \beta_2, \dots, \beta_k)$ que mejor se ajusten a la expresión. El procedimiento de estimación de estos coeficientes se basa en el método dé máxima verosimilitud.

La parte de «regresión» en regresión logística podría dar la impresión de que se trata de una técnica de regresión. Sin embargo, recibe este nombre por motivos históricos. Al principio de la neurona, se hace una combinación lineal que se parece mucho a una regresión lineal. Y luego se aplica la función logística. Así que así surgió el nombre, regresión logística.

b) Máquinas Vector Soporte

La Máquina Vector Soporte (MVS) se fundamenta en la teoría de aprendizaje estadístico desarrollada por Vapnik y pertenecen a la categoría de los clasificadores lineales ya que establecen separadores lineales o hiperplanos ya sea en el espacio original de los datos de entrada, o en un espacio transformado.(Carmona 2016)

El SVM se hace inicialmente con los datos de entrenamiento para los cuales se conoce el valor de la variable dependiente, luego de este entrenamiento se define un modelo que permitirá clasificar otro individuo.(Carmona 2016)

Las Máquinas de Vectores de Soporte (Support Vector Machines) permiten encontrar la forma óptima de clasificar entre varias clases. La clasificación óptima se realiza maximizando el margen de separación entre las clases. Los vectores que definen el borde de esta separación son los vectores de soporte. En el caso de que las clases no sean linealmente separables, podemos usar el truco del kernel para añadir una dimensión nueva donde sí lo sean.

c) Árboles de Decisión

Este modelo surge en el campo del (ML) y la (IA) y que, a partir de una base de datos, fabrica diagramas lógicos que solucionan un problema.

Otro nombre por el que lo conoce es segmentación jerárquica.

Se considera que este tipo de clasificación es explicativa y des composicional que usa una división secuencial, iterativa y descendente tomando como punto inicial la variable dependiente, crea grupos homogéneos definidos mediante combinaciones de las variables independientes estas incluyen la totalidad de los casos recogidos en la muestra. (Parra Rodríguez 2017b)

Los árboles de decisión son una técnica de aprendizaje automático muy utilizada. Sus ventajas principales son:

- Son fáciles de entender y explicar a personas que todavía no están familiarizadas con técnicas de Inteligencia Artificial
- Se adaptan a cualquier tipo de datos
- Descubren cuáles son los atributos relevantes

También tienen sus desventajas:

- No extrapolan bien fuera del rango de entrenamiento
- Tienen la tendencia a sobre ajustar, sobre todo si no se regularizan

d) Clasificador Bayesiano

Naive Bayes es uno de los clasificadores más utilizados por su simplicidad y rapidez. Se trata de una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados, en base al Teorema de Bayes, también conocido como teorema de la probabilidad condicionada:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

1.1.3.2 Validación de los resultados del modelo

Las métricas de evaluación de los modelos de clasificación en machine learning que se usaran en el proceso de validación son: Matriz de confusión y Validación cruzada K-Folds.

a) Matriz de confusión

Como resultado de aplicar un método de clasificación, se cometerán dos errores, en el caso de una variable binaria que toma valores 0 y 1, habrá ceros que se clasifiquen incorrectamente como unos y unos que se clasifiquen incorrectamente como ceros. A partir de este recuento se puede construir el siguiente cuadro de clasificación:

Tabla 1

Matriz de confusión

Valor real Y_i **Valor estimado $\hat{Y}_i$**	$Y_i = 0$	$\hat{Y}_i = 1$
$\hat{Y}_i = 0$	P_{11}	P_{12}
$\hat{Y}_i = 1$	P_{21}	P_{22}

Fuente:(Parra Rodríguez, 2017)

Donde P11 y P22 corresponderán a predicciones correctas (valores 0 bien predichos en el primer caso y valores 1 bien predichos en el segundo caso), mientras que P12 y P21 corresponderán a predicciones erróneas (valores 1 mal predichos en el primer caso y valores 0 mal predichos en el segundo caso). A partir de estos valores se pueden definir los índices que aparecen en la siguiente tabla 2:

Tabla 2

Métricas Matriz de confusión

Índice	**Definición**	**Expresión**
Tasa de aciertos	Cociente entre las predicciones correctas y el total de predicciones.	$\frac{P_{11}+P_{22}}{P_{11}+P_{12}+P_{21}+P_{22}}$
Tasa de errores	Cociente entre las predicciones incorrectas y el total de predicciones.	$\frac{P_{12}+P_{21}}{P_{11}+P_{12}+P_{21}+P_{22}}$
Especificidad	Proporción entre la frecuencia valores cero correctos y el total de valores cero observados.	$\frac{P_{11}}{P_{11}+P_{21}}$
Sensibilidad	Proporción entre la frecuencia de valores uno correctos y el total de valores uno observados.	$\frac{P_{22}}{P_{12}+P_{22}}$
Tasa de falsos ceros	Proporción entre la frecuencia de valores cero incorrectos y el total de valores cero observados	$\frac{P_{21}}{P_{11}+P_{21}}$
Tasa de falsos unos	Proporción entre la frecuencia de valores uno incorrectos y el total de valores uno observados.	$\frac{P_{12}}{P_{12}+P_{22}}$

Fuente: (Parra Rodríguez, 2017)

Las métricas de precisión y recuperación se utilizan para evaluar el rendimiento de los modelos predictivos. Precisión mide el porcentaje de estudiantes que el modelo predijo correctamente como deserción. Recuperación calcula el porcentaje de estudiantes en riesgo de deserción identificados correctamente a partir del conjunto de datos de la prueba. Estas métricas son más eficaces para evaluar el rendimiento en conjuntos de datos desequilibrados que métricas alternativas como la clasificación precisión y característica operativa del receptor (Davis y Goadrich 2006) citado en (Kiesling, 1971)

Podemos dividir en cuatro pasos la investigación:

Paso1. Construir el conjunto de datos de entrenamiento e ingresar los datos de la predicción.

Paso 2. Usar los datos en el entrenamiento de los modelos.

Paso 3. Usar otra parte de los datos para entrenar los modelos de predicción generados previamente Paso 4. Usar las muestras obtenidas de los modelos de predicción en el conjunto de prueba y evaluar los resultados de la predicción generados.(Tan y Shao, 2015)

Los métodos de ML permiten extraer información clave de los datos. Un requisito previo para obtener conocimiento y descubrimientos novedosos es a partir de los datos de observación y simulados. Existen tres elementos relevantes: transparencia, interpretabilidad y explicabilidad.(Roscher et al., 2020)

Las máquinas de vectores de soporte (SVM) han demostrado ser una buena alternativa en comparación con otras técnicas de aprendizaje automático específicas para problemas de clasificación. Sin embargo, también tienen sus debilidades.(Farquad et al., 2009)

b) Validación Cruzada K-Folds

En la validación cruzada k-fold, dividimos aleatoriamente el conjunto de datos de entrenamiento en k pliegues sin reemplazo. Aquí, k – 1 pliegues, los

llamados pliegues de entrenamiento, se utilizan para el entrenamiento del modelo, y un pliegue, el llamado pliegue de prueba, se utiliza para la evaluación del rendimiento. Este procedimiento se repite k veces para que obtengamos k modelos y estimaciones de rendimiento.

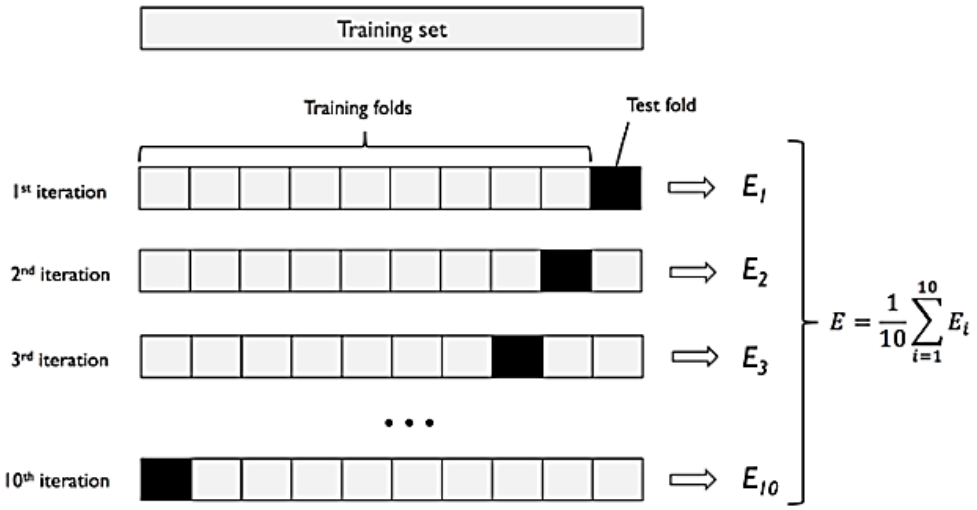


Figura 1. Funcionamiento de la validación cruzada K-Folds.

Nota: Tomado del libro "Machine Learning with Pytorch and Scikit-Learn Develop Machine Learning and Deep Learning Models with Python." De (Raschka et al., 2022)

Luego, calculamos el rendimiento promedio de los modelos en función de los pliegues de prueba diferentes e independientes para obtener una estimación de rendimiento que sea menos sensible a la subpartidación de los datos de entrenamiento en comparación con el método de holdout. Normalmente, utilizamos la validación cruzada k-fold para el ajuste del modelo, es decir, encontrar los valores óptimos de hiperparámetro que producen un rendimiento de generalización satisfactorio, que se estima a partir de la evaluación del rendimiento del modelo en los pliegues de prueba.

Una vez que hemos encontrado valores de hiperparámetro satisfactorios, podemos volver a entrenar el modelo en el conjunto de datos de entrenamiento completo y obtener una estimación final del rendimiento utilizando el conjunto de datos de prueba independiente. La razón detrás de la adaptación de un modelo a todo el conjunto de datos de entrenamiento después de la validación cruzada k-fold es que, en primer lugar, normalmente estamos interesados en un modelo único y final (frente a k modelos

individuales), y en segundo lugar, proporcionar más ejemplos de entrenamiento a un algoritmo de aprendizaje generalmente da como resultado un modelo más preciso y robusto.(Raschka et al., 2022)

c) Curva ROC

La curva ROC (acrónimo de Receiver Operating Characteristic, o Característica Operativa del Receptor) es una medida del rendimiento para problemas de clasificación, indica cuánto es capaz un modelo de distinguir entre diferentes clases. Cuanto mayor sea el AUC, mejor será el modelo en su predicción. Esta curva representa dos parámetros: Tasa de verdaderos positivos (TPR) y Tasa de falsos positivos (FPR).

1.2 Antecedentes

A nivel internacional: En la University of Technology Gauteng, South África, se estudia el crecimiento de la tasa de deserción como un problema común a las instituciones de educación superior. En esta publicación, se usan las técnicas de clasificación para determinar los estudiantes en "riesgo de" abandonar los estudios usando sus calificaciones. Explican los beneficios de poseer esta información a tiempo. El estudio se hace usando las técnicas de machine learning conocidas como Random Forest, Support Vector Machines, Decision Trees, Naïve Bayes, K-Nearest Neighbor y Logistic Regression. Usando los registros y expedientes de 4419 estudiantes. Se concluyó que el método de Random Forest (94,14%) tuvo mejor desempeño para determinar a los estudiantes que podrían abandonar los estudios. (Lottering et al., 2020)

En la universidad de Bologna, Italia se investigó sobre la manera de desarrollar una herramienta que, mediante la explotación de técnicas de aprendizaje automático, permita predecir la deserción de un estudiante de pregrado, se usó como muestra el conjunto de datos seudoanonimizados de 15000 estudiantes matriculados en varios cursos 2016-2017 de lo cual se determinó que el escalamiento de los datos no tiene ningún efecto sobre el rendimiento final de análisis discriminante lineal y Random Forest, por otro lado, el escalado afecta el rendimiento de Máquina vector soporte, pero no parece añadir ningún beneficio. Los resultados obtenidos muestran que partiendo de datos sin ningún valor pedagógico o didáctico se puede intentar mitigar el problema de la deserción.(Del Bonifro et al., 2020a)

Según se investigó en el Departament of Computer Science Tshwane University of Technology Gauteng, South África y cuyo objetivo fue aplicar las técnicas de clasificación para identificar los estudiantes en riesgo de abandono a partir de 4419 registros de estudiantes de una base de datos institucional relacionada con los estudiantes del diplomado matriculado en la Facultad de información, Comunicación y Tecnología, usando como instrumentos los métodos de clasificación machine learning, Python y las librerías de Scikit Learn los resultados revelan que la tasa de precisión general de Random Forest (94,14 %) fue mejor que la de otros algoritmos para identificar a los estudiantes con riesgo de abandono y aunque esta investigación puede predecir estudiantes en riesgo de deserción, la conclusión no se puede generalizar ya que los datos son de una Universidad Tecnológica específica.(Lottering et al., 2020)

En una investigación desarrollada en el Departamento de Tecnologías de la Información y la Comunicación en el Politécnico del noroeste de Malasia que como objetivo fue presentar un modelo de análisis predictivo utilizando métodos de aprendizaje automático supervisado que predice la calificación final (FG) del estudiante en función de su rendimiento académico histórico de los estudios. Bajo una población que utilizó un conjunto de datos recopilados de 489 estudiantes del Departamento de Tecnología de la Información y la Comunicación en el Politécnico del noroeste de Malasia durante los últimos cuatro años, de 2016 a 2019 usando como instrumentos los métodos de clasificación machine learning, métricas de evaluación y software WEKA concluye que J48 fue el mejor modelo de análisis predictivo con una tasa de precisión más alta de 99.6%. así como los modelos de clasificación de machine learning son confiables para la identificación de estudiantes en riesgo de deserción. Los resultados pueden servir como alerta temprana a los administradores y decisiones oportunas para reducir los índices de deserción.(Abdul Bujang et al., 2021)

Un estudio hecho en la universidad Ural State University, Rusia cuyo objetivo fue dar una visión general del desarrollo del Machine Learning hasta la actualidad, diversos algoritmos de aprendizaje automático, aplicaciones y desafíos se obtuvieron como resultados la digitalización y revolución de internet han llevado a un volumen creciente de datos estructurados y no estructurados que deben utilizarse para el análisis y que el aprendizaje automático como impulsor tecnológico clave abarca el poder inteligente para aprovechar el conocimiento de los datos disponibles, también se presentó una revisión exhaustiva del proceso y los algoritmos de aprendizaje automático.(Alzubi et al., 2018)

En un artículo de investigación desarrollado en The Nelson Mandela African Institution of Science and Technology, Aruska, Tanzania que tenía como objetivo proporcionar recomendaciones de algoritmos basados en datos a los investigadores actuales sobre el tema. Usando como muestra datos recopilados sobre el aprendizaje a nivel de país en Twaweza, el conjunto de datos consta de 18 características y aproximadamente 61340 muestras. Aplicando las técnicas de muestreo al conjunto de entrenamiento y realizaron el experimento para construir el modelo usando 60% de datos para entrenamiento y 20% para la validación y otros 20% para la evaluación del modelo. Para medir el rendimiento de los algoritmos su usaron la métrica de evaluación de algoritmos, así como la validación cruzada con k=5. Concluye que los dos clasificadores LR y MLP demostraron ser superiores al lograr las métricas de rendimiento más altas y se demostró que la contribución más alta fue dada por el género de los estudiantes y se determinaron dos clasificadores con el mayor rendimiento predictivo, el ajuste de los hiperparámetros mejora el rendimiento de cada algoritmo en comparación con su configuración de referencia.(Mduma et al., 2019)

Un estudio alemán indica que debe estudiarse la detección temprana de los estudiantes en riesgo de abandonar la universidad como un medio de aprovechamiento eficiente de los fondos públicos y privados puestos que estos son escasos. Se debe prevenir este fenómeno y debe haber intervención de las autoridades mediante tutorías o apoyo institucional. Los estudios de prevención usando las técnicas de machine learning son más eficaces en materia de prevención ya que estas técnicas determinan las características básicas de aquellos estudiantes que pueden desertar.(Isphording y Raabe, 2019)

En un estudio realizado en el Instituto de Tecnología de la Universidad de Hawassa, Etiopía cuyo objetivo fue desarrollar un modelo de predicción para la predicción de la deserción escolar de los estudiantes utilizando técnicas de aprendizaje automático donde los datos históricos se recopilaron de hawassa University Student Information Systems Portal (HUSIS) URL: https://sis.hu.edu.et/ con credenciales relevantes asignadas al registrador de la Escuela de Informática. Los datos contienen información académica de B.Sc. estudiantes del Instituto de Tecnología de la Universidad de Hawassa, Etiopía. Los datos históricos abarcaron desde septiembre de 2017 hasta julio de 2020. El conjunto data set eran los datos de 472 estudiantes matriculados en la Facultad de Estudiantes de Informática Sistema de Información (SIS) para el curso "Avanzado Sistemas de bases de datos". El HUSIS proporciona interacción estudiante-maestro a través de sus portales de

aprendizaje en línea, y está disponible para estudiantes y maestros asignados a una Facultad en particular. Este estudio ha llevado a cabo experimentos para evaluar el rendimiento y la utilidad de diferentes algoritmos de clasificación para predecir el estado académico de los estudiantes. usando Aprendizaje profundo y tensorflow se encontró que este estudio contribuye a abordar los desafíos globales actuales de los estudiantes que abandonan su estudio. El modelo de predicción desarrollado permite a las instituciones de educación superior dirigirse a los estudiantes que probablemente abandonen la escuela e intervenir oportunamente para mejorar las tasas de retención y la calidad de la educación. El modelo de predicción de análisis de aprendizaje permitiría a las instituciones de educación superior dirigirse a los estudiantes que probablemente abandonen la escuela e intervenir oportunamente para mejorar las tasas de retención y la calidad de la educación.(Amare y Simonova, 2021)

A nivel de Latinoamérica y el Caribe se analizaron las causas, implicancias y las formas de mejora se determinó la magnitud de la deserción estudiantil sobre 15 países. Se encontró que la deserción alcanza en promedio el 57,5% y que las universidades privadas tienen mayor incidencia y entre la población varonil, siendo las áreas de humanidades e ingeniería las que tienen mayor frecuencia. Las razones que dominan son las socioeconómicas y culturales como externas, en las internas la carencia de ayudas estudiantiles y apoyo académico, así como académicas y personales. Para reducir la alta incidencia se determina mejorar los diagnósticos, la información y articulación a nivel institucional e identificar de manera temprana los probables desertores y modificar los procesos de enseñanza aprendizaje.(Fiegehen y Díaz, 2016)

Según se establece en un estudio hecho en la Universidad de Cundinamarca, ingeniería de sistemas, se seleccionaron cinco métodos: regresión logística(RL), vecinos más cercanos (KNN), árboles de decisión (DT), arboles aleatorios (random Forest, RF) y máquinas vector soporte (MVS), evaluaron los modelos según las métricas de precisión, matriz de confusión y las métricas adicionales logrando determinar que el modelo con mejor desempeño fue regresión logística (RL) y concluyeron que mientras más atributos se tengan en cuenta, la precisión de los modelos mejora.(Ayala-yaguara y Valenzuela-sabogal, 2020)

Para América Latina, se determina que las tasas de deserción se encuentran entre 40% y 75%. Cuyas razones de ello son múltiples, pero es importante detectar el riesgo que se

relaciona con este nivel de deserción. Usando ML permite tener una ayuda en la toma de decisiones mediante sus técnicas. Se muestra el estudio de los estudiantes de Ingeniería Industrial que tienden a desertar en la Universidad Distrital Francisco de Caldas en el periodo 2003-I a 2018-I. Se selecciona el algoritmo que más se adapta usando la comparación de técnicas de aprendizaje automatizado en Azure Machine Learning.(Zea et al., 2019)

Un estudio hecho en el Tecnológico de Costa Rica analiza el rendimiento de cuatro algoritmos de aprendizaje automático con diferentes perspectivas para definir archivos de datos, en la predicción de la deserción de estudiantes universitarios. Los algoritmos utilizados fueron: Random Forest, Neural Networks, Support Vector Machines y Logistic Regression. Se encontró que el algoritmo Random Forest con 10 variables muestreados aleatoriamente como candidatos en cada división, fue el mejor para predecir las deserciones y que la perspectiva ideal para entrenar el algoritmo es utilizar información sobre todos los semestres que los estudiantes toman dentro de un período de tiempo determinado, utilizando una variable de clasificación que define al no desertor como el estudiante graduado. En una primera muestra de validación, este enfoque predijo correctamente el 91% de los abandonos, con una sensibilidad del 87%.(Solis et al., 2018a)

En la Universidad Distrital Francisco José Caldas, Bogotá Colombia se hizo una investigación con el objetivo de identificar los estudiantes en riesgo de deserción de ingeniería industrial de la universidad distrital José Caldas entre los semestres 2003-1 al 2018-1. Usando como muestra a los alumnos de la carrera de ingeniería industrial entre 2003 y 2018 en un total de 3201 y 24 variables académicas, usando como instrumentos las técnicas de aprendizaje automático de Azure Machine Learning Studio. Los resultados obtenidos son confiables con una precisión de 90.3% y una precisión de 93.6% los resultados de la predicción no especifican la variable académica o social que provoca la deserción y la identificación de herramientas de Machine Learning, como Azure Machine Learning para usar en estos procesos de identificación del riesgo de deserción.(Zea et al., 2019)

En la Facultad de Ciencias de la universidad Técnica Estatal de Quevedo en Ecuador durante el periodo 2012 – 2013 se propone un nuevo clasificador bayesiano simple (SBND) el cual es comparado con otros clasificadores bayesianos y hacer uso de los modelos gráficos probabilísticos en el campo de la enseñanza con el objetivo de predecir

la deserción estudiantil en las universidades. Se usa la información socioeconómica de los alumnos matriculados en el periodo establecido, comparando la información obtenida y sus combinaciones. Se utiliza el software Weka de acceso libre y gratuito.(Oviedo Bayas y Zambrano-Vega, 2019)

En un artículo de revisión sistemática sobre la predicción de la deserción escolar realizada de febrero a marzo de 2020 en la universidad de Sao Paulo, Brasil y cuyo objetivo era hacer una revisión sistemática de la literatura, realizada sobre la aplicación de una técnica de aprendizaje automático para predecir la deserción estudiantil en las instituciones de educación superior. Concluyeron que el alto número de características consideradas en cada estudio es motivo para que los investigadores utilicen técnicas de reducción de la dimensionalidad, los datos completos funcionan mejor que los datos que se redujeron las características. La mayoría de los artículos implementan más de tres técnicas de aprendizaje automático para poder comparar modelos entre sí. Los árboles de decisión son los más populares y son usados en más del 50% de los artículos, los artículos indican que se usó más de una métrica de evaluación, accuracy, precisión, recall y la puntuación F1.(Silva y Roman, 2021)

En la universidad de Cundinamarca, Colombia se investigó con el objetivo de obtener un modelo de minería de datos aplicado al problema de la deserción universitaria en el programa de ingeniería de sistemas cuya población comprende los estudiantes del programa de ingeniería de sistemas y los resultados indican que se evaluaron los resultados de cada modelo en las métricas de precisión, matriz de confusión y métricas adicionales de la matriz de confusión y se ajustaron los parámetros del modelo al graficar su curva de aprendizaje, el modelo de minería de datos se sustenta bajo la técnica de regresión logística debido a que se presentó los mejores índices de precisión en cada una de las métricas de rendimiento planteadas y que entre más atributos se tengan los índices de precisión de los modelos mejoran.(Ayala-yaguara y Valenzuela-sabogal, 2020)

En el artículo de investigación de Caselli Gismondi y Urrelo Huiman, (2021) hecha en la Universidad Nacional del Santa, Huancavelica y que tenía por objetivo proponer las características que deberán formar parte de un modelo de predicción basado en machine learning que contribuya a la adopción de medidas oportunas durante el seguimiento académico. Se usó como población a todos los estudiantes de la UNS, que abarcó a los estudiantes de las cuatro escuelas profesionales de la Facultad de Ingeniería de esta

institución en los semestres de los años entre el 2004 y 2018. Basado en los instrumentos como la revisión documental y métodos estadísticos. Se utilizaron los softwares libres: Python y Pandas. Concluye que el análisis de las variables más frecuentemente empleadas en los estudios de predicción, la data maestra de la Universidad depurada y la propuesta de características que deberán integrar el modelo y los resultados de abandono académico sin haber obtenido Grado académico ni Título profesional que muestra la UNS y la calidad del promedio global de las asignaturas en todas sus carreras ponen luz roja a esta problemática evidenciando la necesidad de su estudio. Las variables de tipo individual, académica, institucional y socioeconómica halladas durante el análisis de la literatura, así como los resultados del procesamiento y depuración de las datas de la UNS constituyeron los referentes para la propuesta de las variables a emplear en el modelo de predicción académica.

Sánchez-Hernández et. al., (2017) Realizaron un estudio de la deserción en los primeros ciclos de la Universidad Privada del Norte, sede Trujillo en los años 2008 a 2012. realizaron una recopilación de información usando encuestas para determinar el índice de deserción semestral (IDS) y los factores que están involucrados, la conclusión es que la deserción no es conveniente y que es vital determinar la deserción universitaria.

Alban Taipe (2019) En este trabajo se plantean nuevas variables que provocan la deserción universitaria y que surgen a partir de la revisión exhaustiva de la literatura y plantea un punto de vista diferente al referirse que hay pocos trabajos que consideren el problema de la deserción a partir del actor, es decir, del alumno. También, propone que en los estudios no deberían limitarse a solo predecir la deserción, sino que también deberían proponer las medidas que permitan evitarla.

Sanchez Nina (2017) Elabora un modelo estadístico predictivo usando estadística multivariada y considerando la regresión logística binaria logrando una efectividad en la predicción del 91.2% empleando para su estudio los factores socioeconómicos, académicos y culturales lo que permitiría determinar si un estudiante abandonará la escuela profesional elegida y así proveer planes y acciones para afrontar este problema de la deserción en las escuelas profesionales de la UNA - PUNO.

Sobre la deserción y los factores asociados en un artículo hecho en la universidad privada del norte con sede en Trujillo (Perú) entre los años 2008- 2012, se recopilo información por medio de una encuesta de servicios y también sobre la dependencia de la matricula

con lo que se determinó el índice de deserción semestral y se valoraron la relación entre los factores de la encuesta de satisfacción resultando que hubo un incremento anual del índice de deserción estudiantil y se concluyó que era necesario determinar otros factores que pronostiquen la retención estudiantil.(Sánchez-Hernández et al., 2017)

En un estudio hecho en la Universidad Continental de Huancayo sobre detección de patrones de éxito en los estudios universitarios, se tuvo como objetivo encontrar patrones en la información académica y sociodemográfica del primer ciclo de estudios. Se determino que los estudiantes que abandonan la universidad no superaron el quinto ciclo, también se determinó que las variables que influyen en la deserción son: con quien vive el estudiante, el estado civil del estudiante, satisfacción del desempeño del docente y estado civil de los padres.(Gamarra et al., 2018)

En un trabajo sobre minería de datos, se analiza el paradigma del aprendizaje para atacar el problema de la deserción estudiantil, determina que los alumnos tienden a desertar cuando su promedio del semestre es muy bajo. Se recopilan las principales variables y se usa un modelo de clasificación, luego, se realiza una exploración de los datos para definir la arquitectura del modelo. Las derivadas del algoritmo se evalúan para observar el aprendizaje de la máquina y comparar los resultados con nuevos datos y así determinar la calidad del modelo al predecir la deserción. (Mamani Padilla, 2019)

Según el artículo de Martínez y Mateus, (2020) realizado en el Politécnico Colombiano Jaime Isaza Cadavid, Facultad de Ingeniería, Medellín. Cuyo objetivo fue proponer un modelo predictivo que sirva como apoyo a las Universidades colombianas para el análisis de la deserción en estudiantes, principalmente, en programas de pregrado en modalidad virtual, hecha tomando eventos históricos con distintas variables de tipo social, académico, personal, laboral, ingresos a las plataformas e-learning, etc. y posteriormente, a estas variables se le aplican algoritmos de aprendizaje profundo usando el aprendizaje profundo, tensorflow logra diseñar la propuesta de un modelo de aprendizaje profundo para la predicción de deserción estudiantil en las instituciones de educación superior en modalidad virtual. El modelo analiza y hace recomendaciones, con el objetivo de apoyar la toma decisiones apropiadas y a tiempo con respecto a la tasa de deserción estudiantil principalmente en el ámbito de la educación virtual en instituciones de educación superior, siendo un proyecto continuo y a largo plazo. Por lo que se pudo realizar un afinamiento del modelo incluyendo atributos adicionales tanto de tipo categórico como

numérico. Y mejorar la arquitectura de la red neuronal, asignando capas adicionales; y con esto lograr resultados de mayor precisión y más cercano a la realidad. Incluso, se puede comparar con un modelo utilizando otra técnica de inteligencia artificial y comparar los resultados y la eficiencia.

En la Facultad de Ingeniería y Arquitectura con los datos de los estudiantes desde el 2009 hasta la actualidad, extraídos de la Dirección General de Tecnología de Información bajo la autorización de secretaria general y cuyo objetivo fue determinar el nivel de eficacia del modelo de aprendizaje supervisado para el pronóstico de la deserción de estudiantes de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión – Lima. usando los datos personales, académicos y financieros de los estudiantes de la FIA del periodo de Enero del 2009 hasta Julio de 2019, siendo un total 3161 registros por medio de los Modelos de machine learning, Validación cruzada, métricas de evaluación de modelos y la curva ROC, software RapidMiner se obtuvo un diccionario de datos de 26 variables finales manteniéndose los factores, los algoritmos de aprendizaje Naive Bayes y DecisiónTree fueron las más viables para todas las carreras de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión – Lima, también con respecto al tercer objetivo, el algoritmo de aprendizaje supervisado KNN es el que se adecúa mejor a las carreras de Ingeniería de Alimentos, Ingeniería Ambiental y Arquitectura. El algoritmo Random Forest se adecúa mejor a la carrera de Ingeniería de Sistemas, y el algoritmo Decision Tree, se adecúa mejor a la carrera de Ingeniería Civil. La técnica de balanceo de datos que usaron los algoritmos KNN y Decision Tree fue el UnderSampling, y el algoritmo Random Forest, SmoteTomek.(Morales, 2020).

Ramírez y Grandón,(2018) En la Universidad de la Costa - CUC. buscaron presentar una clasificación basada en árboles de decisión (CBAD) con parámetros optimizados para predecir la deserción de los estudiantes universitarios. analizando 5288 casos de estudiantes pertenecientes a una universidad pública chilena. Usando como instrumentos los Modelos de machine learning, Validación cruzada, métricas de evaluación de modelos y la curva ROC, software RapidMiner. Se concluye que el resultado de la aplicación de esta técnica con parámetros optimizados logro una razón de precisión de un 87.27%. Se concluye que el uso de técnicas de CBAD con optimización de parámetros resulta en una mejor precisión en comparación a otras investigaciones con un número similar de datos.

En una investigación realizada en Universidad de la Costa - CUC. Cuyo objetivo principal

de este proyecto de investigación fue crear un modelo para la predicción de la deserción de estudiantes de pregrado en la Universidad de la Costa - CUC, a partir del análisis de diferentes factores socioeconómicos y académicos. Durante la fase de caracterización se construyó un conjunto de datos (dataset), a partir de la compilación de los datos demográficos, culturales, sociales, familiares, educativos, estatus socioeconómico y perfil psicológico de cada estudiante, de los periodos comprendidos entre 2013-1 y 2018-2. Tal información fue recopilada a partir de los formatos de inscripción que diligencian los estudiantes cuando ingresan a la institución, un total de 1.606 registros únicos de estudiantes fueron recopilados. Los instrumentos utilizados fueron los Modelos de machine learning, Validación cruzada, métricas de evaluación de modelos y la curva ROC. Se concluyo que el mejor algoritmo para clasificar la deserción estudiantil en la Universidad de la Costa CUC es RandomForest con datos balanceados, la verificación de la efectividad de la técnica se obtuvo luego de realizar un proceso de prueba (test) mediante cross-validation utilizando igual 25 pliegues. La predicción del algoritmo RandomForrest fue la mejor de las alternativas evaluadas (Maquinas de soporte vectorial y Redes Bayesianas). Del conjunto de datos disponible en el proyecto (1606), fue posible obtener los resultados, arrojando inicialmente una exactitud (accuracy) del 78.1% para el clasificador ADTree (Camargo, 2020)

Rivera Vergaray (2021) investiga en la Universidad Nacional Intercultural de la Amazonía. Investigó la predicción de la deserción académica de estudiantes en la Universidad Nacional Intercultural de la Amazonía. Basados en un dataset extraído de la base de datos del sistema de gestión académica de la universidad, que contiene datos socioeconómicos y de rendimiento académico los cuales fueron procesados y formateados utilizando técnicas de one hot encoding para así poder aplicar los modelos predictivos ya mencionados. Para el procesamiento y formateo de datos se utilizó consultas Transac Sql y la aplicación de los modelos predictivos se hizo a través del Software Knime y utilizando Python a través de Google Colab.

Conclusiones vinculadas a los objetivos: Los modelos KNN y Arboles de decisión son los que registran un mejor ajuste y tienen un Accuracy aceptable de 88.840% y 88.4%.

En la Facultad de ingeniería eléctrica y electrónica de la Universidad de los Andes. Se buscó lograr clasificar e intentar predecir el comportamiento de la deserción en un histórico de 10 años (20 semestres evaluados). Usando la base de datos desarrollada

cuenta con descriptores relacionados a 1962 estudiantes. Los instrumentos utilizados: Modelos de machine learning, Validación cruzada, métricas de evaluación de modelos y la curva ROC. Se concluye que los tres métodos de clasificación Random Forest, Multi-Layer Perceptron y Logistic Regression fueron comparados por medio de los 4 indicadores que miden el rendimiento de las predicciones junto con los valores de las matrices de confusión. Se construyó una base de datos con características propias de cada uno de los estudiantes cuya información se tenía registrada. (Molina et al., 2022)

Quiñones Huatangari et al. (2020) investigaron en la Universidad Nacional de Jaén cuyo objetivo del trabajo de investigación fue emplear la minería de datos para determinar modelos que estimen la deserción de estudiantes Awajún y Wampis de la UNJ. La población es igual a la muestra, cuando la muestra coincide con la población se está en presencia de una muestra censal. Ha sido el elemento de registro de información socioeconómica, académica y personal de los cuarenta y nueve (49) estudiantes Awajún y Wampis en los períodos 2012 – 2019 en la UNJ. Usaron la validación cruzada con 10 pliegues, software Weka, métricas de evaluación de modelos y la curva ROC. Concluyeron que el empleo de los algoritmos J48, Ridor y PART de clasificación han permitido obtener tres modelos basados en dos reglas por cada uno con un porcentaje de instancias bien clasificadas de 87.8%, de esta manera siendo los que tienen mejor comportamiento.

En un estudio hecho en la Institución Educativa Departamental General Carlos Albán del Municipio de Alban- Cundinamarca. Que buscaba Evaluar la incidencia del uso de machine learning en la predicción del desempeño en el espacio proyectivo del pensamiento espacial. El estudio fue hecho en una población donde funciona el bachillerato con 299 estudiantes y la parte administrativa; la sede de básica primaria Policarpa Salavarrieta cuenta con 199 estudiantes y la sede del Jardín Infantil que atiende 31 niños. En el sector rural tiene tres sedes: La María con 35 estudiantes; y dos escuelas unitarias Java con 28 estudiantes y Los Alpes con 7 niños. Se emplearon dos instrumentos: la prueba del simulacro y la encuesta sobre factores sociodemográficos. Estos resultados dan cuenta del desempeño académico en el desarrollo del espacio proyectivo del pensamiento espacial y de las características sociodemográficas de cada estudiante. Las conclusiones fueron que los algoritmos, el machine learning, las simulaciones, la extracción, procesamiento y almacenamiento de datos se configura en una metodología necesaria e imprescindible en el desarrollo de estrategias y

conocimientos que promuevan y potencien las actividades en el aula de clase y todos los procesos pedagógicos que esto conlleva. También, La aplicación de técnicas de machine learning y simulación para el desarrollo de un modelo que permita la predicción del desempeño de los estudiantes de Educación Básica y Media se constituyen en una herramienta eficaz para el docente, ya que pueden clasificar a los estudiantes y conocer con alto grado de precisión las categorías de DESEMPEÑO_BAJO y DESEMPEÑO_ALTO de los aprendices; esta es una ventaja que permite a los docentes crear estrategias en cada una de sus asignaturas para orientar al desarrollo del pensamiento espacial a partir de la Teoría del Desarrollo del Conocimiento Espacial de Piaget.(Mendez Aguirre y López Martinez, 2019)

Quintero (2022) en su investigación hecha en la Facultad de Ingeniería de la Universidad de Antioquia y cuyo objetivo fue proponer un modelo que permita predecir la deserción temprana en los programas presenciales de pregrado en la facultad de ingeniería de la U. de A. haciendo uso de los métodos y las técnicas del learning analitycs. Estudio hecho usando las Bases de datos Universidad de Antioquia y bases de datos ICFES, Si bien se tenían datos de la universidad desde el 1996 y hasta 2019, para realizar el cruce con los datos del ICFES (Instituto Colombiano para la Evaluación de la Calidad de la Educación) se trabajó a partir del 2000. Los Instrumentos utilizados fueron dos técnicas de machine learning como son redes neuronales artificiales (RNA) y Xtreme gradient boosting (XG Boost) se entrenaron diferentes modelos. Se identificaron las técnicas del machine learning que comúnmente son utilizadas en la tarea de predecir la deserción y se seleccionaron dos técnicas con gran capacidad predictiva, se identificaron los parámetros que consiguieron el mejor desempeño de cada técnica para los datos que se tenía disponibles. también se determinaron dos clasificadore con el mayor rendimiento predictivo, el ajuste de los hiperparámetros mejora el rendimiento de cada algoritmo en comparación con su configuración de referencia.

Garcia (2019) Investigación en la Universidad Peruana Unión, Facultad de Ingeniería y Arquitectura Escuela Profesional de ingeniería de Sistemas cuyo objetivo fue Implementar un modelo computacional basado en las reglas de clasificación para la predicción de la deserción estudiantil en la Universidad Peruana Unión filial Juliaca. El trabajo utilizó 12000 registros de una base de datos Oracle 11g, desde el semestre 2016-1 a 2019 usando los métodos de clasificación machine learning, Python y las librerías de Scikit Learn. Luego de la modelación la predicción con nuevos datos es de 99.78% para

alumno que no deserta y 92.03% para alumno que deserta también se determina los factores que influyen en la deserción usando XGBoost seleccionando 12 factores.

Sanchez Nina (2017) en su tesis cuyo objetivo fue determinar el modelo estadístico para la deserción estudiantil en las escuelas profesionales de la UNA-Puno donde la población estuvo formada por 18499 estudiantes de la Universidad Nacional del altiplano Puno matriculados en el semestre 2017-II y la muestra 325 estudiantes de los cuales 92 son de ingeniería, 60 de biomédicas y 173 de sociales. Usando como instrumentos la Regresión logística binaria, software estadístico SPSS V.22 pruebas estadísticas Chi cuadrado. Obtuvo las siguientes conclusiones que el porcentaje de clasificación correcta usando el modelo es del 91.2%. se determinaron los coeficientes significativos del modelo de clasificación binaria también se determinó y codifico las variables para el modelo estadístico de deserción estudiantil en las escuelas profesionales de la UNA-PUNO usando la regresión logística binaria.

Candia Oviedo (2019) En su tesis cuyo objetivo fue predecir el rendimiento académico de los estudiantes de la UNSAAC del primer semestre a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático y que tomo como población o muestra a los estudiantes ingresantes a la USAAC desde el semestre 2014-I hasta el 2018-I y el tamaño de la muestra 12698 alumnos en sus diferentes modalidades. Usando como instrumentos la Información proporcionada por la unidad de cómputo de la universidad conteniendo encuestas a los postulantes e información de los estudiantes y sus promedios ponderados y los algoritmos de clasificación de aprendizaje automático. Obtuvo como resultados que el algoritmo Random Forest fue el de mejor performance y los factores clave son las notas de ingreso, escuela profesional, el semestre, el género entre los principales y el algoritmo de clasificación arboles aleatorios alcanzó el 69% de efectividad y los factores influyentes fueron determinados por medio del análisis estadístico chi cuadrado y el coeficiente de correlación de Pearson.

Otra investigación hecha en ETH Zurich-Departament of Management, Technology, and Economics, que como objetivo buscaba proporcionar una introducción estructurada y accesible al tema de aprendizaje automático interpretable y usaron como población o muestra de estudio de dos casos un conjunto de datos de Boston Housing y procesamiento del lenguaje natural y clasificación de los comentarios del youtube, se concluye que los resultados del primer estudio de caso muestra que la implementación existente de Python

de explicaciones contrafactuales no permite controlar la escasez y la viabilidad de las explicaciones. Los resultados del segundo estudio de casos muestran que la comprensibilidad de las explicaciones de LIME depende entre otras cosas; de la estructura de instancia de texto a explicar. (Benk y Ferrario, 2021)

Otra investigación hecha en Delhi Technological University, Delhi, India que buscaban dar una visión general del desarrollo del Machine Learning hasta la actualidad, diversos algoritmos de aprendizaje automático, aplicaciones y desafíos; como resultados importantes se concluye que la digitalización y la revolución de internet han llevado a un volumen creciente de datos estructurados y no estructurados que deben utilizarse para el análisis, el aprendizaje automático como impulsor tecnológico clave abarca el poder inteligente para aprovechar el conocimiento de los datos disponibles y finalmente se presentó una revisión exhaustiva del proceso y los algoritmos de aprendizaje automático.(Alzubi et al., 2018)

En la universidad de la Coruña, España se investiga acerca del aprendizaje automático y con el objetivo de indagar acerca del desarrollo de modelos de clasificación sobre datos con distribuciones muy desbalanceadas y para esta investigación se usa los bancos de datos de los repositorios mencionados, concluyeron que la librería SHAP de Python basado en el artículo de Lundberg and Lee (2017) y sirve para interpretar modelos. SHAP significa Shapley y Additive exPlanations y está basado en el valor de Shapley, es decir, en la media de las contribuciones marginales de cada variable al modelo también se interpretan los efectos locales y globales se concluye además, los modelos basados en árboles de decisión suelen obtener buenos resultados en los problemas de clasificación.(Piñeiro, 2022)

Otra de las investigaciones hechas en el Centro de innovación y desarrollo tecnológico en cómputo de la unidad de informática en México cuyo objetivo fue presentar los fundamentos teóricos de un nuevo modelo de clasificación que se basa en el enfoque asociativo de reconocimiento de patrones: clasificador de Heaviside que usando los bancos de datos de los repositorios KEEL Data-Mining Software Tool y UCI Machine Learning Repository. University of California. En esta investigación se concluye que el clasificador logra el 100% del rendimiento, validando con 10 fold cross y el peor rendimiento poco más del 50% estos resultados fueron validados, además, por la prueba no paramétrica de Wilcoxon.(Floriano, 2019)

También, en la Escuela técnica superior de ingeniería industrial de Barcelona se investiga acerca del aprendizaje automático y cuyo objetivo es entender que es el aprendizaje automático, comparar y entender teóricamente los algoritmos de clasificación, trabajar con los algoritmos de manera práctica usando programación Python y la librería Scikit-Learn, y concluyen que el algoritmo Máquina vector soporte alcanza una precisión de 0.99, Árboles de decisión con una precisión de 0.864 y finalmente Naive bayes una precisión de 0.909. El algoritmo Máquina vector soporte y Naive bayes presentan la mejor precisión. (Marrugat y Ginebra, 2020)

En la Universidad Nacional Agraria de la Selva de Tingo María-Perú, buscaron desarrollar un modelo de Machine Learning para la clasificación de imágenes satelitales en la Amazonia peruana y que usaron las imágenes satelitales Landsat 8 se usan los algoritmos de clasificación supervisada de Machine Learning se concluyen que Máquina vector soporte alcanza una precisión de 0.99, Árboles de decisión una precisión de 0.864 y Naive bayes presentaron la mejor precisión, por lo que los algoritmos de Máquina vector soporte y Naive bayes presentan la mejor precisión.(Chucos Baquerizo y Vega Ventocilla, 2022)

Zarate Valderrama, (2019) en su trabajo de tesis que buscaba identificar el patrón que siguen los estudiantes que desertan de sus estudios universitarios tomando como muestra no probabilística de 150 estudiantes de la escuela profesional de ingeniería de sistemas de la universidad nacional de san Agustín usando como instrumentos encuestas tipo cuestionario, utilizando preguntas abiertas y cerradas, técnicas de minería de datos. Concluye que un árbol de decisión mejorado para la clasificación y las principales características que influyen en la deserción y la identificación del patrón de deserción estudiantil, entrenamiento y validación de los modelos de clasificación e identificación de atributos, el algoritmo para la construcción de un árbol de decisión mejorado como el de mejores resultados según las métricas. Los atributos que influyen fueron los créditos actuales, las notas y el tipo de modalidad de examen de admisión.

La deserción estudiantil es un problema latente en nuestro país y esta involucra muchos factores entre los cuales podemos mencionar factores psicológicos, económicos, sociológicos, organizacionales y los que son de interacción entre el estudiante y la institución. Se considera que dichos factores permitirían predecir y prevenir el abandono y la perseverancia estudiantil universitaria. Dado que las variables pueden ser

modificadas para lograr disminuirla y prevenirla de manera significativa. La deserción universitaria es un problema que genera pérdidas económicas tanto en las universidades privadas como en universidades públicas, en las primeras a los padres puesto que ellos asumen la totalidad de los gastos que implican estudiar en una universidad privada y las universidades nacionales este problema le genera al estado un forado económico que le afecta al tesoro público.

Buscamos establecer el mejor modelo en ML para predecir la deserción estudiantil de los alumnos de la Universidad Nacional de Moquegua. El diseño de la investigación es no experimental y tipo de investigación explicativo correlacional, la población está constituida por 2305 alumnos y la muestra 762 los métodos a emplear son: Regresión Logística (RL), Máquina Vector Soporte (MVS), Arboles de Clasificación (DT) y Clasificador Bayesiano (NB); el resultado es obtener el mejor modelo para establecer las variables más influyentes en la deserción estudiantil.

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1 Identificación del problema

La deserción se puede definir como el proceso de abandono, voluntario o forzoso de la carrera en la que se matricula un estudiante, por la influencia positiva o negativa de circunstancias internas o externas a él o ella (Universidad de la República de Uruguay, 2003) (Fiegehen y Díaz, 2016)

Según un estudio realizado a nivel Latinoamérica y el Caribe, la deserción se analiza en tres factores: Los procesos de enseñanza, las características de los estudiantes tanto económicas, sociales y culturales y la interacción múltiple entre ellos. Se concluye que es un problema de gran interés para estudiarlo y enfrentarlo a nivel Latinoamérica y Caribe se menciona en este artículo que la deserción es un problema relevante y de creciente impacto social. En general, las universidades están en un proceso de análisis incipiente lo que impide comprender cabalmente el problema a nivel de países y buscar soluciones apropiadas. La superación de la deserción implicaría un cambio profundo en las metodologías docentes transitando de una pedagogía centrada en la enseñanza a una formación activa focalizada en el aprendizaje del estudiante. En síntesis, el gran desafío es, por una parte, incrementar la cobertura, la que en América Latina y el Caribe ha crecido a menor ritmo que en otros continentes y, por otra, disminuir la repitencia y deserción. Asimismo, es necesario mejorar la transición entre la formación académica y el empleo tanto al nivel de egreso como durante la carrera. Ello implica definir mejor los perfiles de egreso, mejorar la vinculación con el sector productivo e incrementar la formación basada en competencias. (Fiegehen y Díaz, 2016)

Existe un Importante estudio realizado por Logros.edu.pe revela dramática cifra que tiende a crecer año a año. "La deserción hizo que se pierdan 200 millones de dólares en los últimos dos años debido a la imposibilidad de cubrir las pensiones mensuales y en otros casos debido al simple desinterés por lo que estudian. Con ese monto los padres de familia podrían completar el tramo que le falta a la línea 1 del tren eléctrico." Explica de manera contundente el Ing. Rafael Plasencia, director del portal Logros quien ha realizado un completo estudio sobre las pérdidas causadas por la deserción universitaria. Este completo estudio señala, además, que, de continuar la tendencia, en los próximos diez años, se podrían perder más de 2 mil 100 millones de dólares, que es la cantidad que el estado comprometió para reconstruir las zonas afectadas por el terremoto del sur. El monto invertido (y perdido) por los padres de familia en pensiones estudios universitarios en el 2004 fue de 99.23 millones de dólares. Mientras que en el año 2005 se perdió más de 100 millones de dólares. En el mismo año no se graduaron 42 mil alumnos y se estima que esta tendencia continúe en aumento en los próximos años hasta superar los 70 mil no graduados para en el 2015. (Plasencia, 2018)

En la universidad nacional de Moquegua, como parte de los informes de autoevaluación la oficina de actividades y servicios informo que la deserción estudiantil en los años 2008 al 2010 oscilaba entre 11% y 20%, pero a la fecha no existe una estadística fehaciente. La universidad nacional de Moquegua no cuenta con un modelo predictivo que identifique los patrones predictivos para predecir futuros alternativos de deserción y que asista a la toma de decisiones para proponer protocolos y minimizar la tasa de deserción.(Diego, 2019)

2.2 Enunciado del problema

¿Cuál es el mejor modelo Machine Learning para explicar la deserción estudiantil de los alumnos de la Universidad Nacional de Moquegua?

2.3 Justificación

El problema de la deserción no es un problema exclusivo de Perú y mucho menos de una sola universidad y cada región tiene sus características, es un problema antiguo y que tiene muchas variables, es un problema complicado y de grandes repercusiones tanto para el estudiante como para instituciones y saber si un alumno va a desertar podría aliviar costos sociales y económicos.(Del Bonifro et al., 2020b)

El problema de la deserción universitaria provoca un incremento de alumnos con educación superior incompleta lo que perjudica a los estudiantes, a su entorno, al país y consecuentemente a la universidad afectando su presupuesto, las consecuencias de la deserción son graves ya que reflejan en una menor oportunidad laboral, estigma social, retribuciones económicas muy bajas, incremento de la delincuencia y otros que agravan nuestro entorno regional y nacional.(Buabeng-andoh, 2022)

Ocuparse de la deserción universitaria es muy complejo, pero importante ya que está considerado como un indicador de calidad de la gestión universitaria. (Tito, 2020).

Existen estudios realizados sobre las técnicas de aprendizaje supervisado y las técnicas de clasificación también técnicas de reducción de características las cuales pueden usarse para estudiar el problema de la deserción y su predicción.(Ayala-yaguara y Valenzuela-sabogal, 2020)

Las formas de analizar la deserción universitaria teniendo datos reales y usando los métodos de clasificación supervisada de Machine learning no es única y la metodología aplicar es diversa no se establece una única forma de hacer el análisis para lograr el objetivo y dependerá de diversos factores y en ocasiones un modelo conviene en relación a otros por lo que cada realidad es diferente, además no existe una única forma de hacer la reducción y selección de características que influyen en la deserción, existen nuevos métodos que son automáticos que son efectivos basados en algoritmos clásicos como la metodología Random Forest.(Mduma et al., 2019)

La presente investigación se realiza para conocer las variables particulares que intervienen en la deserción estudiantil universitaria en la Universidad Nacional de Moquegua y establecer un modelo de predicción, que servirá para poder proponer posteriormente políticas para evitar y prevenir la deserción.

En razón a lo expuesto, consideramos importante estudiarlo ya que al detectar las principales variables y predecir sus efectos puede mejorarse las condiciones y brindar tutorías focalizadas y también mejorar la malla curricular y establecer mecanismos de retención universitaria como respuesta a este problema y que esto redunde en la mejora de la calidad académica, una mejor mano de obra calificada y mayor oportunidad laboral y calidad de vida de los habitantes de la región Moquegua.

2.4 Objetivos

2.4.1 Objetivo general

Determinar el mejor modelo Machine Learning para analizar y predecir la deserción estudiantil de los alumnos de la Universidad Nacional de Moquegua.

2.4.2 Objetivos específicos

- Seleccionar las variables más significativas que explican la deserción estudiantil de los alumnos de la Universidad Nacional de Moquegua
- Obtener el mejor modelo de Regresión logística, Árboles de decisión, Máquinas vector soporte y Naive bayes para la deserción estudiantil en la Universidad Nacional de Moquegua.
- Obtener el mejor modelo de cuatro algoritmos/modelos de ML comparando las métricas de desempeño del modelo.

2.5 Hipótesis

2.5.1 Hipótesis general

El mejor modelo de clasificación supervisada para predecir la deserción estudiantil es regresión logística para los tipos de variables identificados.

2.5.2 Hipótesis específicas

- Las técnicas automáticas de Random forest y Featurewiz seleccionan las variables significativas que se utilizaran en el modelamiento de la deserción estudiantil de los alumnos de la Universidad nacional de Moquegua sede Mariscal Nieto.
- Las técnicas de clasificación supervisada de machine learning permiten obtener modelos adecuados para predecir la deserción estudiantil
- Las métricas permiten evaluar modelos Machine learning con mejor desempeño.

CAPÍTULO III

MATERIALES Y MÉTODOS

3.1 Lugar de estudio

El estudio se hará en la ciudad de Moquegua en la provincia Mariscal Nieto lugar de ubicación de la Universidad Nacional de Moquegua. La importancia radica en que, al lograr disminuir la deserción, esto se traduce a lograr más mano de obra calificada apta para trabajar y crear empleos competitivos de acuerdo con el contexto y realidad social, así como mejorar la calidad de vida de los habitantes de la región que se encuentra bajo la influencia de varios proyectos mineros en curso y en actividad.

3.2 Población

La población en la presente investigación estará constituida por tres Escuelas Profesionales, dos del área de ingeniería de la Universidad Nacional de Moquegua y una en el área de Ciencias, es decir que nuestra población está constituida por los estudiantes matriculados ingresantes y estudiantes que han reprobado o retirado por algún motivo semestres anteriores, entre 2009 hasta el 2019. Las Escuelas Profesionales en mención: Escuelas Profesionales de Ingeniería de Minas, Escuelas Profesionales de Ingeniería Agroindustrial y Escuelas Profesionales de Gestión Pública y Desarrollo Social. La característica principal de la población es que todas las Escuelas Profesionales desarrollan todos los semestres desde el primero hasta el décimo.

La población es un grupo de sujetos u objetos con características definitorias diversas. Por consiguiente, el investigador definirá la población en el supuesto de que todos los miembros o elementos tengan diferente valor como fuentes de información.

La población para la presente investigación estuvo conformada por 2305 y distribuida según se describe en la siguiente tabla:

Tabla 3

Tamaño de población

Alumnos	Tamaño de población
Alumnos que desertaron	765
Alumnos que no desertaron	1540
Total	2305

3.3 Muestra

La muestra se determinó por muestreo no aleatorio, quedando de la siguiente manera

Tabla 4

Tamaño de muestra

Alumnos	Tamaño de muestra
Alumnos que desertaron	109
Alumnos que no desertaron	220
Total	329

3.4 Método de investigación

Método hipotético deductivo, procedimiento que sigue el investigador para hacer de su actividad una práctica científica. El método hipotético-deductivo tiene varios pasos esenciales: observación del fenómeno a estudiar, creación de una hipótesis para explicar dicho fenómeno, deducción de consecuencias o proposiciones más elementales que la propia hipótesis, y verificación o comprobación de la verdad de los enunciados deducidos comparándolos con la experiencia.

Aprendizaje automático, El objetivo principal del aprendizaje automático es buscar patrones y construir algoritmos que puedan aprender de los datos anteriores para hacer predicciones sobre nuevos datos de entrada.

La investigación, de acuerdo con los objetivos que se plantea, corresponde a los siguientes tipos de investigación: Según su nivel, la investigación pertenece al tipo descriptivo y

correlacional, según el tipo de datos corresponde al tipo cuantitativo y de acuerdo con el propósito, la investigación es de tipo aplicada.

Para comparar los modelos de clasificación Logística, Árboles de decisión, Machine Vector Support y Naive Bayes, se consideraron 40 variables (8 numéricas, 31 categóricas y 1 objetivo) y 329 datos (109 desertores y 220 no desertores), luego de codificar algunas variables se consideró 94 variables (8 numéricas, 86 categóricas y 1 objetivo) y la variable objetivo es la deserción (1 si deserta y 0 si no deserta), también en el proceso se consideró 70% de datos para entrenamiento y 30% para evaluación. Las evaluaciones se realizaron con la mayor cantidad de atributos, pero existen varios autores indican que esto puede afectar el desempeño del modelo (Raschka et al., 2022). Se aplican dos métodos de selección de características o atributos el feature importance que se fundamenta en ramdon forest, mide la importancia del atributo mientras se van reduciendo los datos con baja correlación (Raschka et al., 2022); el segundo es featurewiz que es una nueva librería de Python para crear modelos de alto rendimiento una vez que haya completado la selección de los mejores atributos, featurewiz examinará numerosas variables de este tipo y encontrará solo las características menos correlacionadas y más relevantes para su modelo, 'featurewiz' utiliza el algoritmo MRMR (Minimum Redundancy Maximum Relevance) como base para su selección de características(AutoViML/featurewiz: use estrategias avanzadas de ingeniería de funciones y seleccione las mejores funciones de su conjunto de datos con una sola línea de código., 2022).

El objetivo final de la reducción de características es aumentar la precisión predictiva y reducir la complejidad de los resultados.(Sivakumar et al., 2016)

3.5 Descripción detallada por objetivo especifico

Con el propósito de cumplir los objetivos: general y específicos de la investigación, se procedió de la siguiente manera:

En primer lugar, se recopiló la información mediante dos técnicas de recolección de datos: observación y encuesta, utilizando el instrumento: ficha socioeconómica del estudiante y que se encuentra en la dirección de bienestar universitario de la Universidad Nacional de Moquegua.

En segundo lugar, teniendo en cuenta las características del trabajo el diseño es descriptivo transversal, la metodología CRISP-DM en el procesamiento y preparación de

los datos, se eliminaron las variables con muchos datos faltantes y se consideró reemplazar la mediana o la moda de acuerdo a si el dato era categórico o numérico respectivamente, cambiando la escala de los datos donde sea necesario y haciendo que los datos numéricos se ajusten a una distribución normal para un mejor tratamiento de los mismos.

En tercer lugar, se usó el software Python 3.9.12 debido a su codificación legible, conciso y de fácil aprendizaje, además Python ofrece una gran cantidad de bibliotecas para ML como Scikit-learn, Tensor Flow y Keras, NumPy para el análisis de datos y rendimiento científico, SciPy para informática avanzada, Seaborn para visualización de datos y Pandas para el análisis de datos.

En cuarto lugar, para comparar los modelos de clasificación Logística, Árboles de decisión, Machine Vector Support y Naive Bayes, se consideraron 40 variables (8 numéricas, 31 categóricas y 1 objetivo) y 329 datos (109 desertores y 220 no desertores), luego de codificar algunas variables numéricas se consideró 94 variables (8 numéricas, 86 categóricas y 1 objetivo) y la variable objetivo es la deserción (1 si deserta y 0 si no deserta), en la evaluación de los modelos y evitar el sobre entrenamiento se usa la técnica de retención que consiste en dividir los datos en 70% de datos para entrenamiento y retención de 30% para evaluación, también la técnica de la validación cruzada con K=5 y K=10 Folds, adicionalmente usamos el Área de la curva ROC.

En quinto lugar, se realizó la selección de variables más importantes para el modelamiento de los datos, utilizando la Random Forest (RF) y Featurewiz(FW) que permiten seleccionar todas las características y su contribución, por medio de un conjunto de datos de entrenamiento y otro de prueba, obteniendo los datos codificados de la tabla 55.

Posteriormente se evaluaron cuatro algoritmos: Regresión logística, Árboles de decisión, Máquina vector soporte y Naive Bayes utilizando parámetros de ajuste convenientes. Finalmente, de acuerdo a la evaluación se mostraron los resultados obtenidos con la utilización de los algoritmos para la investigación.

Técnicas e Instrumentos:

Las técnicas e instrumentos usados son:

Para la variable deserción la técnica es la observación y el Instrumento es la Ficha socio económica (Ver ANEXO 1).

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1 Identificación de las variables

Diccionario de datos inicial de acuerdo con cada ficha socio económica (Ver ANEXO 1) según los siguientes aspectos:

I. Datos Generales del Estudiante
II. Antecedentes Académicos
III. Aspecto Económico
IV. Aspecto de Vivienda
V. Aspecto de Salud
VI. Diagnóstico Social

Algunos de las características no se consideraron debido a que la información no estaba completa, como código, apellidos y nombres, fecha de nacimiento, dirección domiciliaria, celular, composición familiar, del estudiante foráneo, aspecto de salud y el diagnóstico social. Al respecto en las fichas socioeconómicas se observan que muchos de los datos solicitados no son respondidos o proporcionados por los estudiantes quizás porque no se entiende lo que se solicita o no lo consideraron importantes.

4.2 Codificación de variables

La siguiente tabla muestra las variables y su codificación luego de haber hecho la limpieza y la imputación de datos.

Tabla 5

Variables Consideradas y su codificación.

ATRIBUTO	DESCRIPCIÓN
DESERTA	1= SI deserta; 0= No deserta, binario
ESCUELA PROFESIONAL	EP_MINAS.EP_AGRO, EP GPDS, binario 0=No, 1=Si
CICLO	CICLO12, CICLO34, CICLO56, CICLO78, CICLO910, binario 0=No, 1=Si
DATOS GENERALES	
SEXO	1: femenino y 2: masculino
EDAD	Variable normalizada EDAD_N
ESTADO CIVIL	ESTCIV_S, ESTCIV_C, ESTCIV_OT binarias 0=No, 1=Si
LUGAR DE NACIMIENTO	LNAC_MOQ, LNAC_ILO, LNAC_SC, LNAC_OTRO binarias 0=No, 1 =Si
DIR ZONA	DIRZON_URB, DIRZON_RUR, DIRZON_OTR binarias 0=No, 1=Si
PROVINCIA	PROV_MCLN, PROV_ILO, PROV_SC Binarias 0=No, 1= Si
ANTECEDENTES ACADÉMICOS	
TIPO COLEGIO	TIPO_PRIM, TIPO_SEC, 1=Nacional, 2=Particular
PREPARACIÓN PREUNIVERSITARIA	PRE_PP, PREU_AC, PREU_CEPRE, PREU_SOL, binarias 0=No, 1= Si
MODALIDAD DE INGRESO	MD_ORDIN, MD_CEPRE, MD_EXTR, binarias 0=No,1=Si
ASPECTO ECONOMICO	
COMPOSICIÓN FAMILIAR	COM_HOGAR número de personas que componen el hogar, numérica.
GRADO DE INSTRUCCIÓN DE LOS PADRES	1=SIN INSTRUCCIÓN,2=PRIMARIA,3=SECUNDARIA, 4=SUPERIOR
SOSTIENE EL HOGAR	1=PADRE Y MDRE,2=PADRE,3=MADRE,4=FAM. TUTOR, 5=ALUMNO
MODALIDAD DE INGRESO ECONÓMICO	1=MENSUAL,2=QUINCENAL,3=SEMANAL,4=DIARIO
CONDICION LABORAL DEL ESTUDIANTE	1=TRABAJA, 0=NO TRABAJA
INGRESO ECONOMICO FAMILIAR	TOTAL_INGRE
VIVIENDA DEL ESTUDIANTE	VIV_CPADRES, VIV_C1PAD, VIV_ALOJ, VIV_CUID, VIVCALQ, binarias 0=No, 1=Si
CARGA FAMILIAR	CARGA_FAM, NUMERO DE MIEMBROS
HIJOS QUE CURSAN E. SUPERIOR	HIJOS_SUP, NÚMERO HIJOS EDUCACIÓN SUPERIOR
DEL ESTUDIANTE	
DEPENDENCIA ECONÓMICA	1=AMBOSPADRES,2= SÓLO PADRE, 3=SÓLO MADRE, 4= UN FAMILIAR, 5 =DE SÍ MISMO
RIESGO FAMILIAR	1=HIJO DE PADRES VIVOS,2=HUERFANO DE MADRE, 3=HUERFANO DE PADRE, 4=HUERFANO DE PADRE Y MADRE, 5=VIVE SÓLO
ASPECTO DE VIVIENDA	
TENENCIA DE VIVIENDA	1=PROPIA, 2=ALQUILADA,3=INVASIÓN,4=OTRO
TIPO DE COSNTRUCCIÓN	1=NOBLE,2=MIXTO,3=RÚSTICO,4=PRECARIO
TIPO DE VIVIENDA	1=INDEPENDIENTE,2=DEPARTAMENTO, 3=CONVENTILLO, 4=OTRO
SERVICIOS	SERV_AGUA, SERV_DESAG, SERV_FONO, binarias 0=No, 1=Si

NÚMERO DE PISOS	N_PISOS
NUMERO DE DORMITORIOS	N_DORM
NÚMERO DE COCINAS	NCOCINA
NÚMERO DE BAÑOS	NBANO
NÚMERO DE SALAS	NSALA
NÚMERO DE COMEDOR	NCOMEDOR
TV_COLOR	binaria, 0=No, 1=Si
RADIO	binaria, 0=No, 1=Si
EQUIP_SONIDO	binaria, 0=No, 1=Si
PLANCHA	binaria, 0=No, 1=Si
CELULAR	binaria, 0=No, 1=Si
LAPTOP	binaria, 0=No, 1=Si
CABLE	binaria, 0=No, 1=Si
ROPERO	binaria, 0=No, 1=Si
REFRIGERADOR	binaria, 0=No, 1=Si
INTERNET	binaria, 0=No, 1=Si
BIBLIO_PERS	binaria, 0=No, 1=Si
COMPUTADORA	binaria, 0=No, 1=Si

4.3 Análisis Descriptivo Univariado

El estudio se realizó con una muestra de 109 fichas socio económicas de los estudiantes que desertaron y 220 de los que no desertaron de las tres escuelas profesionales Ing. de Minas, Ing. De Agroindustrial y de Gestión Pública y Desarrollo Social, correspondiente a la provincia Mariscal Nieto entre los años 2009 y 2019.

Para el análisis descriptivo en el indicador Escuela profesional según la Tabla 6 de 109 fichas, la mayor parte de fichas que corresponden a los alumnos en deserción, el mayor porcentaje es para la EP de Ing. De Minas con 38%, seguido de Gestión Pública y Desarrollo Social con 29% y la EP de Ing. Agroindustrial que tiene 32%.

Tabla 6

Deserción según la Escuela profesional

EP	Frecuencia	Porcentaje	Porcentaje Acumulado
EP_AGRO	35	32%	32%
EP_GPDS	32	29%	61%
EP_MINAS	42	38%	100%
Total	109	100%	

En la Tabla 7, de manera similar los alumnos que permanecen son en su mayoría de la EP Ing. De minas seguidos de las otras dos escuelas Ing. Agroindustrial y Gestión Pública ambos con el mismo porcentaje 33%.

Tabla 7

Permanencia según la Escuela profesional

EP	Frecuencia	Porcentaje	Porcentaje Acumulado
EP_AGRO	73	33%	33%
EP_GPDS	73	33%	66%
EP_MINAS	74	34%	100%
Total	220	100%	

En la Figura 2. Se aprecia las comparaciones entre los alumnos que desertan y los que permanecen según la escuela profesional, es claro que los que permanecen siempre serán en volumen mayor a los que no permanecen.

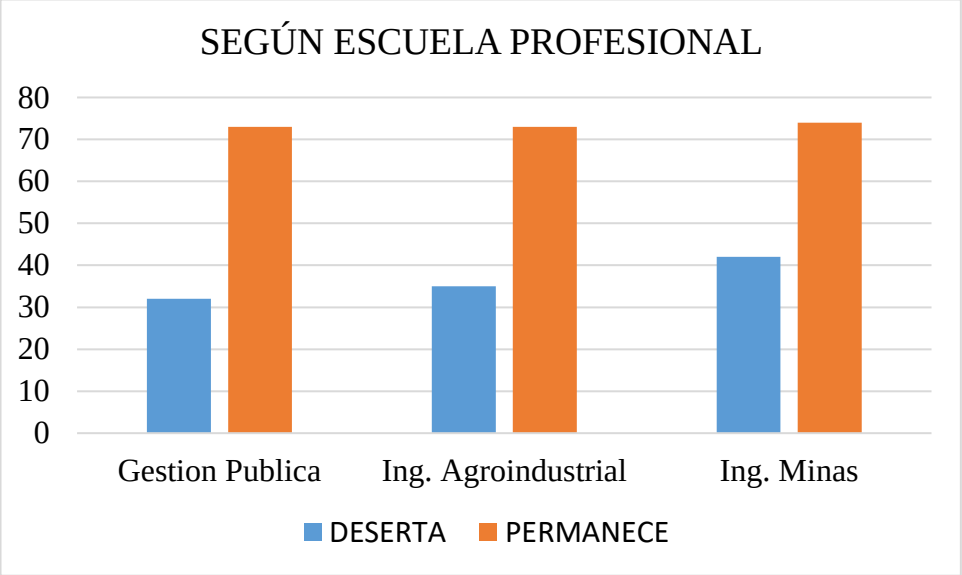


Figura 2. Deserción según Escuela Profesional

También, en la Tabla 8. según el ciclo de estudios observamos que en el caso de la variable CICLO los que desertan hasta en 90% en el primer ciclo y 10% en Segundo, tercero y quinto.

Tabla 8

Deserción Según el Ciclo de estudios

CICLO	Frecuencia	Porcentaje	Porcentaje Acumulado
Primero	98	90%	90%
Segundo	6	6%	96%
Tercero	2	2%	98%
Quinto	3	2%	100%
Total	109	100%	

La Tabla 9, sobre la permanencia de los alumnos según el ciclo de estudios observamos que en el caso de la variable CICLO los que permanecen disminuye en relación con la tabla 8 de 90% a 55% en el primer ciclo y 17%, 11% y 12% en Segundo, tercero y cuarto y la diferencia en los demás ciclos.

Tabla 9

Permanencia según el Ciclo de estudios

CICLO	Frecuencia	Porcentaje	Porcentaje Acumulado
Primero	55	25%	25%
Segundo	39	17%	42%
Tercero	24	11%	53%
Cuarto	27	12%	65%
Quinto	19	9%	74%
Sexto	17	8%	82%
Sétimo	10	5%	87%
Octavo	22	10%	97%
Noveno	7	3%	100%
Total	220	100%	

La Figura 3, muestra que la mayor incidencia en cuanto a deserción se encuentra en el primer ciclo y algo menos en el segundo ciclo.

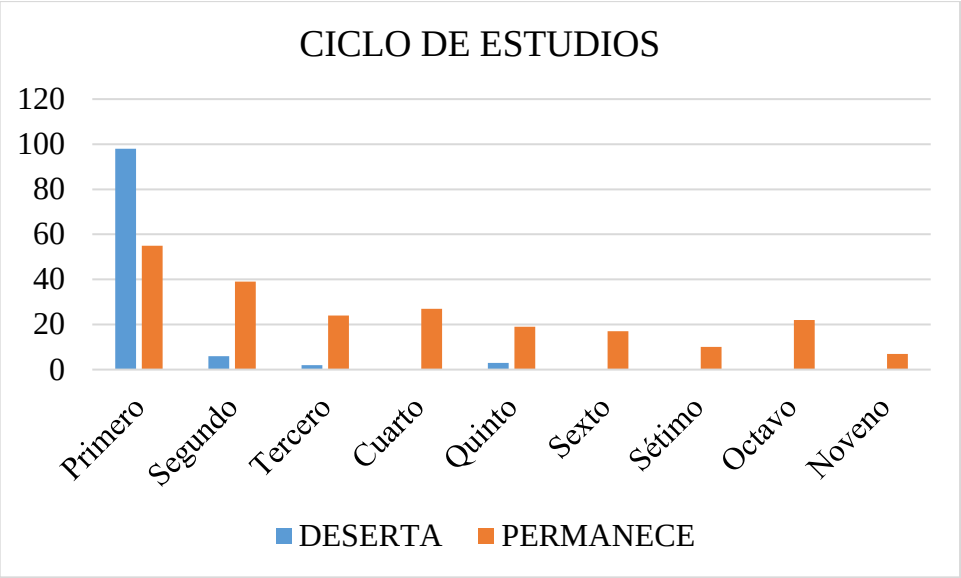


Figura 3. Según el Ciclo de Estudios

Para el caso de la Tabla 10 según el GÉNERO la deserción incide en un 63% para el caso MASCULINO ósea es mayor y en el caso FEMENINO el resto 37%

Tabla 10

Deserción Según el Género

GÉNERO	Frecuencia	Porcentaje	Porcentaje Acumulado
Femenino	40	37%	37%
Masculino	69	63%	100%
Total	109	100%	

La Tabla 11 según el GÉNERO la permanencia incide en un 51% para el caso MASCULINO ósea es mayor y en el caso FEMENINO el resto 49%

Tabla 11

Permanencia Según el Género.

GÉNERO	Frecuencia	Porcentaje	Porcentaje Acumulado
Femenino	107	49%	49%
Masculino	113	51%	100%
Total	220	100%	

La Figura 4 muestra de manera gráfica la deserción y permanencia según el género la mayor cantidad de alumnos varones tienen la tendencia a desertar.

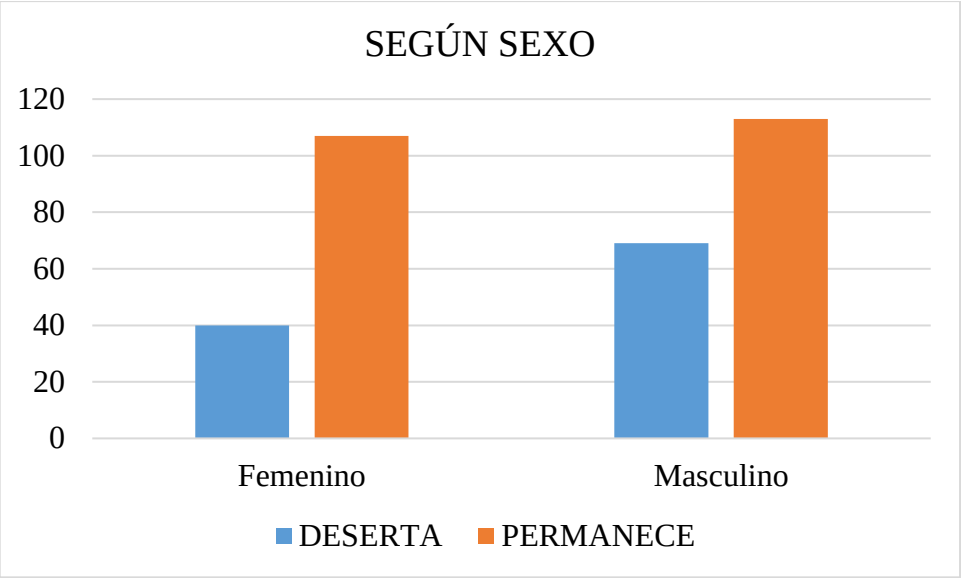


Figura 4. Deserción Según el Género

En la Tabla 12 el porcentaje mayor 97% corresponde a la categoría Soltero y la diferencia que es 3% está distribuido en el rubro de Casado.

Tabla 12

Deserción Según el ESTADO CIVIL

EST_CIV	Frecuencia	Porcentaje	Porcentaje Acumulado
Casado	3	3%	3%
Soltero	106	97%	100%
Total	109	100%	

La Tabla 13, la permanencia, el porcentaje mayor 97% corresponde a la categoría Soltero y la diferencia que es 3% está distribuido en el rubro de Casado y otro.

Tabla 13

Resumen permanencia según estado civil

EST_CIV	Frecuencia	Porcentaje	Porcentaje Acumulado
Casado	5	2%	2%
Soltero	214	97%	99%
Otro	1	0%	100%
Total	109	100%	

La Figura 5 ilustra que los alumnos que desertan como los que permanecen son en su mayoría solteros.

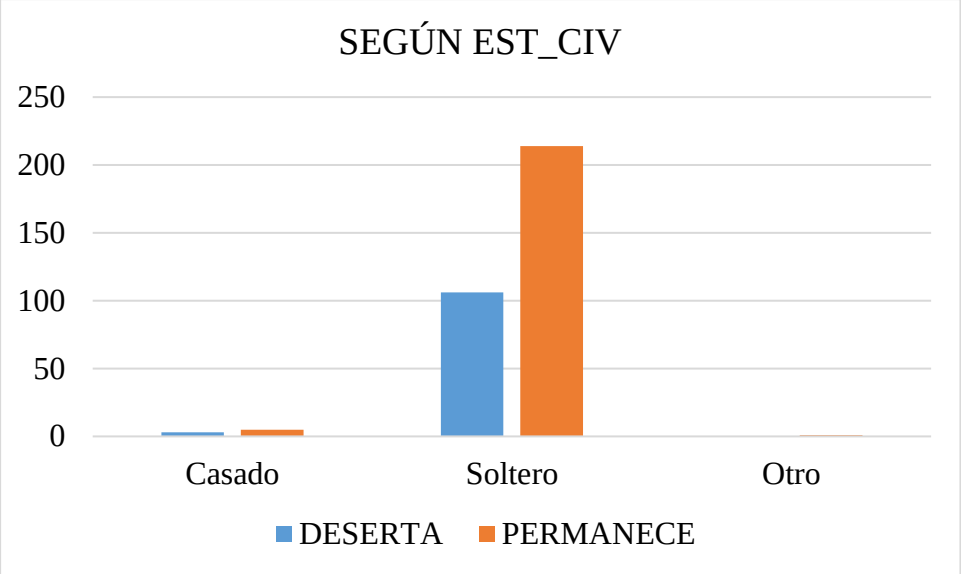


Figura 5. Según el Estado Civil

En la Tabla 14, la deserción según la categoría DIR_ZONA tiene a Cercado con 30% seguido de San Antonio con 28% y a continuación Otro con el 20% seguido de otros centros poblados y distritos distribuido la diferencia algo del 22%.

Tabla 14

Deserción según DIR_ZONA

DIR_ZONA	Frecuencia	Porcentaje	Porcentaje Acumulado
Cercado	33	30%	30%
Chen chen	6	6%	36%
Los Ángeles	4	4%	39%
Otro	20	18%	58%
Samegua	6	6%	63%
San Antonio	28	26%	89%
San Francisco	11	11%	100%
Total	109	100%	

La Tabla 15, la permanencia según la categoría DIR_ZONA tiene a Cercado con 30% seguido de San Antonio con 26% y a continuación Otro con el 18% seguido de otros centros poblados y distritos distribuido la diferencia algo del 27%.

Tabla 15

Resumen permanencia según DIR_ZONA

DIR_ZONA	Frecuencia	Porcentaje	Porcentaje Acumulado
Cercado	87	30%	30%
Chen chen	13	6%	36%
Los Ángeles	2	4%	39%
Otro	51	18%	58%
Samegua	19	6%	63%
San Antonio	39	26%	89%
San Francisco	9	11%	100%
Total	220	100%	

La Figura 6, muestra que la deserción tiene mayor incidencia en alumnos que viven en el cercado san Antonio y otro que correspondería a alumnos que no viven en Moquegua.

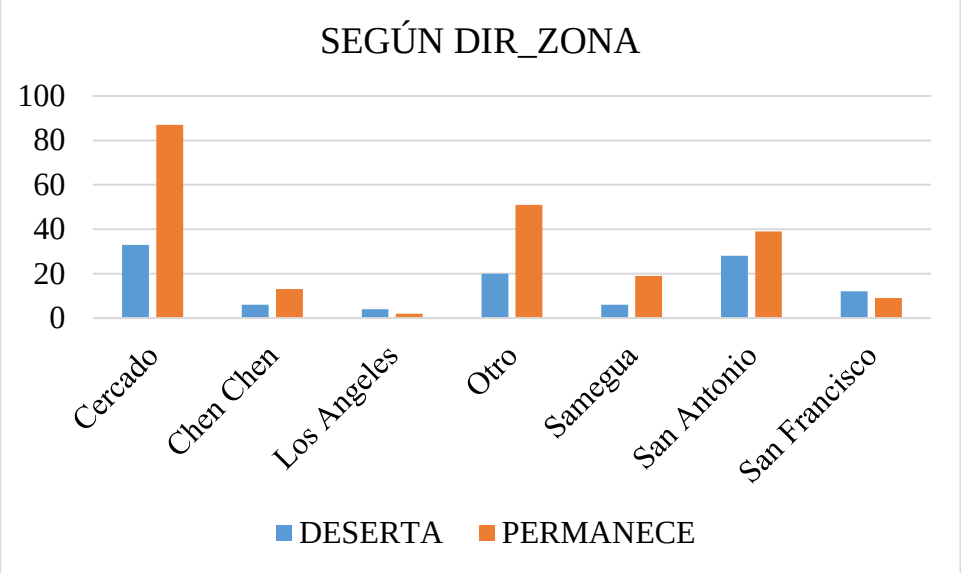


Figura 6. Gráfico según DIR_ZONA

En la Tabla 16, la deserción según la categoría PROVINCIA tiene a Moquegua con 69% seguido de Otro con 19% y el resto distribuido entre Sánchez Cerro e Ilo.

Tabla 16

Deserción según PROVINCIA

PROVINCIA	Frecuencia	Porcentaje	Porcentaje Acumulado
Ilo	6	6%	6%
Moquegua	75	69%	74%
Sánchez Cerro	21	6%	80%
Otro	7	19%	100

Total	109	100%

En la Tabla 17, la permanencia según la categoría PROVINCIA tiene a Moquegua con 65% seguido de Otro con 29% y el resto distribuido entre Sánchez Cerro e Ilo.

Tabla 17

Resumen permanencia según PROVINCIA

PROVINCIA	Frecuencia	Porcentaje	Porcentaje Acumulado
Ilo	9	4%	4%
Moquegua	143	65%	69%
Sánchez Cerro	5	2%	71%
Otro	63	29%	100%
Total	220	100%	

La Figura 7 grafica una mayor concentración de alumnos que desertan en Moquegua capital.

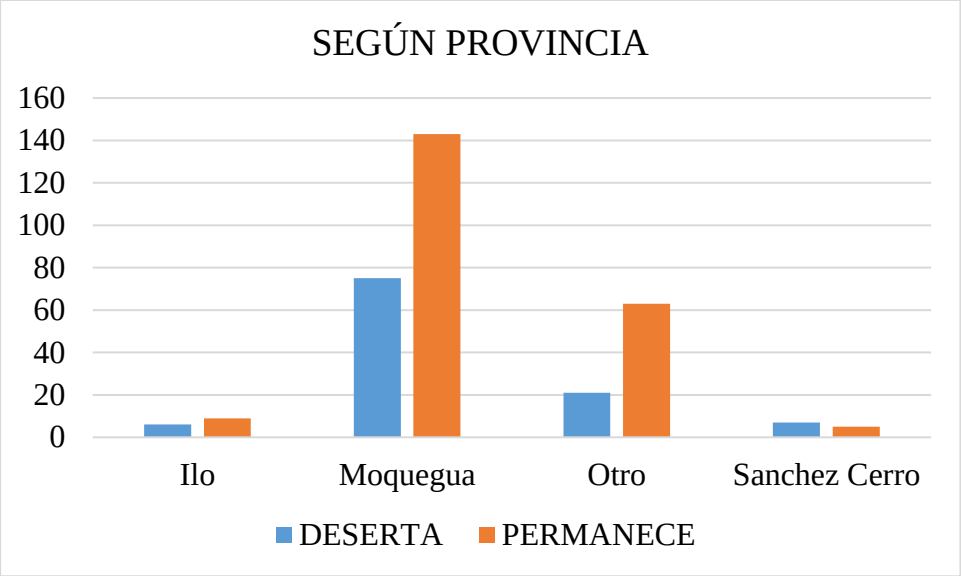


Figura 7. Gráfico Según PROVINCIA

En la Tabla 18, la deserción según el tipo de Primaria incide en mayor porcentaje los alumnos provenientes de colegios de primaria estatal con 94% y con el rubro de privado en 6% que incluirían a los colegios parroquiales.

Tabla 18

Deserción según TIPO_PRIM

TIPO_PRIM	Frecuencia	Porcentaje	Porcentaje Acumulado
ESTATAL	103	94%	94%
PRIVADO	6	6%	100%
Total	109	100%	

En la Tabla 19, la permanencia según el tipo de Primaria incide en mayor porcentaje los alumnos provenientes de colegios de primaria estatal con 92% y con el rubro de privado en 8% que incluirían a los colegios parroquiales.

Tabla 19

Permanencia según TIPO_PRIM

TIPO_PRIM	Frecuencia	Porcentaje	Porcentaje Acumulado
ESTATAL	202	92%	92%
PRIVADO	18	8%	100%
Total	220	100%	

La Figura 8, ilustra la permanencia y deserción según TIPO_PRIM.

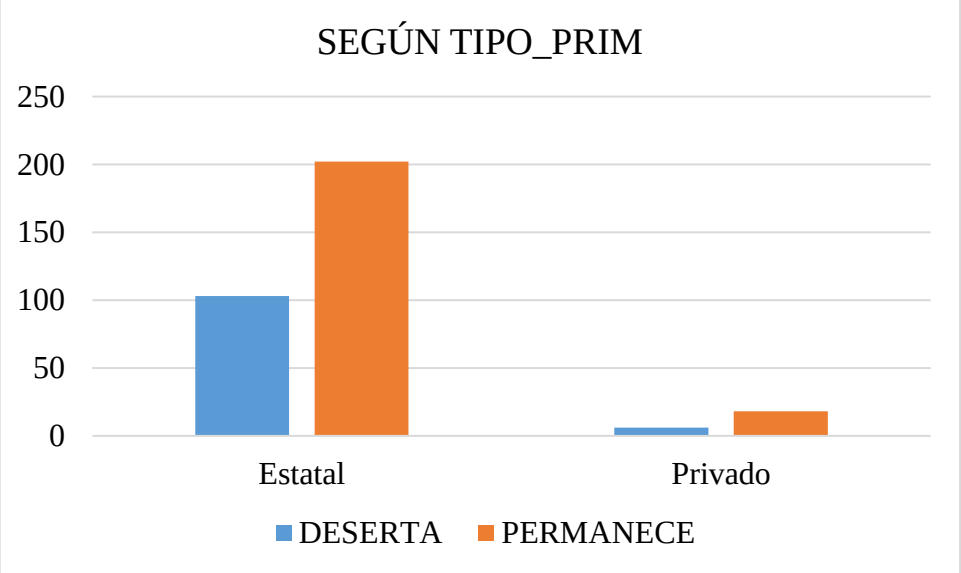


Figura 8. Según el Tipo de Primaria

En la Tabla 20, deserción según el tipo de secundaria se observa que un 87% estudia en un colegio estatal y en colegio Privado 13%.

Tabla 20

Deserción según TIPO_SEC

TIPO_SEC	Frecuencia	Porcentaje	Porcentaje Acumulado
Estatal	95	87%	87%
Privado	14	13%	100%
Total	109	100%	

También, la permanencia según tipo de secundaria la mayor incidencia está colegio secundario estatal con 90% seguido del rubro secundario privado con 10%.

Tabla 21

Permanencia según TIPO_SEC

TIPO_SEC	Frecuencia	Porcentaje	Porcentaje Acumulado
Estatal	197	90%	90%
Privado	23	10%	100%
Total	220	100%	

La Figura 9, muestra la proporción de la permanencia y deserción según el tipo de secundaria correspondiendo la mayor a colegio secundario estatal.

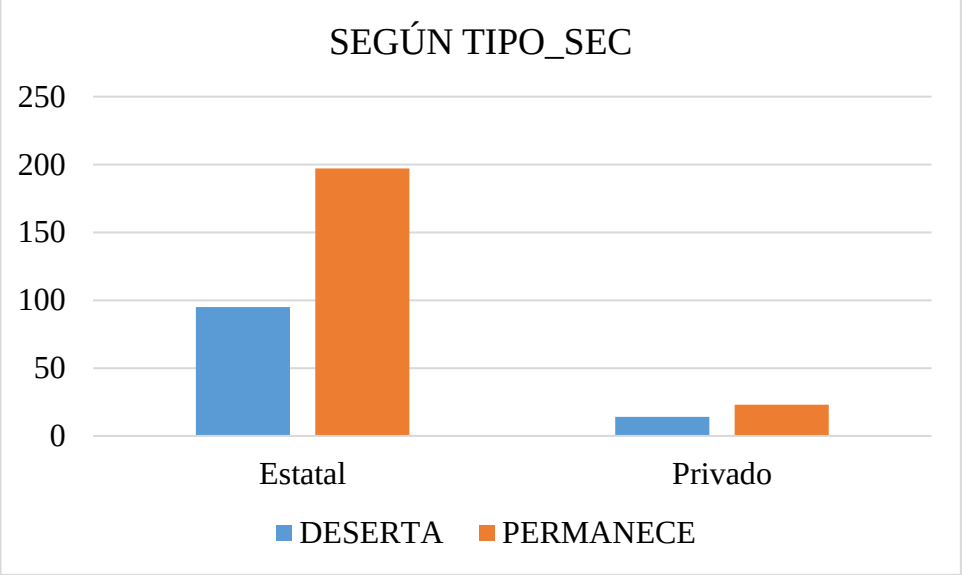


Figura 9. Según el Tipo de Secundaria

Considerando la deserción según la variable Preparación Preuniversitaria y según se observa en la Tabla 22 la mayor incidencia corresponde a preparación por su cuenta con 55% seguido de CEPRE con 39% y academia particular con 6%.

Tabla 22

Deserción según PREP_PREU

PREP_PREU	Frecuencia	Porcentaje	Porcentaje Acumulado
Acad. Particular	6	6%	6%
Cepre	43	39%	45%
Por su cuenta	60	55%	100%
Total	109	100%	

Según permanencia en la variable Preparación Preuniversitaria y se observa en la Tabla 23 la mayor incidencia corresponde a preparación por su cuenta con 53% seguido de CEPRE con 45% y academia particular con 3%.

Tabla 23

Permanencia según PREP_PREU

PREP_PREU	Frecuencia	Porcentaje	Porcentaje Acumulado
Acad. Particular	6	3%	3%
Cepre	98	45%	47%
Por su cuenta	116	53%	100%
Total	220	100%	

Gráficamente la Figura 10 presenta que la mayor parte de alumnos se prepara por su cuenta.

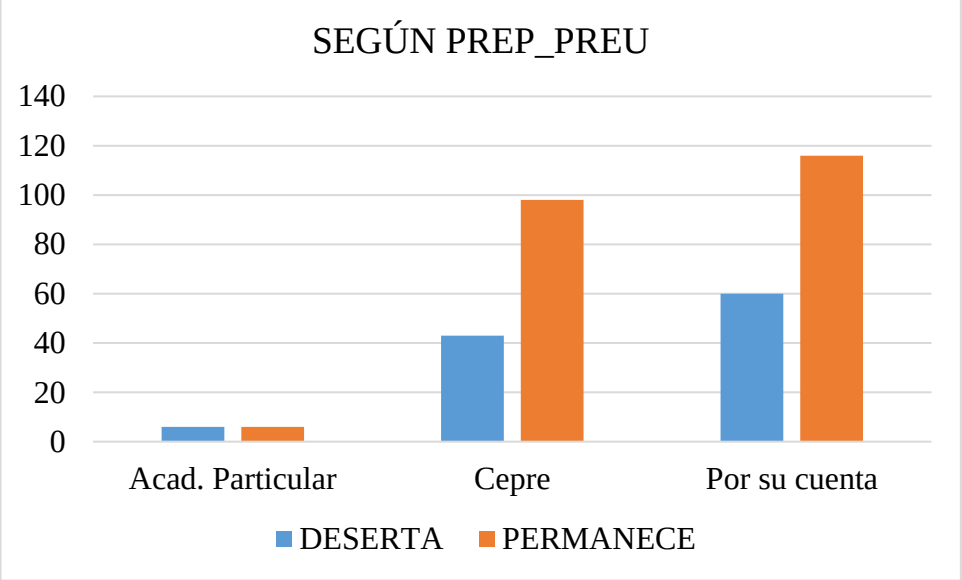


Figura 10. Según Preparación universitaria

En la Tabla 24 se observa que la deserción, según modalidad de ingreso predominante es el examen ordinario 75% y la modalidad CEPRE con 22% en segunda instancia y el resto 3% en Extraordinario.

Tabla 24

Deserción Según la Modalidad de Ingreso

MOD_ING	Frecuencia	Porcentaje	Porcentaje Acumulado
Cepre	24	22%	22%
Extraordinario	3	3%	25%
Ordinario	82	75%	100%
Total	109	100%	

La Tabla 25, la permanencia, según modalidad de ingreso predomina examen ordinario 70% y la modalidad CEPRE con 27% en segunda instancia y el resto 2% en Extraordinario.

Tabla 25

Permanencia según MOD_ING

MOD_ING	Frecuencia	Porcentaje	Porcentaje Acumulado
Cepre	60	27%	27%
Extraordinario	5	2%	30%
Ordinario	154	70%	100%
Por su cuenta	1	0%	100%
Total	109	100%	

La Figura 11 representa las frecuencias respectivas sobre la modalidad de ingreso para los alumnos que desertan y permanecen.

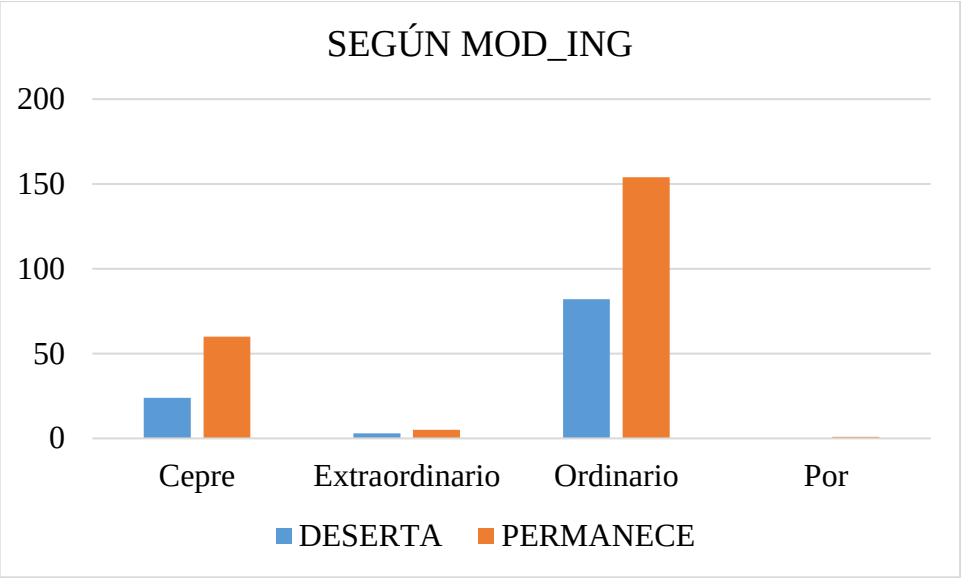


Figura 11. Según Modalidad de Ingreso

La Tabla 26 resume la información relativa a la preparación de los padres y observamos que predomina la instrucción superior 60% seguido de 29% que corresponde a Secundaria y una cantidad menor de 11% al rubro de Primaria en el caso del padre.

Tabla 26

Deserción según INSTRUC_PADRE

INSTRUC_PADRE	Frecuencia	Porcentaje	Porcentaje Acumulado
Primaria	12	11%	11%
Secundaria	32	29%	40%
Superior	65	60%	100%
Total	109	100%	

La Tabla 27, la permanencia relativa a la preparación de los padres y observamos que predomina la instrucción superior 64% seguido de 28% que corresponde a Secundaria y una cantidad menor de 8% al rubro de Primaria en el caso de la madre.

Tabla 27

Permanencia según INSTRUC_PADRE

INSTRUC_PADRE	Frecuencia	Porcentaje	Porcentaje Acumulado
Primaria	17	8%	8%
Secundaria	62	28%	36%
Superior	141	64%	100%
Total	220	100%	

La Figura 12, representa las frecuencias de instrucción de los padres de manera gráfica mostrando una mayor predominancia a que los padres tienen instrucción superior.

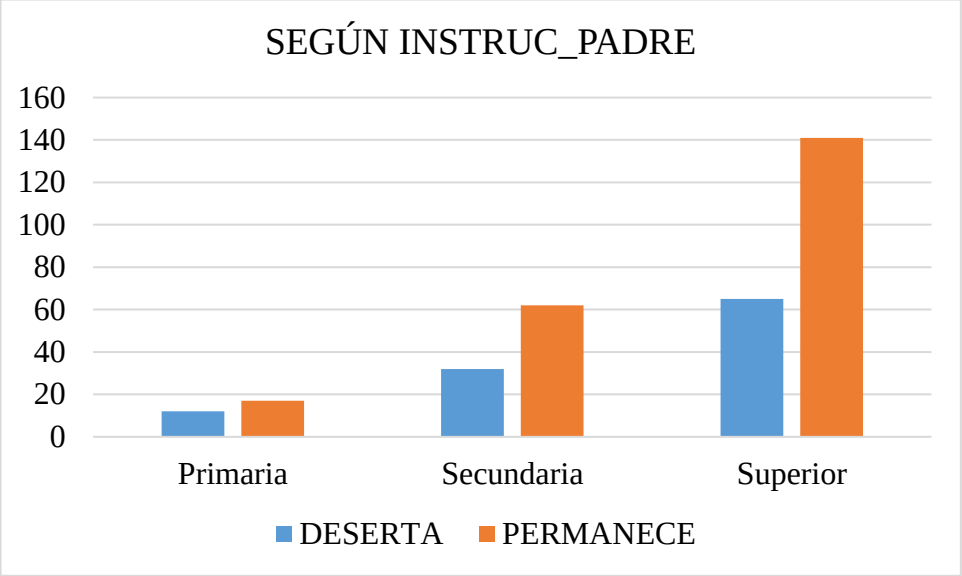


Figura 12. Según Instrucción del Padre

En la Tabla 28, según la deserción se resume la información referente a la instrucción de la madre indicando que el mayor porcentaje está en secundaria con 44% seguido de instrucción superior con 28% y esta tabla aparece Primaria con 28%.

Tabla 28

Deserción según INSTRUC_MADRE

INSTRUC_MADRE	Frecuencia	Porcentaje	Porcentaje Acumulado
Primaria	30	28%	28%
Secundaria	49	44%	72%
Superior	30	28%	100%
Total	109	100%	

En la Tabla 29, según la permanencia se resume la información referente a la instrucción de la madre indicando que el mayor porcentaje está en secundaria con 60% seguido de instrucción superior con 25% y esta tabla aparece Primaria con 16%.

Tabla 29

Permanencia según INSTRUC_MADRE

INSTRUC_MADRE	Frecuencia	Porcentaje	Porcentaje Acumulado
Primaria	35	16%	16%
Secundaria	131	60%	75%
Superior	54	25%	100%

Total	220	100%

A continuación, la Figura 11 representa las frecuencias gráficamente sobre la preparación de la madre según el grado de instrucción.

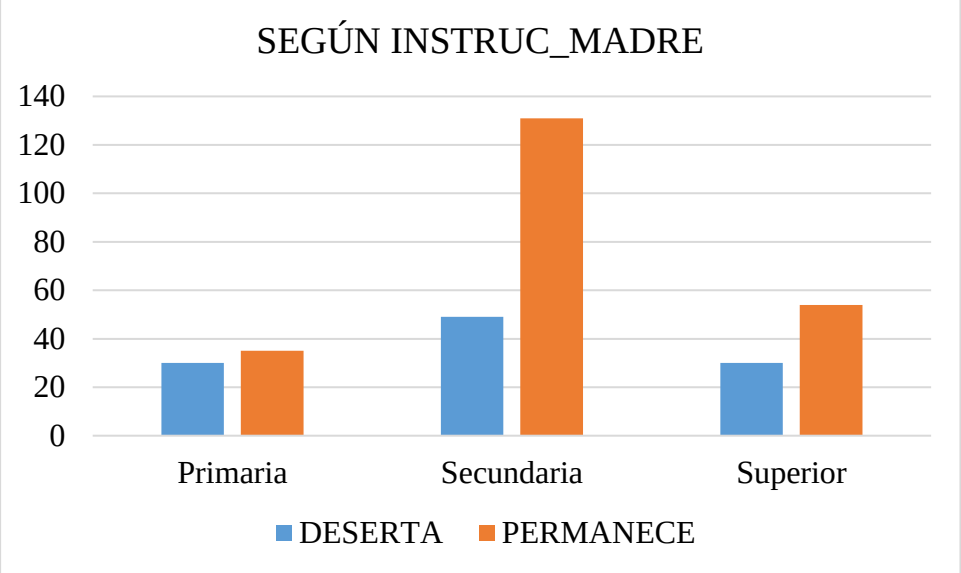


Figura 13. Según Instrucción de la Madre

Respecto a quien sostiene el hogar el 47%, en la Tabla 30, indica que son los padres quienes sostienen el hogar con 47% y secunda el padre con 25%, 17% la Madre y en porcentaje mínimo de 6% para Familiar/Tutor y Alumno.

Tabla 30

Deserción según SOST_HOGAR

SOST_HOGAR	Frecuencia	Porcentaje	Porcentaje Acumulado
Alumno	7	6%	6%
Familiar/ Tutor	6	6%	12%
Madre	18	17%	28%
Padre	27	25%	53%
Padre y Madre	51	47%	100%
Total	109	100%	

Respecto a permanencia sobre quien sostiene el hogar el 36%, en la Tabla 31, indica que son los padres quienes sostienen el hogar y secunda la madre con 27%, luego padre con 19%, y en porcentaje mínimo de 7% para Familiar/Tutor y Alumno 11%.

Tabla 31

Permanencia según SOST_HOGAR

SOST_HOGAR	Frecuencia	Porcentaje	Porcentaje Acumulado
Alumno	25	11%	11%
Familiar/ Tutor	15	7%	18%
Madre	59	27%	45%
Padre	42	19%	64%
Padre y Madre	79	36%	100%
Total	109	100%	

La Figura 14, representa las frecuencias de la categoría SOST_HOGAR de manera gráfica tanto para los que desertan como para los que permanecen.

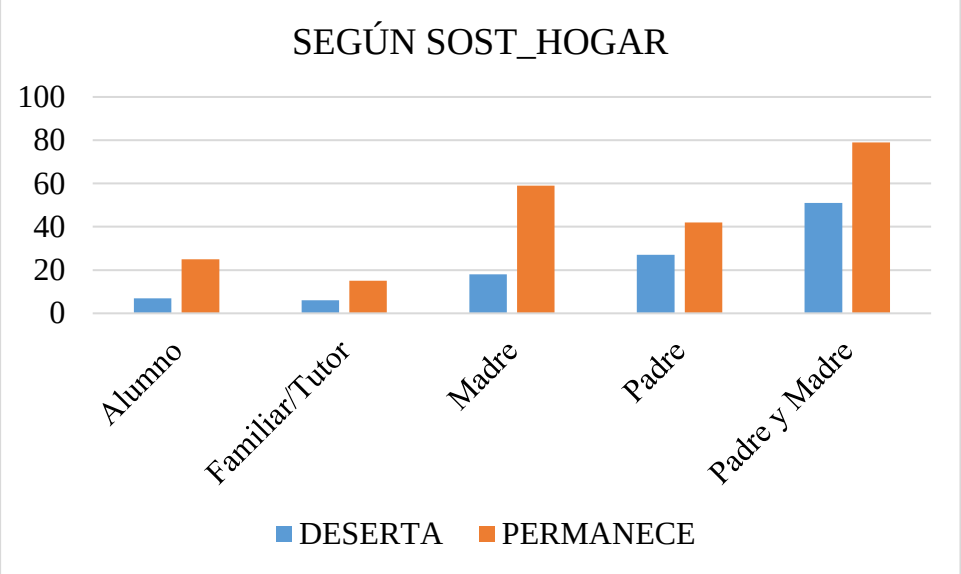


Figura 14. Según Sostiene el Hogar

En la Tabla 32, la deserción, presenta la condición ocupación donde predomina el rubro dependiente con 66%, seguido de independiente con 28%, es decir que trabaja en algún rubro del sector estatal o empresa privada.

Tabla 32

Deserción según COND_OCUP

COND_OCUP	Frecuencia	Porcentaje	Porcentaje Acumulado
Ambos	1	1%	1%
Dependiente	72	66%	67%
Desocupado	5	5%	72%
Independiente	31	28%	100%
Total	109	100%	

En la Tabla 33, la permanencia, presenta la condición ocupación donde predomina el rubro dependiente con 55%, seguido de independiente con 35%, es decir que trabaja en algún rubro del sector estatal o empresa privada, desocupado en menor proporción 3%.

Tabla 33

Permanencia según COND_OCUP

COND_OCUP	Frecuencia	Porcentaje	Porcentaje Acumulado
Ambos	17	8%	8%
Dependiente	120	55%	62%
Desocupado	6	3%	65%
Independiente	77	35%	100%
Total	220	100%	

La Figura 15 grafica e ilustra la deserción y permanencia según la condición u ocupación de los padres.

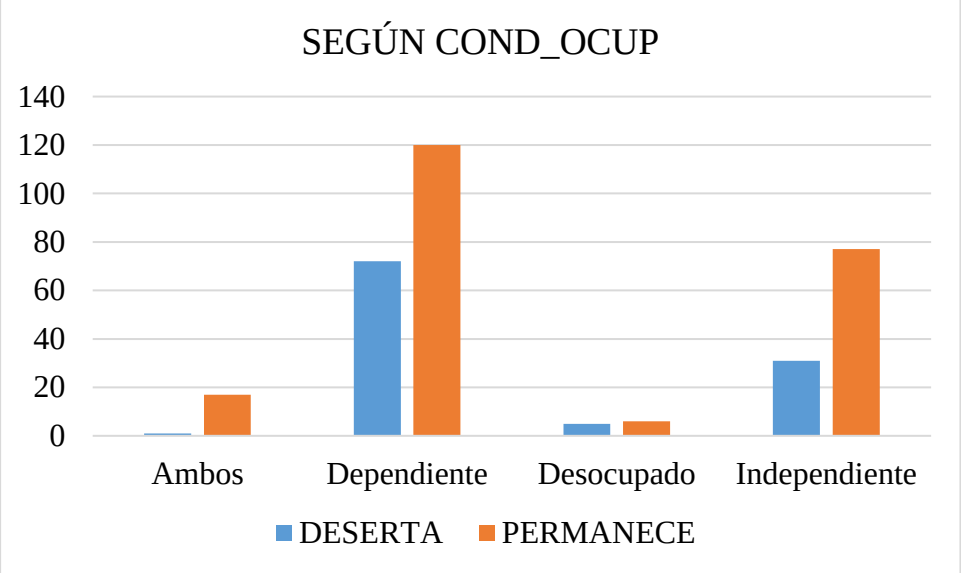


Figura 15. Según Condición Ocupacional

Respecto a la modalidad de ingreso económico M_ING_ECO, la Tabla 34 muestra que la tasa que predomina es mensual con 81%, seguido de la modalidad Semanal con 13% y la diferencia mínima 5% y 2% en diario y quincenal respectivamente.

Tabla 34

Deserción según Modalidad de ingreso económico

M_ING_ECO	Frecuencia	Porcentaje	Porcentaje Acumulado
Diario	5	5%	5%
Mensual	88	81%	85%
Quincenal	2	2%	87%
Semanal	14	13%	100%
Total	109	100%	

La Tabla 35, la permanencia, según modalidad de ingreso económico muestra que la tasa que predomina es mensual con 74%, seguido de la modalidad Semanal con 13% y la diferencia mínima 9% y 5% en diario y quincenal respectivamente.

Tabla 35

Permanencia según M_ING_ECO

M_ING_ECO	Frecuencia	Porcentaje	Porcentaje Acumulado
Diario	20	9%	9%
Mensual	162	74%	83%
Quincenal	10	5%	87%
Semanal	28	13%	100%
Total	220	100%	

La Figura 16 ilustra que la modalidad de ingreso económico en ambos deserta y permanece predomina la modalidad mensual.

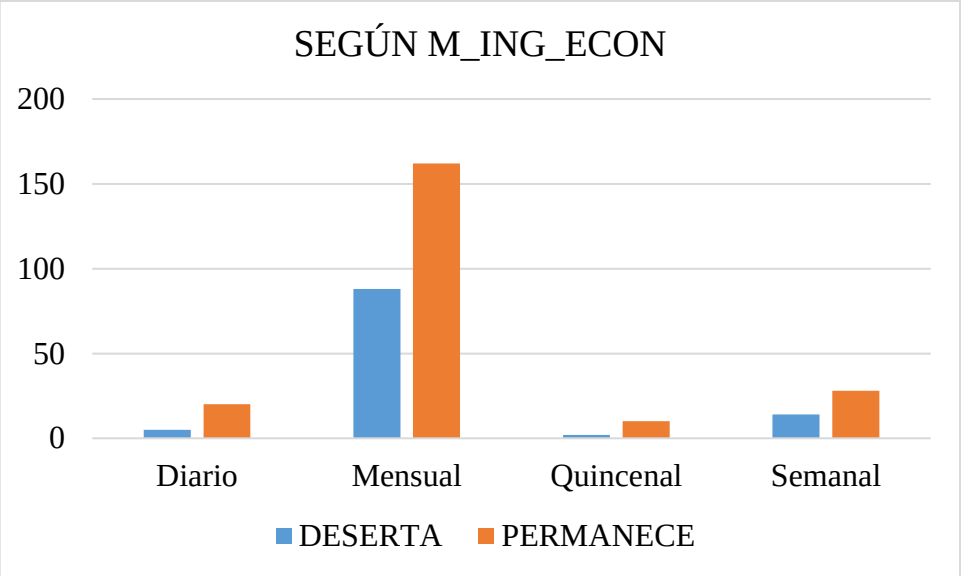


Figura 16. Gráfico deserción según modalidad de ingreso.

Según la Tabla 36, sobre la condición laboral del estudiante C_LAB_EST observamos que el 77% de los mismos no Trabaja frente a un 25% que si lo hace.

Tabla 36

Deserción según la condición laboral del estudiante.

C_LAB_EST	Frecuencia	Porcentaje	Porcentaje Acumulado
No trabaja	84	77%	77%
Trabaja	25	23%	100%
Total	109	100%	

Según la Tabla 37, permanencia, sobre la condición laboral del estudiante C_LAB_EST observamos que el 70% de los mismos no Trabaja frente a un 30% que si lo hace.

Tabla 37

Permanencia según C_LAB_EST

C_LAB_EST	Frecuencia	Porcentaje	Porcentaje Acumulado
No trabaja	154	70%	70%
Trabaja	66	30%	100%
Total	220	100%	

En la Figura 17 se observa que predomina No trabaja en ambas deserta y permanece.

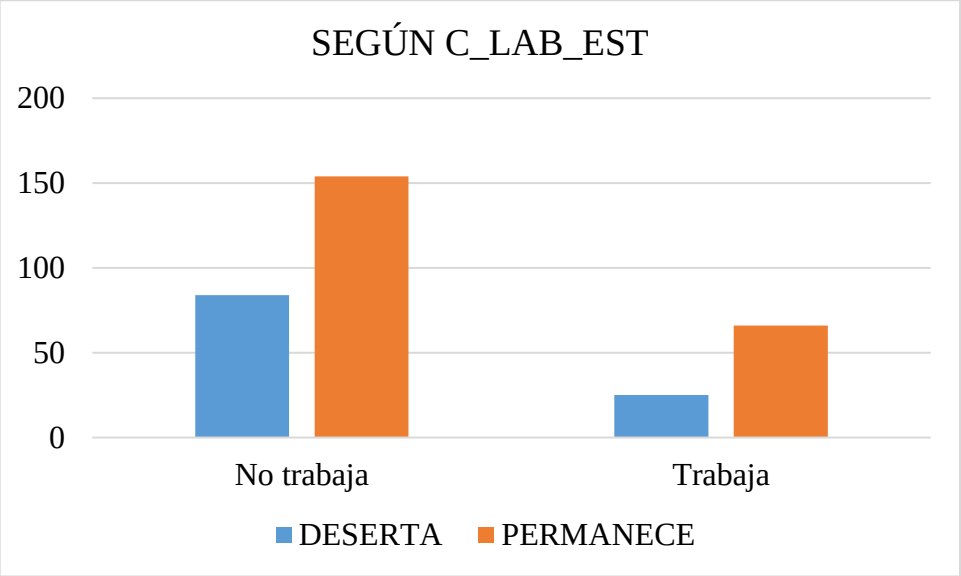


Figura 17. Gráfico deserción según la condición laboral del estudiante.

La Tabla 38, deserción, muestra que la vivienda del estudiante en un 58% que vive con sus padres seguido del 16% que vive con uno de sus padres, luego el rubro Cuarto alquilado con 12% y alojado con 11% y el resto 4% como cuidante es mínimo.

Tabla 38

Deserción según VIVIENDA_EST

VIVIENDA_EST	Frecuencia	Porcentaje	Porcentaje Acumulado
Como alojado	12	11%	11%
Como cuidante	4	4%	15%
Con sus padres	63	58%	73%
Con uno de sus padres	17	16%	89%
Cuarto alquilado	13	12%	100%
Total	109	100%	

La Tabla 39, permanencia, muestra que la vivienda del estudiante en un 43% que vive con sus padres seguido del 25% que vive con uno de sus padres, luego el rubro Cuarto alquilado con 19% y alojado con 11% y el resto 2% como cuidante es mínimo.

Tabla 39

Permanencia según VIVIENDA_EST

VIVIENDA_EST	Frecuencia	Porcentaje	Porcentaje Acumulado
Como alojado	25	11%	11%
Como cuidante	4	2%	13%
Con sus padres	95	43%	56%
Con uno de sus padres	55	25%	81%
Cuarto alquilado	41	19%	100%
Total	220	100%	

La Figura 18, representa la información sobre la vivienda tanto del que deserta como del que permanece de manera gráfica.

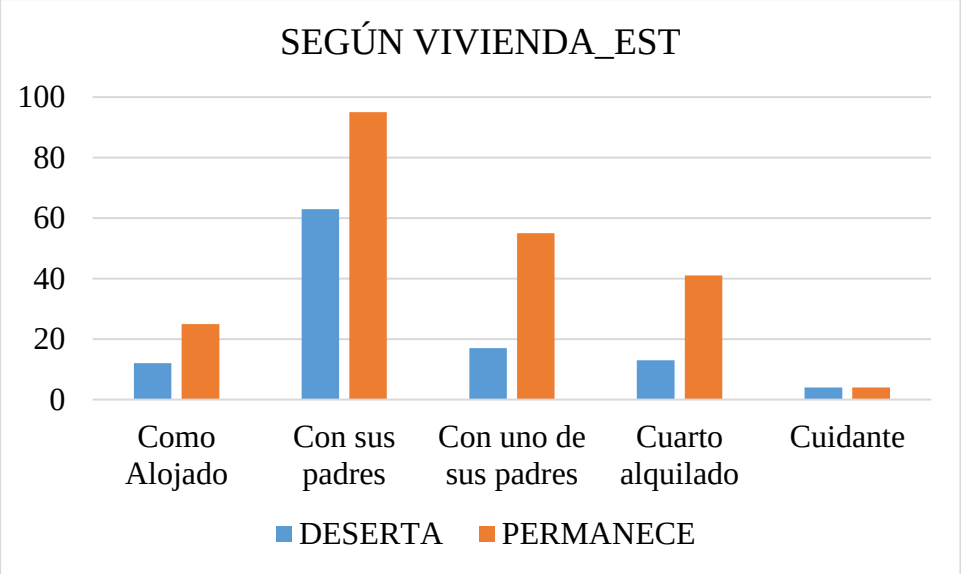


Figura 18. Según Vivienda del Estudiante

La Tabla 40, presenta los índices respecto al rubro dependencia económica liderando con 53% que indica que los estudiantes dependen económicamente de ambos padres, y que los que indican con uno de los padres es de 16% los que dependen de su madre y 15% dependen del padre respectivamente y con un familiar el 6% y los que viven solos con un 11%.

Tabla 40

Deserción según Dependencia Económica

DEP_ECON	Frecuencia	Porcentaje	Porcentaje Acumulado
Ambos padres	58	53%	53%
Sólo Madre	17	16%	69%
Sólo Padre	16	15%	83%
Un familiar	6	6%	89%
Vive solo	12	11%	100%
Total	109	100%	

La Tabla 41, permanencia según dependencia económica liderando con 43% que indica que los estudiantes dependen económicamente de ambos padres, y que los que indican con uno de los padres es de 22% los que dependen de su madre y 14% dependen del padre respectivamente y con un familiar el 4% y los que viven solos con un 18%.

Tabla 41

Permanencia según DEP_ECON

DEP_ECON	Frecuencia	Porcentaje	Porcentaje Acumulado
Ambos padres	94	43%	43%
Sólo Madre	48	22%	65%
Sólo Padre	30	14%	78%
Un familiar	8	4%	82%
Vive solo	40	18%	100%
Total	220	100%	

La Figura 19, a continuación, permite una visualización gráfica de las tablas anteriores destacando que la mayor parte de estudiantes dependen económicamente de ambos padres.

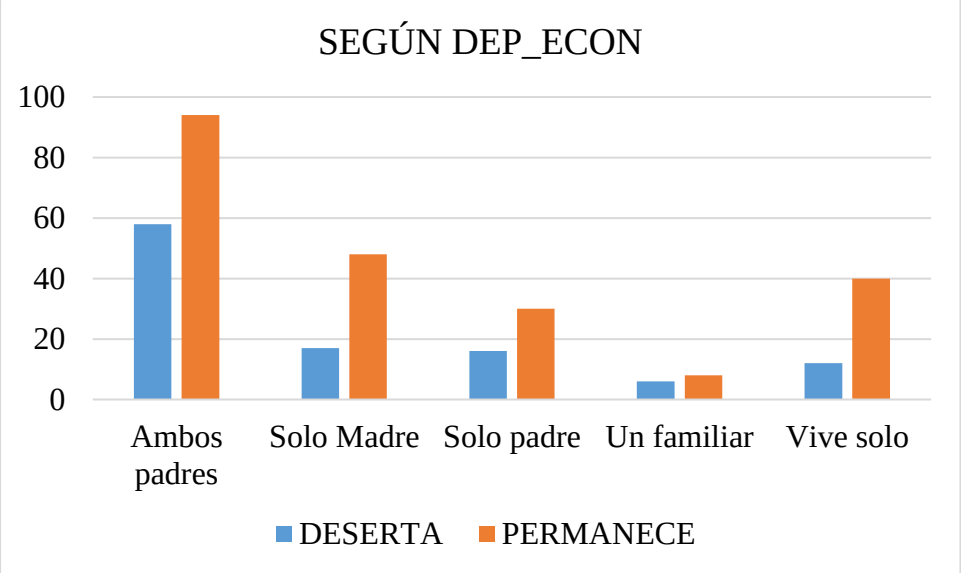


Figura 19. Según Dependencia Económica

Según la Tabla 42, en su mayoría los estudiantes que desertan son hijos de padres vivos en un 83% destacando que hay un 9% que es huérfano de padre y un 3% mínimo que vive sólo.

Tabla 42

Deserción según RIESG_FAM

RIESG_FAM	Frecuencia	Porcentaje	Porcentaje Acumulado
Hijo de padres vivos	91	83%	83%
Huérfano de madre	3	3%	86%
Huérfano de padre.	10	9%	95%

Huérfano de padre y madre	2	2%	97%
Vive solo	3	3%	100%
Total	109	100%	

En la Tabla 43, en su mayoría los estudiantes que permanecen son hijos de padres vivos en un 78% destacando que hay un 9% que vive sólo

Tabla 43

Permanencia según RIESG_FAM

RIESG_FAM	Frecuencia	Porcentaje	Porcentaje Acumulado
Hijo de padres vivos	172	78%	78%
Huérfano de madre	16	7%	85%
Huérfano de padre.	10	5%	90%
Huérfano de padre y madre	3	1%	91%
Vive solo	19	9%	100%
Total	220	100%	

La Figura 20, muestra que la mayor parte de los alumnos son hijos de padres vivos.

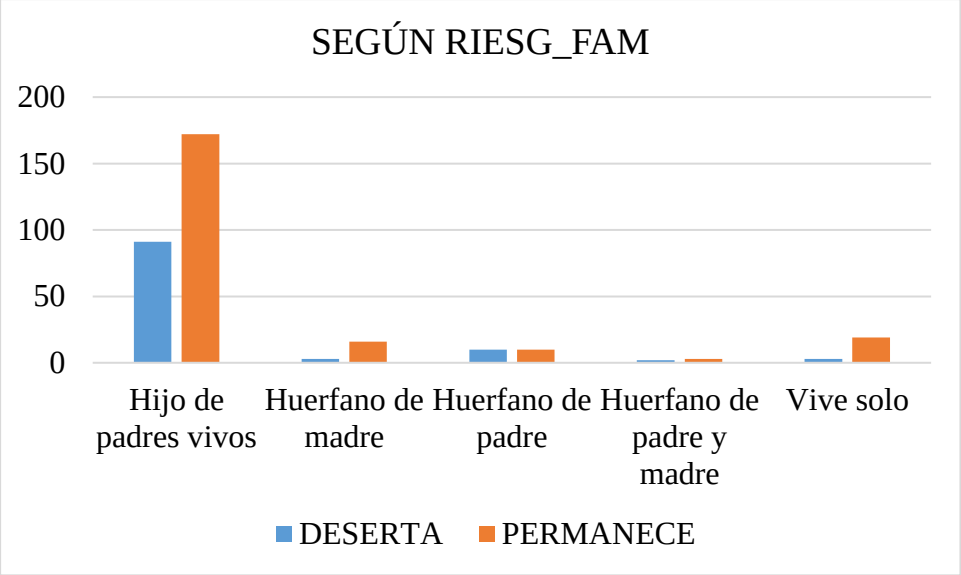


Figura 20. Gráfico según RIESG_FAM

La Tabla 44, sobre deserción en tenencia de vivienda muestra que 81% tiene vivienda propia, 11% vive en alquiler y cantidad mínima en Otro e Invasión con porcentajes de 7% y 1% respectivamente.

Tabla 44

Resumen deserción según la tenencia de Vivienda.

TENEN_VIV	Frecuencia	Porcentaje	Porcentaje Acumulado
Alquilada	12	11%	11%
Invasión	1	1%	12%
Otro	8	7%	19%
Propia	88	81%	100%
Total	109	100%	

La Tabla 45, sobre permanencia en tenencia de vivienda muestra que 65% tiene vivienda propia, 21% vive en alquiler y cantidad mínima en Otro e Invasión con porcentajes de 8% y 5% respectivamente

Tabla 45

Permanencia según TENEN_VIV

TENEN_VIV	Frecuencia	Porcentaje	Porcentaje Acumulado
Alquilada	46	21%	21%
Invasión	12	5%	26%
Otro	18	8%	35%
Propia	144	65%	100%
Total	220	100%	

La Figura 21, muestra que en su mayoría los estudiantes cuentan con una casa propia y en menor proporción alumnos viven en casa alquilada.

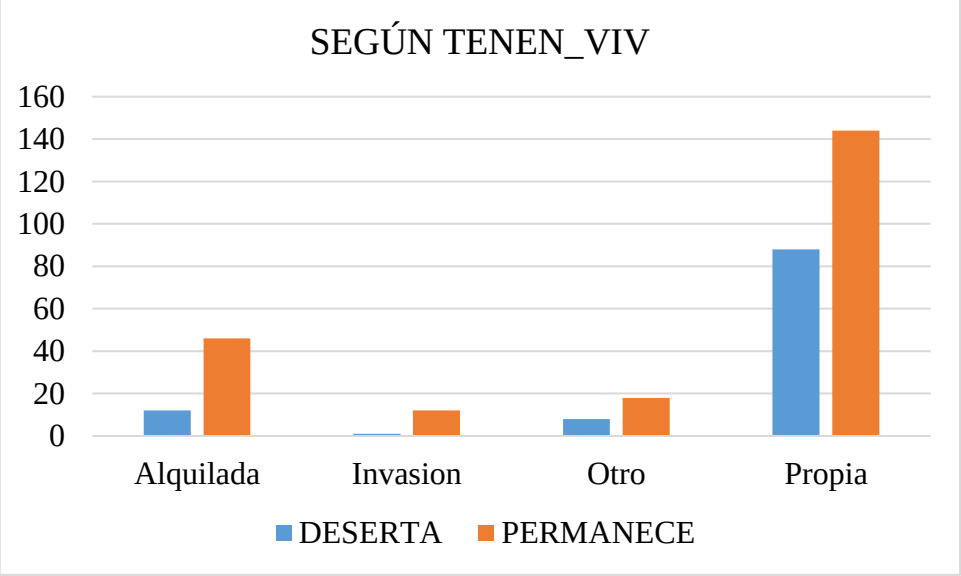


Figura 21. Deserción según tenencia de Vivienda.

En la Tabla 46, se resume la información sobre el tipo de construcción para estudiantes que desertan la mayor parte 67% tiene su casa de material Noble, de material rústico 17% y 16% tiene su casa de material Mixto.

Tabla 46

Resumen deserción según el tipo de construcción.

TIPO_CONSTR	Frecuencia	Porcentaje	Porcentaje Acumulado
Mixto	17	16%	16%
Noble	73	67%	83%
Rústico	19	17%	100%
Total	109	100%	

En la Tabla 47, se resume la información sobre permanencia según el tipo de construcción para estudiante la mayor parte 48% tiene su casa de material Noble, de material rústico 20% y 25% tiene su casa de material Mixto y en construcción precaria un 8%.

Tabla 47

Permanencia según TIPO_CONSTR

TIPO_CONSTR	Frecuencia	Porcentaje	Porcentaje Acumulado
Mixto	55	25%	25%
Noble	105	48%	73%
Precario	17	8%	80%
Rústico	43	20%	100%
Total	109	100%	

La Figura 22, muestra que la mayor parte de estudiantes tiene su vivienda de material Noble y en menor proporción de material mixto.

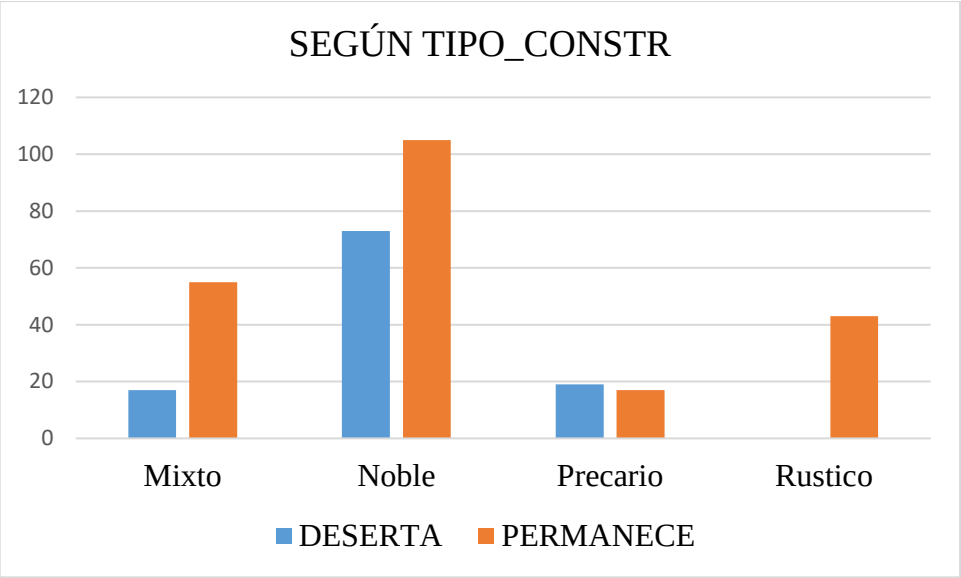


Figura 22. Gráfico deserción según tipo de construcción.

La Tabla 48, la información de estudiantes que desertan sobre el tipo de vivienda muestra que 87% vive independientemente, seguido de 10% en el rubro de Otro y mínimamente 1% y 2% en Conventillo y Departamento respectivamente.

Tabla 48

Resumen deserción según tipo de vivienda.

TIPO_VIV	Frecuencia	Porcentaje	Porcentaje Acumulado
Conventillo	1	1%	1%
Departamento	2	2%	3%
Independiente	95	87%	90%
Otro	11	10%	100%
Total	109	100%	

De manera similar, la Tabla 49, sobre los que permanecen según el tipo de vivienda muestra que 77% vive independientemente, seguido de 11% en el rubro de Otro y mínimamente 10% y 2% en Conventillo y Departamento respectivamente.

Tabla 49

Permanencia según TIPO_VIV

TIPO_VIV	Frecuencia	Porcentaje	Porcentaje Acumulado
Conventillo	21	10%	10%
Departamento	5	2%	12%
Independiente	170	77%	89%

Otro	24	11%	100%
Total	220	100%	

La Figura 23, indica que en mayor proporción los estudiantes viven de manera independiente.

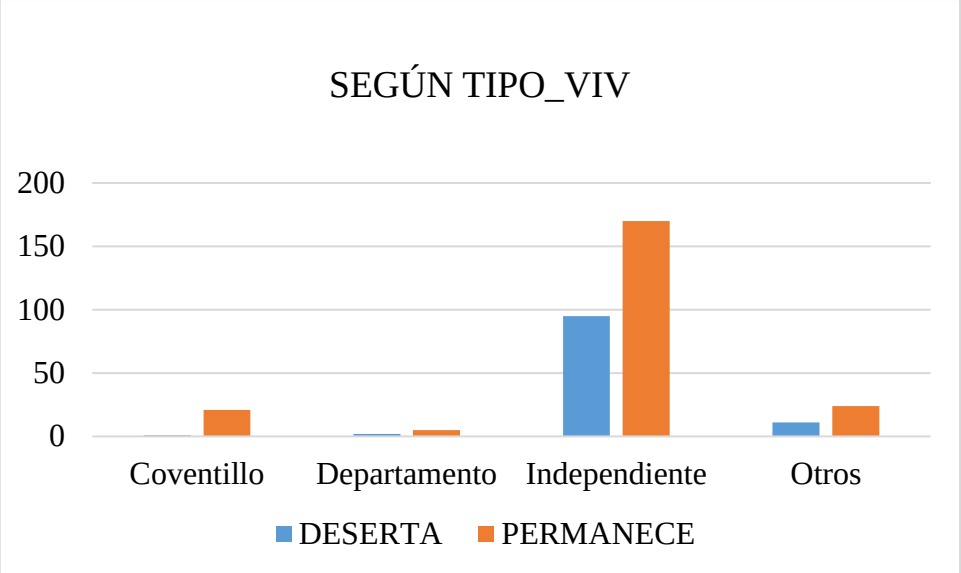


Figura 23. Gráfico deserción según el tipo de vivienda.

La Tabla 50, contiene los estadísticos descriptivos de las categorías COMP_HOGAR, TOTAL_INGRE, CARGA_FAM, HIJOS_SUP, N_PISOS y N_DORM variable numéricas.

Tabla 50

Estadística descriptiva otros

	COMP_HOGAR	TOTAL_INGRE	CARGA_FAM	HIJOS_SUP	N_PISOS	N_DORM
Media	4.54	1564.17	3.5	1.38	1.39	2.98
Error típico	0.14	92.06	0.12	0.07	0.06	0.15
Mediana	4	1491	3	1	1	3
Moda	4	1491	3	1	1	2
Desviación estándar	1.42	961.12	1.22	0.69	0.67	1.59
Varianza	2.01	923742.6	1.49	0.48	0.44	2.52
Curtosis	1.9	1.4	7.08	7.78	7.2	13.47
Asimetría	1.12	1.15	1.58	2.43	2.21	2.45
Rango	8	4690	9	4	4	12

Mínimo	2	110	1	1	1	1
Máximo	10	4800	10	5	5	13
Suma	495	170494.58	381	150	152	325
Cuenta	109	109	109	109	109	109

4.4 Análisis comparativo de los modelos de Clasificación

- Regresión Logística
- Máquina de Vector Soporte
- Arboles de Decisión
- Naive Bayes

El total de datos muestrales es 329 datos y 40 variables de los cuales 109 son desertores (=1) y 220 son no desertores (=0). Los datos faltantes se fueron completando con la media en el caso de la edad y el total de ingresos, la moda en los casos de variables cualitativas. La variable total de ingresos se escaló a unidades de miles y la variable edad se ajustó a una distribución normal junto con las otras variables numéricas.

Para hacer el entrenamiento de consideraron 70% (230 datos) de los datos y 30% (99 datos) para prueba.

4.4.1 Análisis del Rendimiento

Se hicieron pruebas con 40 y 94 atributos o características y luego se aplicó el primer método para reducción de atributos, el método Ramdon forest (contribución mayor o igual al 2%) resultando 12 y 9 atributos importantes respectivamente, se procede con la prueba respectiva y seguidamente aplicamos el Segundo método featurewiz para extraer de manera automática atributos resultando 19 y 24 atributos de importancia respectivamente. Aplicando la tercera prueba.

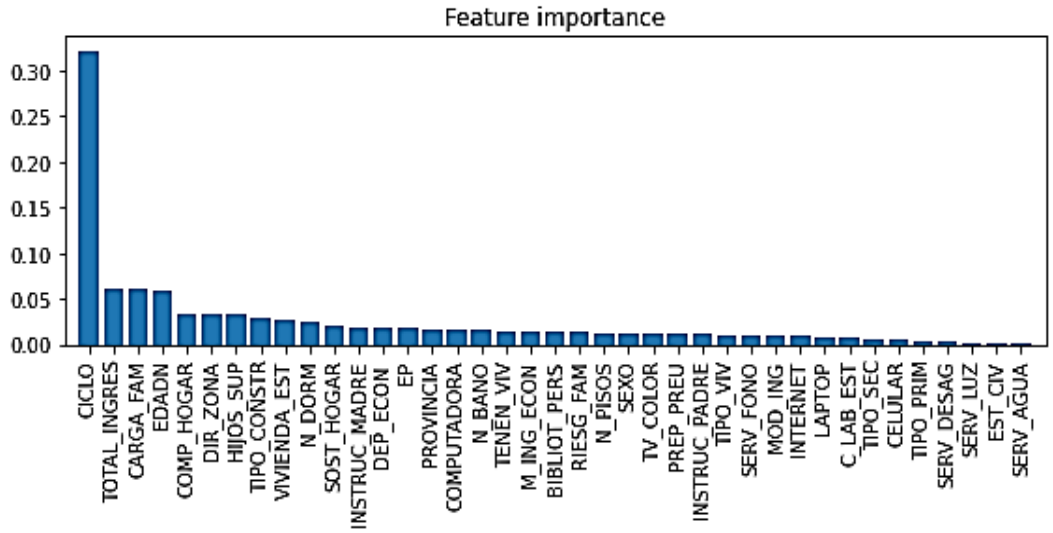


Figura 24. Características según su contribución, técnica Random Forest.

La Tabla 51 muestra las 12 características (11 características y 1 dependiente la Deserción) importantes según la técnica de Random Forest aplicado a los 40 datos los cuales y según su orden corresponden a los siguientes aspectos de la ficha socioeconómica del estudiante:

DATOS GENERALES: CICLO, EDAD, DIR_ZONA son 3 características.
ASPECTO ECONOMICO: TOTAL_INGRE, CARGA_FAM, COMP_HOGAR, VIVIENDA_EST, SOS_HOGAR, HIJOS_SUP. Son 6 características.
ASPECTO VIVIENDA: N_DORM, TIPO_CONSTR. Son 2 características.

Tabla 51

Características aporte mayor o igual al 2% según RF.

	Característica	**%**
1.	CICLO	32%
2.	TOTAL_INGRES	6%
3.	CARGA_FAM	6%
4.	EDADN	6%
5.	COMP_HOGAR	3%
6.	DIR_ZONA	3%
7.	HIJOS_SUP	3%
8.	TIPO_CONSTR	3%
9.	VIVIENDA_EST	3%
10.	N_DORM	2%
11.	SOST_HOGAR	2%

La Figura 25. Presenta las características según la contribución de ellas según la técnica Random Forest.

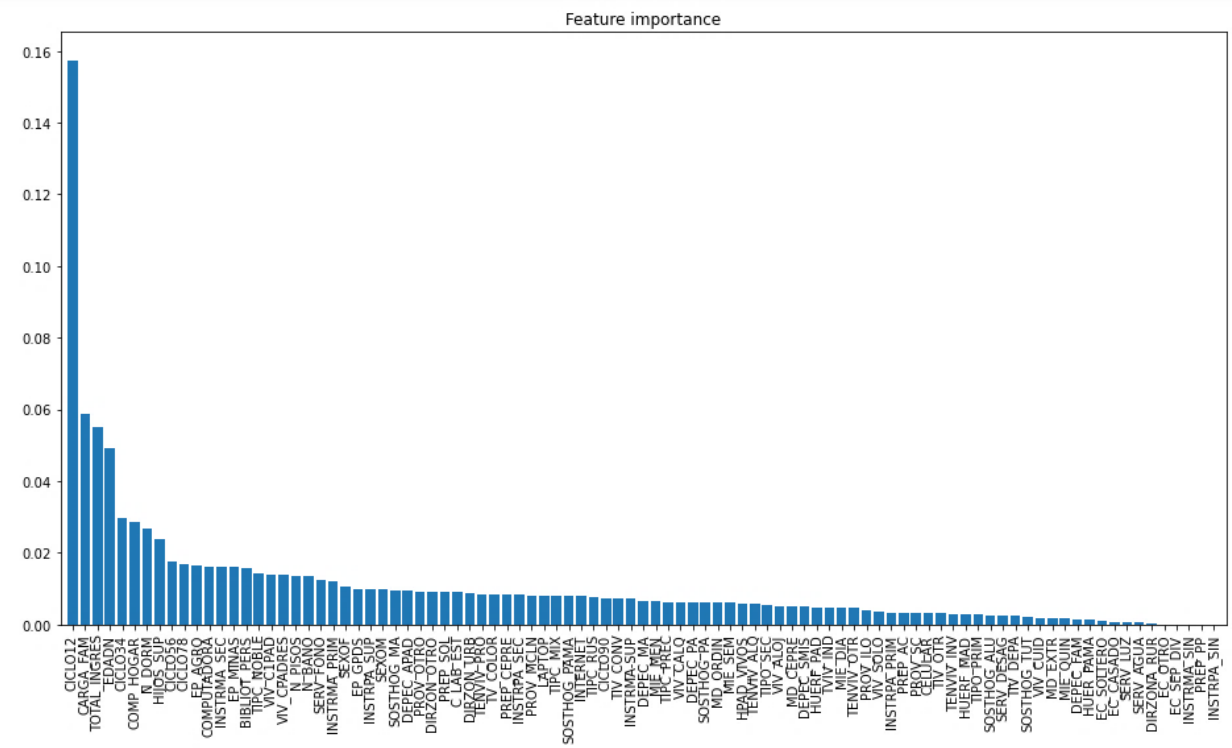


Figura 25. Características y contribución según la técnica de Random Forest.

En la Tabla 52 se resumen según la importancia las 9 (8 variables explicativas y 1 dependiente) características seleccionadas usando la técnica de Random Forest aplicado a 94 atributos agregando variables dummy las cuales y según su orden corresponden a los siguientes aspectos de la ficha socioeconómica del estudiante:

DATOS GENERALES: CICLO12, CICLO34, EDAD_N. son 3 características.

ASPECTO ECONOMICO: TOTAL_INGRE, CARGA_FAM, HIJOS_SUP, COMP_HOGAR. Son 4 características.

ASPECTO VIVIENDA: N_DORM. Son 1 características.

Tabla 52

Características aporte mayor o igual al 2% según Random Forest

	Característica	%
1.	CICLO12	16%
2.	CARGA_FAM	6%
3.	TOTAL_INGRES	6%
4.	EDADN	5%
5.	CICLO34	3%
6.	COMP_HOGAR	3%
7.	N_DORM	3%
8.	HIJOS_SUP	2%

La Tabla 53 muestra la reducción de características y que según Featurewiz se seleccionaron como características importantes luego de aplicar varios modelos de prueba de las 40 características se redujeron a 19 (18 variables explicativas y 1 dependiente) características importantes en los siguientes aspectos según la ficha socioeconómica del estudiante:

DATOS GENERALES: CICLO, EP, PROVINCIA. Son 3 característica.
ASPECTO ECONOMICO: CARGA_FAM, HIJOS_SUP, M_INGR_ECON. Son 3 características.
ASPECTO VIVIENDA: TIPO_VIV, TIPO_CONSTR, COMPUTADORA, TV_COLOR, TENEN_VIV, N_PISOS, N_DORM, LAPTOP, N_BANO, CELULAR, INTERNET, BIBLIO_PERS. Son 12 características.

Tabla 53

Reducción a 18 características importantes. Featurewiz.

	Característica
1.	CICLO
2.	HIJOS_SUP
3.	TIPO_VIV
4.	TV_COLOR
5.	EP
6.	PROVINCIA
7.	N_BANO
8.	TENEN_VIV
9.	CARGA_FAM
10.	COMPUTADORA
11.	INTERNET
12.	BIBLIOT_PERS
13.	TIPO_CONSTR
14.	M_ING_ECON
15.	CELULAR
16.	N_PISOS
17.	LAPTOP
18.	N_DORM

De similar manera al aplicarle dos veces la técnica featurewiz a los datos con variables dummy de 94 características selecciona como importantes a 24 características las cuales se presentan en la siguiente tabla y cuya evaluación en Featurewiz se presenta en el anexo. En la Tabla 54 se muestran las características

según su importancia las cuales corresponden a los siguientes aspectos según la ficha socioeconómica del estudiante:

DATOS GENERALES: CICLO12, EP_MINAS. Son 2 característica.
ASPECTO ECONOMICO: HIJOS_SUP, VIV_CUID, DEPEC_SIMIS, VIV_ALOJ, CARGA_FAM. Son 5 característica.
DEL ESTUDIANTE: HVSOL, HUERF_MAD, HUERF_PAD. Son 2 característica.
ASPECTO VIVIENDA: TIV_CONV, TIPC_PREC, TIPC_MIX, TIV_DEPA, TIV_OTR, TENVIV_INV, TEVIV_PRO, COMPUTADORA, INTERNET, TIPC_NOBLE, TV_COLOR, CELULAR, N_BANO, BIBLIOT_PERS. Son 14 características.

Tabla 54

Selección Featurewiz 23 atributos seleccionados importantes

Características			
1.	CICLO12	13.	TIPC_MIX
2.	TIPC_PREC	14.	TIV_DEPA
3.	TENVIV_INV	15.	CARGA_FAM
4.	HIJOS_SUP	16.	COMPUTADORA
5.	VIV_CUID	17.	INTERNET
6.	EP_MINAS	18.	TEVIV_PRO
7.	HUERF_MAD	19.	TIPC_NOBLE
8.	BIBLIOT_PERS	20.	N_BANO
9.	HUERF_PAD	21.	CELULAR
10.	DEPEC_SIMIS	22.	TV_COLOR
11.	TIV_CONV	23.	TIV_OTR
12.	VIV_ALOJ		

En la Tabla 55 se resumen la reducción aplicada y las características consideradas para modelar la deserción en cada uno de los métodos de clasificación. A los datos codificados con características dummy se le aplicaron dos reducciones obteniendo como numero de características importantes 24.

Tabla 55

Resumen aplicación de reducción de características.

	Datos Codificados	Datos codificados Dummy
Sin reducción de características	40	94
Random Forest	12	9
Featurewiz	19	24

Los métodos de clasificación usados fueron: Regresión Logística (RLC), Arboles Decisión (DTC), Máquina Vector Soporte (MVSC) y Naive Bayes (NBC) y para efectuar el análisis comparativo se emplearon las métricas de evaluación de Scikit-Learn, como matriz de confusión, el puntaje F1, la curva ROC y la validación cruzada con K=5 y K=10 Folds. Estos modelos de clasificación tienen parámetros por defecto, pero en algunos se pueden cambiar algunos como por ejemplo la opción Class-Weight fijada con parámetro “balanced” debido a que el conjunto de datos está desequilibrado el número de muestras positivas (109) es menor que el número de muestras negativas (220).

Luego, debido a la codificación de las variables el número de iteraciones, para el método que minimiza la función, se tuvo que aumentar a 600, ya que el valor por defecto es 100, además para la solución del problema de descenso se seleccionó solver = 'lbfgs', para regresión logística y usamos kernel = 'rbf' para machine vector support.

4.4.1.1 Regresión logística

Usamos el clasificador Regresión logística para 329 datos, 39 variables explicativas y 1 variable dependiente.

La Tabla 56 presenta los resultados de las métricas de regresión logística considerando 39 variables explicativas y 1 variable dependiente tiene una exactitud del 78% es buena debido a que la proporción entre las observaciones que fueron etiquetadas correctamente y número total de observaciones es alto, la precisión indica que la relación entre las observaciones etiquetadas como positivas entre el total de observaciones

etiquetadas como positivas es 58%, el recall o exhaustividad indica que las observaciones etiquetadas correctamente como positivas y el número total de observaciones que realmente fueron positivas es alto 97% en otras palabras el número total de desertores que fueron etiquetados correctamente como desertores es del orden del 97% y finalmente el valor F1 el promedio ponderado entre precisión y recall dado que las clases no están equilibradas es 72%.

Tabla 56

Métricas Regresión Logística 40 características

Métricas	
Accuracy	78%
Precision	58%
Recall	97%
F1	72%

Tabla 57

Matriz confusión Regresión Logística 40 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 48	Falsos Negativos 1
	1: Deserta	Falsos Positivos 21	Verdaderos Positivos 29

La Figura 26. Muestra el Área de la curva ROC para el clasificador Regresión logística y que tiene un valor alto y cercano a la unidad.

```
Sin entrenar: ROC AUC=0.500
Regresión Logística: ROC AUC=0.884
```

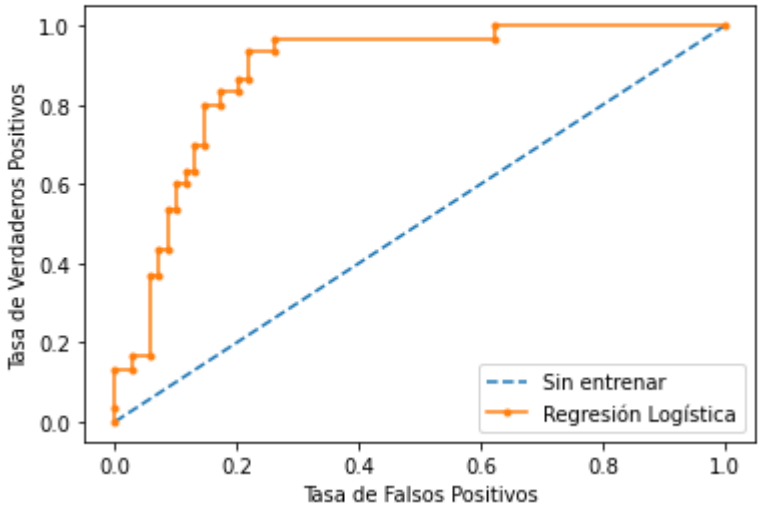


Figura 26. Curva ROC 40 características

Regresión logística usando 329 datos, 11 variables explicativas y 1 variable dependiente. (reducción Random Forest)

Tabla 58

Métrica Regresión logística reducción a 12 características según RF.

Métricas	
Accuracy	81%
Precision	62%
Recall	93%
F1	75%

Tabla 59

Matriz de confusión Regresión logística 12 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 52	Falsos Negativos 2
	1: Deserta	Falsos Positivos 17	Verdaderos Positivos 28

La Figura 27, es la curva ROC del clasificador Regresión logística cuyo valor es mejor cercano a uno.

```
Sin entrenar: ROC AUC=0.500
Regresión Logística: ROC AUC=0.907
```

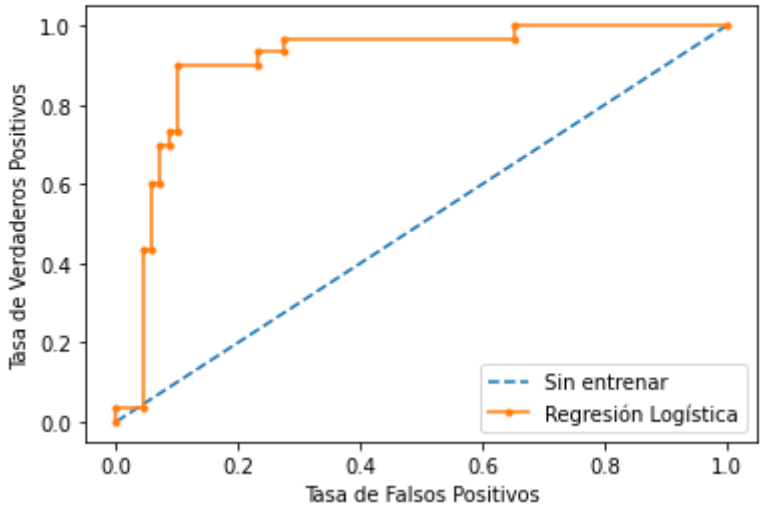


Figura 27. Curva ROC reducción a 12 características según RF

Regresión logística usando 329 datos, 18 variables explicativas y 1 variable dependiente. (reducción Featurewiz)

Tabla 60

Métricas Regresión logística reducción a 19 características.

Métricas	
Accuracy	79%
Precision	59%
Recall	97%
F1	73%

Tabla 61

Matriz de confusión Regresión Logística 19 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 49	Falsos Negativos 1
	1: Deserta	Falsos Positivos 20	Verdaderos Positivos

La Figura 28. Presenta la curva ROC para la reducción a 19 características según Featurewiz y que tiene un valor cercano a uno.

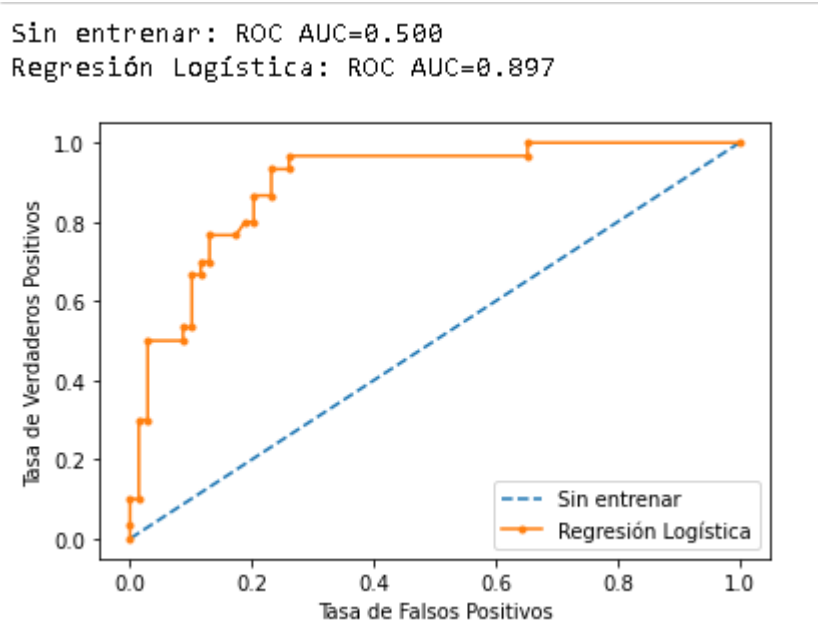


Figura 28. Curva ROC reducción 19 características.

Regresión logística con 329 datos, 93 variables explicativas y 1 variable dependiente.

Tabla 62

Métricas de Regresión logística 94 características

Métricas	
Accuracy	77%
Precision	59%
Recall	80%
F1	68%

Tabla 63

Matriz de confusión Regresión Logística 94 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 52	Falsos Negativos 6
	1: Deserta	Falsos Positivos 17	Verdaderos Positivos 24

Figura 29. Y el área de la curva ROC cuando el número de características es 94 usando variables dummy cuyo resultado también es aceptable puesto que es cercano a la unidad.

```
Sin entrenar: ROC AUC=0.500
Regresión Logística: ROC AUC=0.858
```

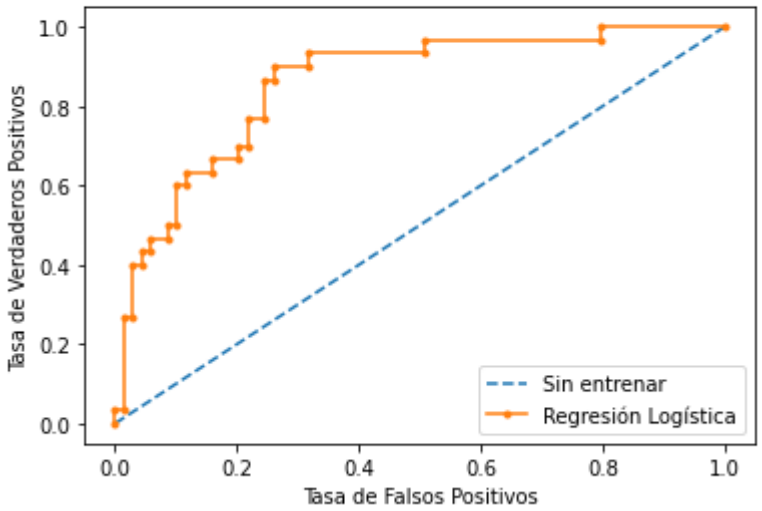


Figura 29. Curva ROC 94 características

Regresión logística usando 329 datos, 10 variables explicativas y 1 variable dependiente. (reducción Ramdon Forest)

Tabla 64

Métricas de Regresión logística reducción a 9 características.

Métricas	
Accuracy	76%
Precision	56%
Recall	97%
F1	71%

Tabla 65

Matriz de confusión Regresión logística 9 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 46	Falsos Negativos 1
	1: Deserta	Falsos Positivos 23	Verdaderos Positivos 29

La Figura 30. Muestra la curva ROC y su valor para cuando las características son 9 y el resultado es aceptable y cercano a 1.

```
Sin entrenar: ROC AUC=0.500
Regresión Logística: ROC AUC=0.832
```

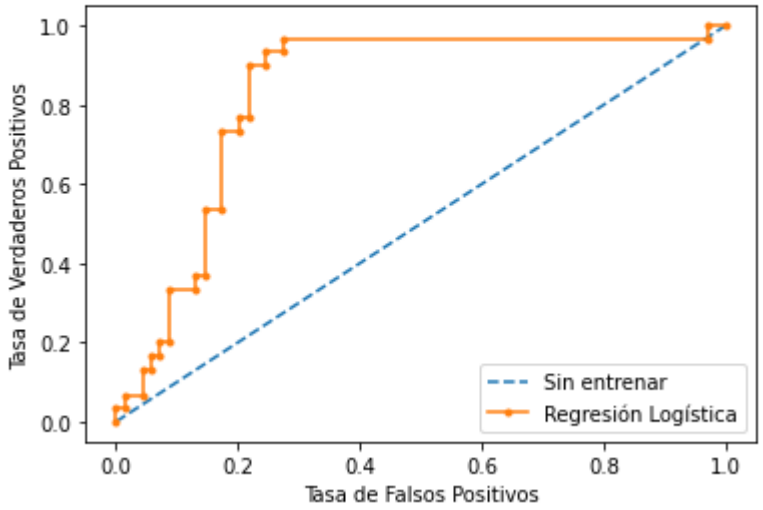


Figura 30. Curva ROC reducción a 9 características.

Regresión logística con 329 datos, 23 variables explicativas y 1 variable dependiente. (reducción Featurewiz)

Tabla 66

Métricas Regresión logística reducción a 24 características

Métricas	
Accuracy	81%
Precision	62%
Recall	93%
F1	75%

Tabla 67

Matriz de confusión Regresión Logística 24 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 52	Falsos Negativos 2
	1: Deserta	Falsos Positivos 17	Verdaderos Positivos 28

La Figura 31 se observa la curva ROC para la reducción a 24 características y cuyo valor es cercano a la unidad por lo que es aceptable.

```
Sin entrenar: ROC AUC=0.500
Regresión Logística: ROC AUC=0.884
```

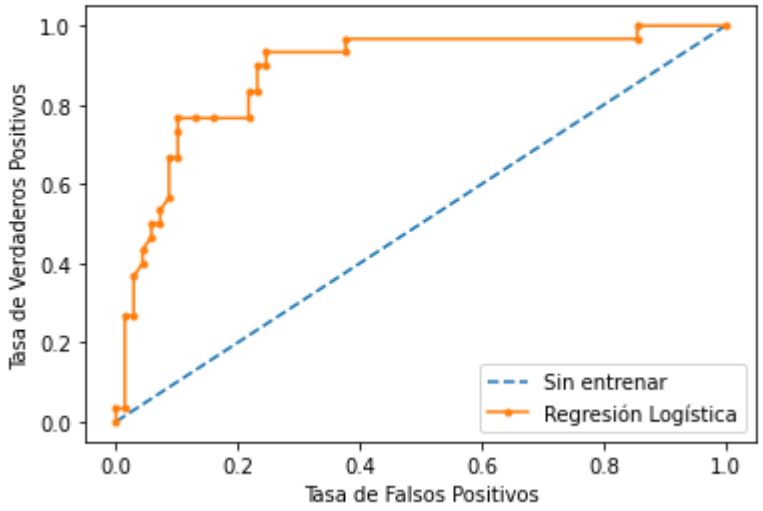


Figura 31. Curva ROC reducción a 24 características

Como se observa en las tablas 56 a 67 de métricas el clasificador Regresión Logística obtiene su mejor precisión cuando la reducción de características usando Random Forest se reduce de 40 características a 12 se logra una Precisión de 81%, es decir, que 81% de los alumnos han sido clasificados correctamente como desertores, también en la reducción de 94 usando Featurewiz a 24 se logra una Precisión de 81%, y el área de la curva ROC es de 0.884 y cuando tenemos la reducción a 12 características el área de la curva ROC es de 0.907 mayor que reducción a 24 características 0.884. Las características que corresponden al mejor modelo de Regresión logística, en orden de importancia, son:

CICLO, DIR_ZONA, EDAD_N. son 3 característica (Datos generales).
TOTAL_INGRES, CARGA_FAM, COMP_HOGAR, HIJOS_SUP, VIVIENDA_EST, SOST_HOGAR. Son 6 característica (Aspecto económico).
HVSOL. Son 1 característica (Del estudiante).
TIPO_CONSTR, N_DORM. Son 2 características (Aspecto de vivienda).
CICLO, EP, PROVINCIA. Son 3 característica. (Dato general)
M_ING_ECO, CARGA_FAM, HIJOS_SUP. Son 3 característica. (Aspecto económico)

TENEN_VIV, TIPO_VIV, TIPO_CONSTR, N_PISOS, N_BANO, TV_COLOR, CELULAR, LAPTOP, INTERNET, BIBLIO_PERS, COMPUTADORA, N_DORM, Son 12 características. (Aspecto de vivienda)
En el caso de 24 características las variables son:
EP_MINAS, CICLO12 son 2 características (Datos generales). VIV_ALOJ, VIV_CUID, CARGA_FAM, HIJOS_SUP. Son 4 características (Aspecto económico). DEPEC_SMIS, HUERF_PAD, HUERF_MAD. Son 3 características (Aspecto del estudiante). TEVIV_PRO, TEVIV_INV, TIPC_NOBLE, TIPC_MIX, TIPC_PREC, TIV_DEPA, TIV_CONV, TIPC_OTR, N_BANO, TV_COLOR, CELULAR, INTERNET, BIBLIO_PERS, COMPUTADORA. Son 14 características (Aspecto de vivienda).

A continuación, la Tabla 68 muestra los coeficientes del modelo Regresión logística con 12 características.

Tabla 68

Coeficientes del modelo 12 características.

	Característica	Valor
1	CICLO	[-1.379026019740131]
2	EDADN	[0.26559378540764284]
3	DIR_ZONA	[0.012698478266252835]
4	COMP_HOGAR	[0.26343924944998]
5	SOST_HOGAR	[-0.14792621430782002]
6	TOTAL_INGRES	[-0.34008911008565923]
7	VIVIENDA_EST	[-0.06848357445986113]
8	CARGA_FAM	[-0.102429192697214]
9	HIJOS_SUP	[0.46529554329550976]
10	TIPO_CONSTR	[-0.4334943549976145]
11	N_DORM	[0.20786119717323914]

También, la Tabla 69 muestra las 24 características y su coeficiente del modelo Regresión logística.

Tabla 69

Coeficientes del modelo 24 características.

	Característica	Valor
1	EP_MINAS	[0.374619633896757]
2	CICLO12	[2.91175473489585]
3	VIV_ALOJ	[0.49629325216819586]
4	VIV_CUID	[0.6657459933430719]
5	CARGA_FAM	[0.25046750407159357]
6	HIJOS_SUP	[0.42555272166732894]
7	DEPEC_SMIS	[0.19477392835451432]
8	HUERF_MAD	[-0.7874435665699228]
9	HUERF_PAD	[0.030585192423532876]
10	TENVIV_PRO	[0.4894106260548982]
11	TENVIV_INV	[-0.5943219834107047]
12	TIPC_NOBLE	[0.4375851237359467]
13	TIPC_MIX	[0.006937492333284856]
14	TIPC_PREC	[-0.916068096780322]
15	TIV_DEPA	[-0.47260174408044114]
16	TIV_CONV	[-1.1200279614921191]
17	TIV_OTR	[0.038156735073808616]
18	N_BANO	[-0.3226771003104284]
19	TV_COLOR	[0.302339489793135]
20	CELULAR	[0.23407102888985784]
21	INTERNET	[-0.6335045867006641]
22	BIBLIOT_PERS	[0.9969739325794424]
23	COMPUTADORA	[0.525414839110219]

4.4.1.2 Árboles de Decisión

Árboles de decisión usando 329 datos, 39 variables explicativas y 1 variable dependiente.

Tabla 70

Métricas de Árboles Decisión 40 características

Métricas	
Accuracy	79%
Precision	64%
Recall	70%
F1	67%

Tabla 71

Matriz de confusión Árboles Decisión 40 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 57	Falsos Negativos 9
	1: Deserta	Falsos Positivos 12	Verdaderos Positivos 21

```
Sin entrenar: ROC AUC=0.500
Arboles de Clasificacion: ROC AUC=0.787
```

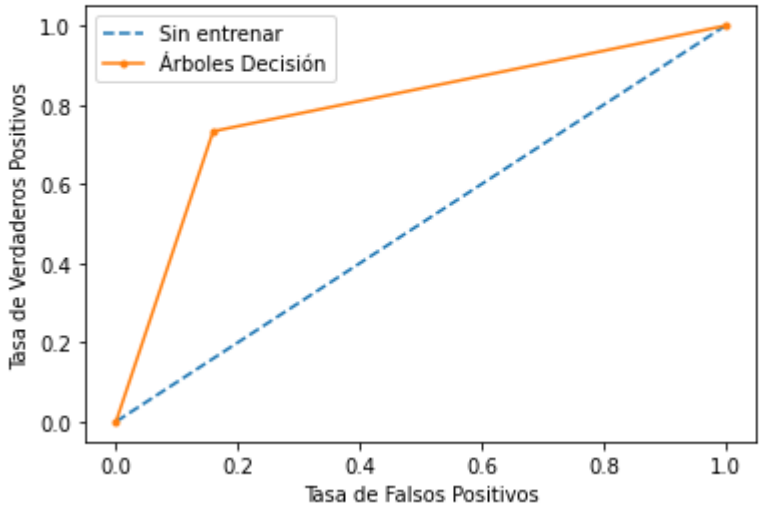


Figura 32. Curva ROC 40 características

Con 329 datos, 11 variables explicativas y 1 variable dependiente. (Random Forest)

Tabla 72

Métricas Árboles Decisión reducción a 12 características.

Métricas	
Accuracy	77%
Precision	61%
Recall	67%
F1	63%

Tabla 73

Matriz de confusión Árboles Decisión 12 características

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 56	Falsos Negativos 10
	1: Deserta	Falsos Positivos 13	Verdaderos Positivos 20

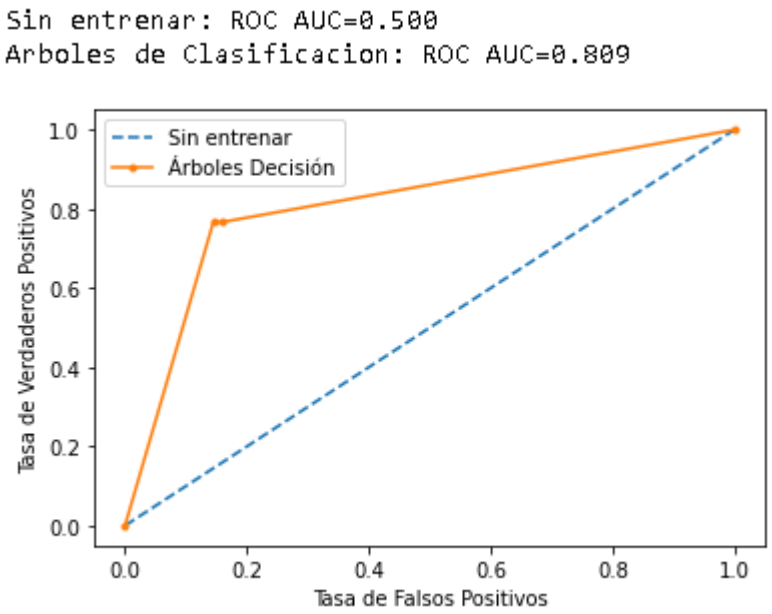


Figura 33. Curva ROC reducción a 12 características

Árboles de decisión con 329 datos, 18 variables explicativas y 1 variable dependiente. (Featurewiz)

Tabla 74

Métricas Árboles Decisión reducción 19 características.

Métricas	
Accuracy	76%
Precision	61%
Recall	72%
F1	66%

Tabla 75

Matriz de confusión Árboles Decisión 19 características

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 52	Falsos Negativos 9
	1: Deserta	Falsos Positivos 15	Verdaderos Positivos 23

```
Sin entrenar: ROC AUC=0.500
Arboles de Clasificacion: ROC AUC=0.744
```

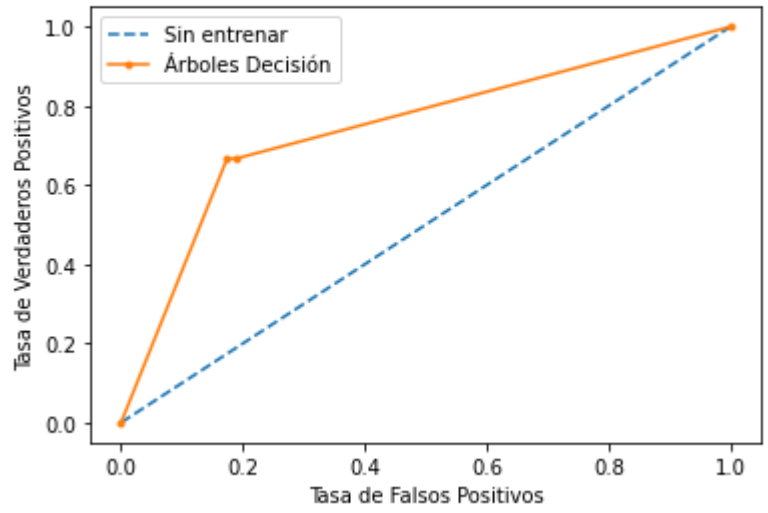


Figura 34. Curva ROC reducción a19 características.

Con 329 datos, 93 variables explicativas y 1 variable dependiente.

Tabla 76

Métricas Árboles Decisión 94 características.

Métricas	
Accuracy	70%
Precision	50%
Recall	63%
F1	56%

Tabla 77

Matriz de confusión Árboles Decisión 94 características

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 50	Falsos Negativos 11
	1: Deserta	Falsos Positivos 19	Verdaderos Positivos 19

La Figura 35 es la curva ROC para el clasificador Árbol de decisión 94 características

```
Sin entrenar: ROC AUC=0.500
Arboles de Clasificacion: ROC AUC=0.660
```

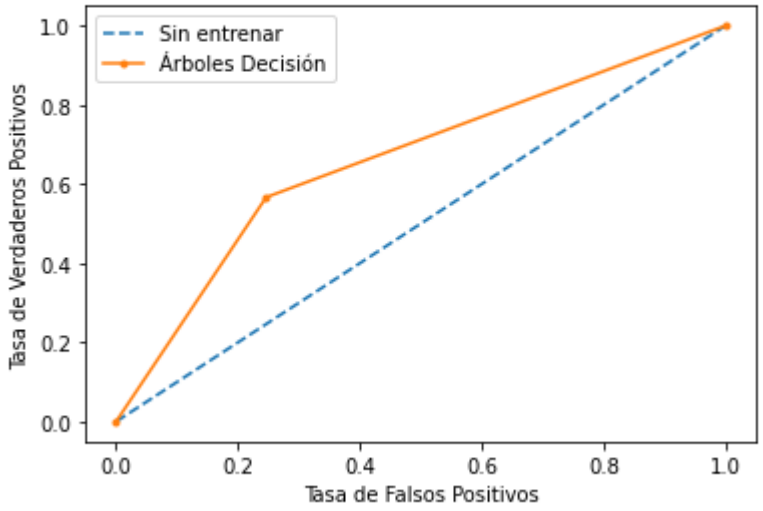


Figura 35. Curva ROC 94 características.

Árboles de decisión con 329 datos, 8 variables explicativas y 1 variable dependiente. (Ramdon Forest)

Tabla 78

Métricas Árboles Decisión reducción a 9 características

Métricas	
Accuracy	78%
Precision	61%
Recall	73%
F1	67%

Tabla 79

Matriz de confusión Árboles Decisión 9 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 55	Falsos Negativos 8
	1: Deserta	Falsos Positivos 14	Verdaderos Positivos 22

La curva ROC de Árboles de decisión para reducción de 94 a 9 características usando Random Forest.

```
Sin entrenar: ROC AUC=0.500
Arboles de Clasificacion: ROC AUC=0.749
```

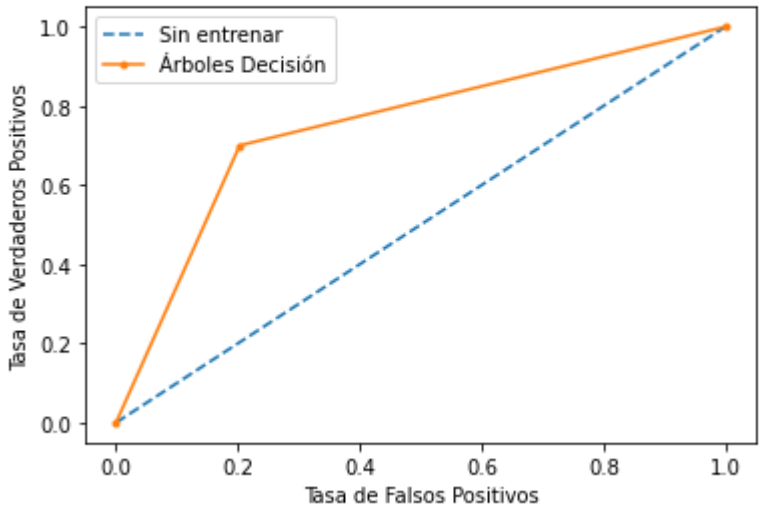


Figura 36. Curva ROC reducción a 9 características.

Árboles de decisión con 329 datos, 23 variables explicativas y 1 variable dependiente. (reducción Featurewiz)

Tabla 80

Métricas Árboles Decisión reducción a 24 características.

Métricas	
Accuracy	77%
Precision	60%
Recall	70%
F1	65%

Tabla 81

Matriz de confusión Árboles Decisión 24 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 55	Falsos Negativos 9

	Falsos Positivos	Verdaderos Positivos
1: Deserta	14	21

La Figura 37 presenta la curva ROC del clasificador Árboles Decisión reducción de 94 a 24 características usando Featurewiz.

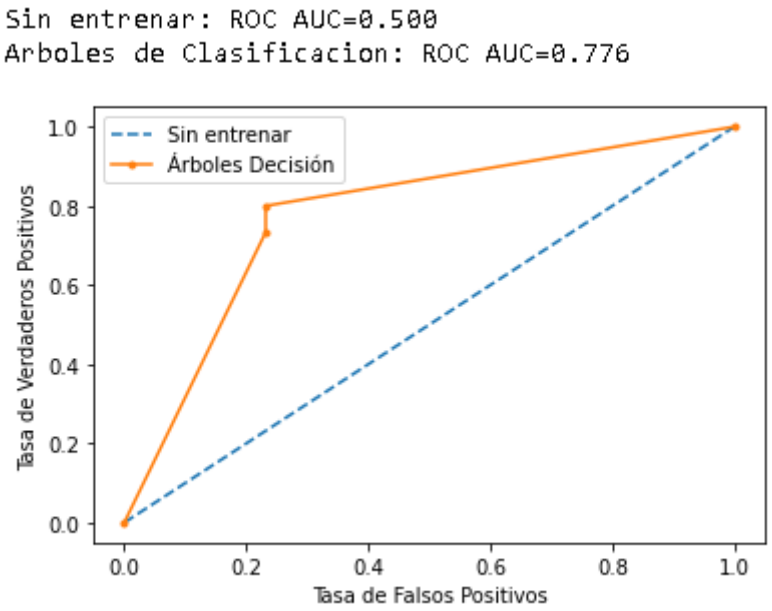


Figura 37. Curva ROC reducción a 24 características.

Como se observa en las tablas 70 a 81 de métricas el clasificador Árboles de Decisión obtiene su mejor precisión en los datos de 9 características se logra una Precisión de 78%, es decir, que 78% de los alumnos han sido clasificados correctamente como desertores, área de la curva ROC es de 0.749. Es de notar que cuando se aplica la técnica de reducción de características las métricas disminuyen levemente, tanto en la reducción por random forest y la técnica Featurewiz. Las características que corresponden al mejor Árbol de decisión son:

CICLO12, CICLO34, EDADN, 3 características en Datos generales. COMP_HOGAR, TOTAL_INGR, CARGA_FAM, HIJOS_SUP, 4 características en el Aspecto Económico. N_DORM, 1 variables en el aspecto de vivienda.

La Figura 38 muestra un recorte del Árbol de decisión y sus ramas, el árbol completo se ve en el ANEXO 2.

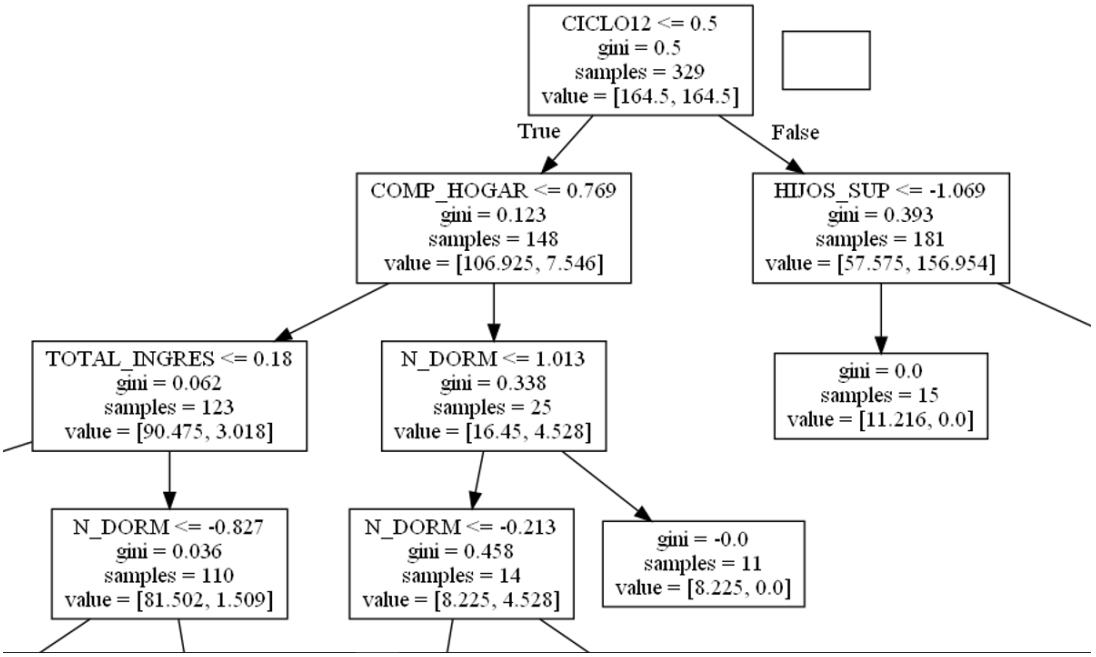


Figura 38. Árbol de decisión reducción a 9 características (Recorte del Árbol)

4.4.1.3 Máquina vector soporte

Máquina vector soporte usando 329 datos, 39 variables explicativas y 1 variable dependiente.

Tabla 82

Métricas Máquina vector soporte 40 características

Métricas	
Accuracy	69%
Precision	60%
Recall	55%
F1	58%

Tabla 83

Matriz de confusión Máquina vector soporte 40 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 47	Falsos Negativos 17
	1: Deserta	Falsos Positivos 14	Verdaderos Positivos

Matriz de confusión Máquina vector soporte 40 características.

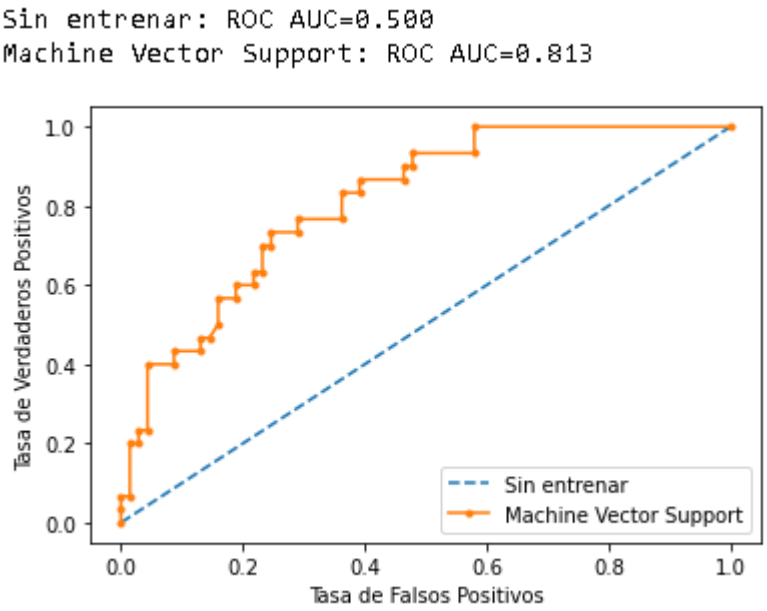


Figura 39. Curva ROC 40 características.

Máquina vector soporte con 329 datos, 11 variables explicativas y 1 variable dependiente. (reducción Random Forest)

Tabla 84

Métricas Máquina vector soporte reducción a 12 características.

Métricas	
Accuracy	74%
Precision	64%
Recall	71%
F1	67%

Tabla 85

Matriz de confusión Máquina vector soporte 12 características

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 46	Falsos Negativos 11
	1: Deserta	Falsos Positivos 15	Verdaderos Positivos 27

Matriz de confusión Máquina vector soporte reducción 12 características.

```
Sin entrenar: ROC AUC=0.500
Machine Vector Support: ROC AUC=0.876
```

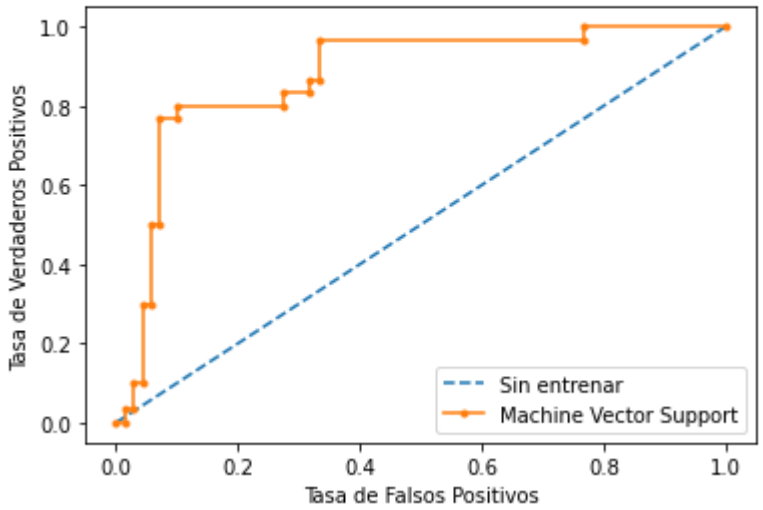


Figura 40. Curva ROC reducción a 12 características.

Máquina vector soporte con 329 datos, 18 variables explicativas y 1 variable dependiente. (reducción Featurewiz)

Tabla 86

Métricas reducción Máquina vector soporte reducción a 19 características

Métricas	
Accuracy	66%
Precision	55%
Recall	55%
F1	55%

Tabla 87

Matriz de confusión Máquina vector soporte 19 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 44	Falsos Negativos 17
	1: Deserta	Falsos Positivos 17	Verdaderos Positivos 21

```
Sin entrenar: ROC AUC=0.500
Machine Vector Support: ROC AUC=0.839
```

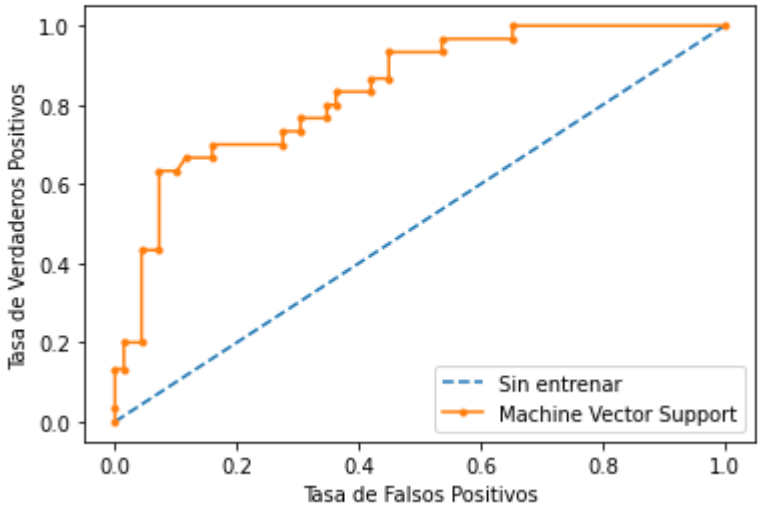


Figura 41. Curva ROC reducción a 19 características

Máquina vector soporte con 329 datos, 93 variables explicativas y 1 variable dependiente.

Tabla 88

Métricas Máquina vector soporte 94 características

Métricas	
Accuracy	71%
Precision	62%
Recall	61%
F1	61%

Tabla 89

Matriz de confusión Máquina vector soporte 94 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 47	Falsos Negativos 15
	1: Deserta	Falsos Positivos 14	Verdaderos Positivos 23

```
Sin entrenar: ROC AUC=0.500
Machine Vector Support: ROC AUC=0.817
```

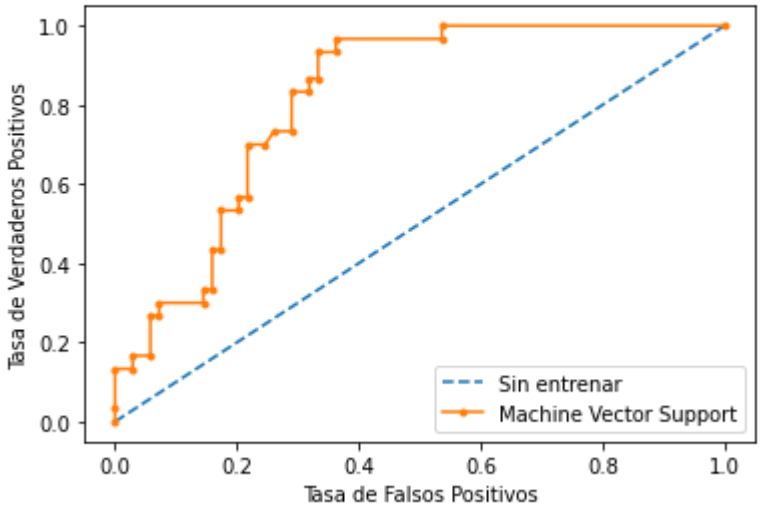


Figura 42. Curva ROC 94 características

Máquina vector soporte usando 329 datos, 8 variables explicativas y 1 variable dependiente. (reducción Ramdon Forest)

Tabla 90

Métricas Máquina vector soporte reducción a 9 características

Métricas	
Accuracy	76%
Precisión	65%
Recall	79%
F1	71%

Tabla 91

Matriz de confusión Máquina vector soporte 9 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 45	Falsos Negativos 8
	1: Deserta	Falsos Positivos 16	Verdaderos Positivos 30

```
Sin entrenar: ROC AUC=0.500
Machine Vector Support: ROC AUC=0.838
```

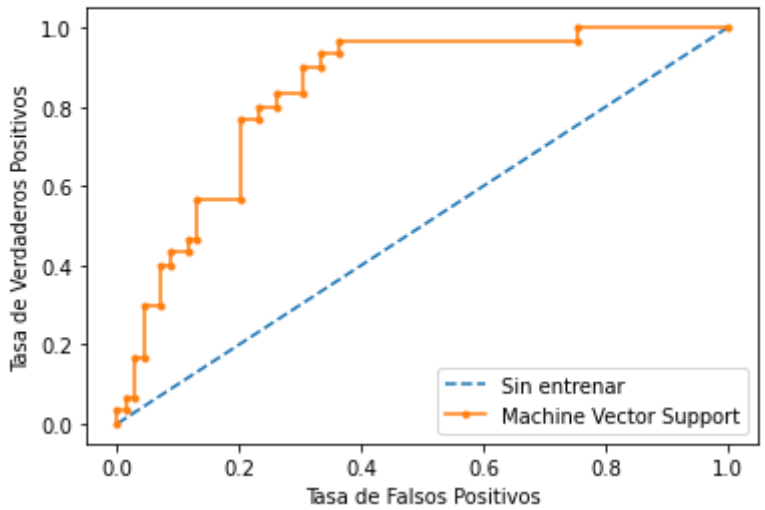


Figura 43. Curva ROC reducción a 9 características

Con 329 datos, 23 variables explicativas y 1 variable dependiente. (Featurewiz)

Tabla 92

Métricas Máquina vector soporte reducción a 24 características.

Métricas	
Accuracy	78%
Precisión	69%
Recall	76%
F1	72%

Tabla 93

Matriz de confusión Máquina vector soporte 24 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 48	Falsos Negativos 9
	1: Deserta	Falsos Positivos 13	Verdaderos Positivos 29

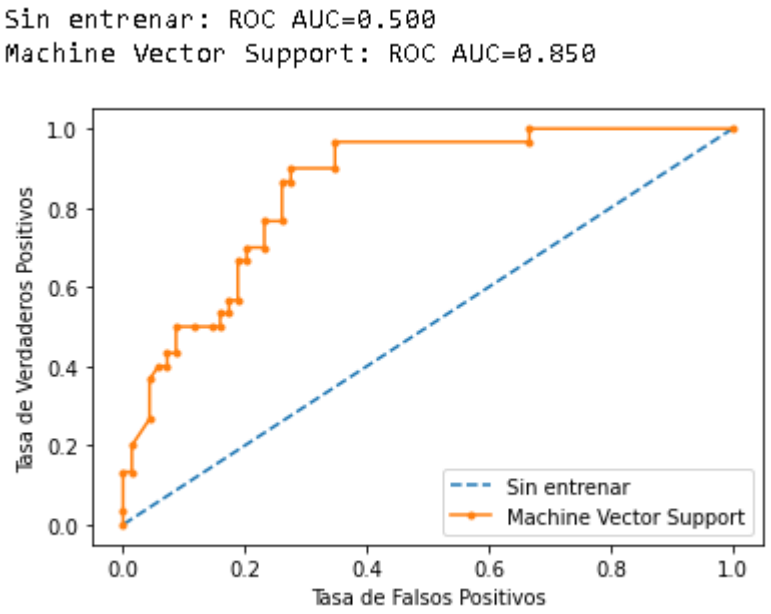


Figura 44. Curva ROC reducción a 24 características

Como se observa en las tablas 82 a 93 de métricas el clasificador Máquina vector soporte obtiene su mejor precisión cuando la reducción de características usando Featurewiz dos veces se reduce de 94 a 24 se logra una Precisión de 78%, es decir, que 78% de los alumnos han sido clasificados correctamente como desertores, área de la curva ROC es de 0.850. Las características que corresponden al mejor modelo de Máquina vector soporte, en orden de importancia, son:

EP_MINAS, CICLO12 son 2 características (Datos generales).

VIV_ALOJ, VIV_CUID, CARGA_FAM, HIJOS_SUP. Son 4 características (Aspecto económico).

DEPEC_SMIS, HUERF_PAD, HUERF_MAD. Son 3 características (Aspecto del estudiante)

TEVIV_PRO TEVIV_INV, TIPC_NOBLE, TIPC_MIX, TIPC_PREC, TIV_DEPA, TIV_CONV, TIPC_OTR, N_BANO, TV_COLOR, CELULAR, INTERNET, BIBLIO_PERS, COMPUTADORA. Son 14 características (Aspecto de vivienda).

4.4.1.4 Naive Bayes

Naive bayes con 329 datos, 39 variables explicativas y 1 variable dependiente.

Tabla 94

Métricas Naïve Bayes 40 características

Métricas	
Accuracy	62%
Precision	50%
Recall	68%
F1	58%

Tabla 95

Matriz de confusión Máquina vector soporte 40 características.

		Predichas	
		0: Permanece	1: Deserta
Reales	0: Permanece	Verdaderos Negativos 35	Falsos Negativos 12
	1: Deserta	Falsos Positivos 26	Verdaderos Positivos 26

.

```
Sin entrenar: ROC AUC=0.500
Naive Bayes: ROC AUC=0.830
```

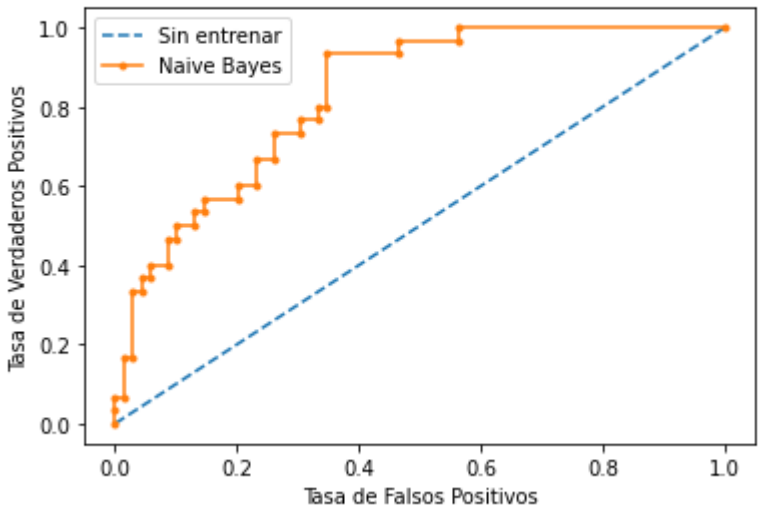


Figura 45. Curva ROC 40 características

Naive bayes con 329 datos, 11 variables explicativas y 1 variable dependiente. (Random Forest)

Tabla 96

Métricas reducción Naïve Bayes reducción a 12 características

Métricas	
Accuracy	76%
Precision	68%
Recall	71%
F1	69%

Tabla 97

Matriz de confusión Naïve Bayes 12 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 48	Falsos Negativos 11
	1: Deserta	Falsos Positivos 13	Verdaderos Positivos 27

```
Sin entrenar: ROC AUC=0.500
Naive Bayes: ROC AUC=0.865
```

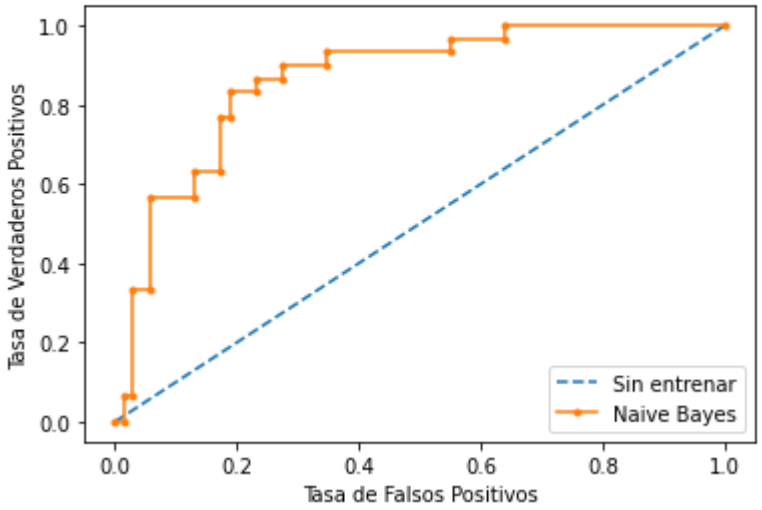


Figura 46. Curva ROC reducción a 12 características

Naive bayes con 329 datos, 19 variables explicativas y 1 variable dependiente. (Featurewiz)

Tabla 98

Métricas reducción Naive Bayes reducción a 19 características

Métricas	
Accuracy	69%
Precision	58%
Recall	66%
F1	62%

Tabla 99

Matriz de confusión Naïve Bayes 19 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 43	Falsos Negativos 13
	1: Deserta	Falsos Positivos 18	Verdaderos Positivos 25

```
Sin entrenar: ROC AUC=0.500
Naive Bayes: ROC AUC=0.851
```

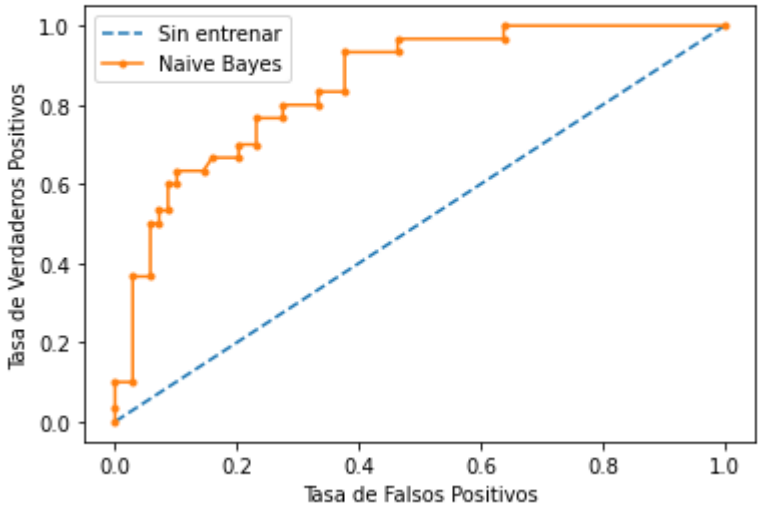


Figura 47. Curva ROC reducción a 19 características

Naive bayes con 329 datos, 93 variables explicativas y 1 variable dependiente.

Tabla 100

Métricas Naive Bayes 94 características

Métricas	
Accuracy	67%
Precisión	54%
Recall	92%
F1	68%

Tabla 101

Matriz de confusión Naïve Bayes 94 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 31	Falsos Negativos 3
	1: Deserta	Falsos Positivos 30	Verdaderos Positivos 35

```
Sin entrenar: ROC AUC=0.500
Naive Bayes: ROC AUC=0.738
```

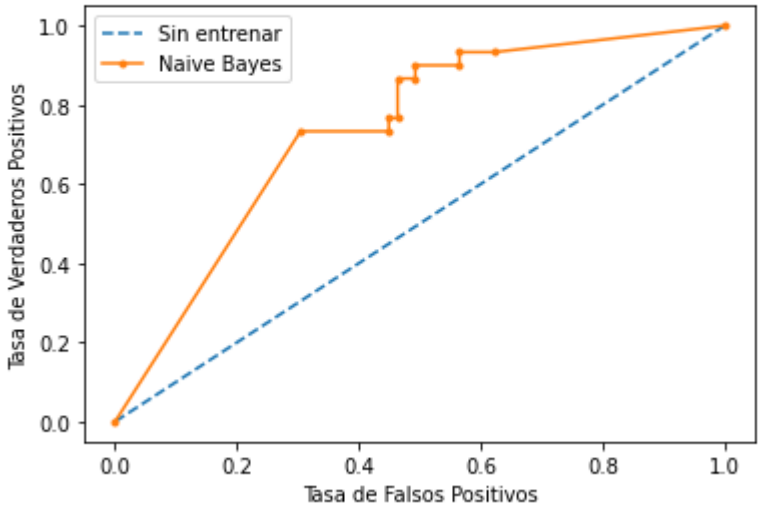


Figura 48. Curva ROC 94 características

Naive bayes con 329 datos, 8 variables explicativas y 1 variable dependiente. (Ramdon Forest)

Tabla 102

Métricas reducción Naive Bayes reducción a 9 características

Métricas	
Accuracy	75%
Precisión	63%
Recall	82%
F1	71%

Tabla 103

Matriz de confusión Naïve Bayes reducción 9 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 43	Falsos Negativos 7
	1: Deserta	Falsos Positivos 18	Verdaderos Positivos 31

La Figura 49 muestra la curva ROC y su área para el clasificador Naive Bayes usando reducción a 9 características.

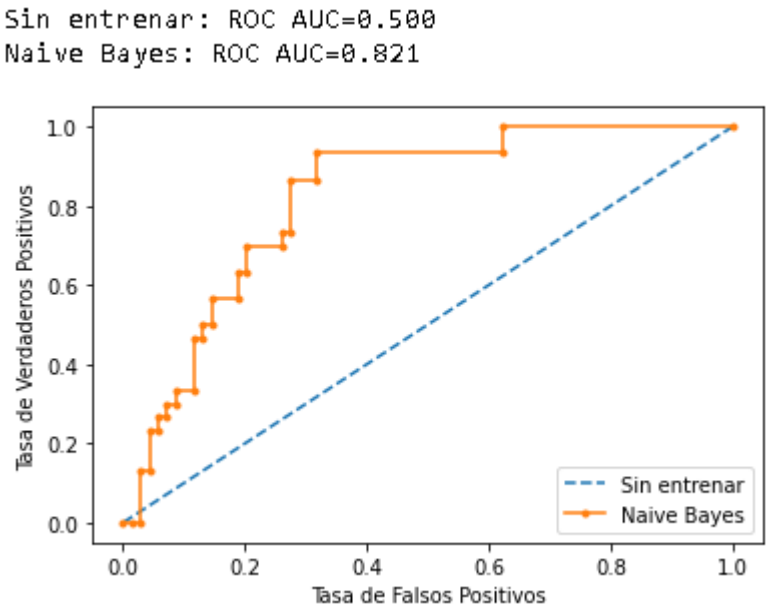


Figura 49. Curva ROC reducción a 10 características

Naive bayes usando 329 datos, 23 variables explicativas y 1 variable dependiente. (Featurewiz)

Tabla 104

Métricas para Naive Bayes reducción a 24 características

Métricas	
Accuracy	52%
Precisión	44%
Recall	97%
F1	61%

Tabla 105

Matriz de confusión Naïve Bayes 24 características.

		Predichas	
		0: Permanece	**1: Deserta**
Reales	0: Permanece	Verdaderos Negativos 14	Falsos Negativos 1

	Falsos Positivos	Verdaderos Positivos
1: Deserta	47	37

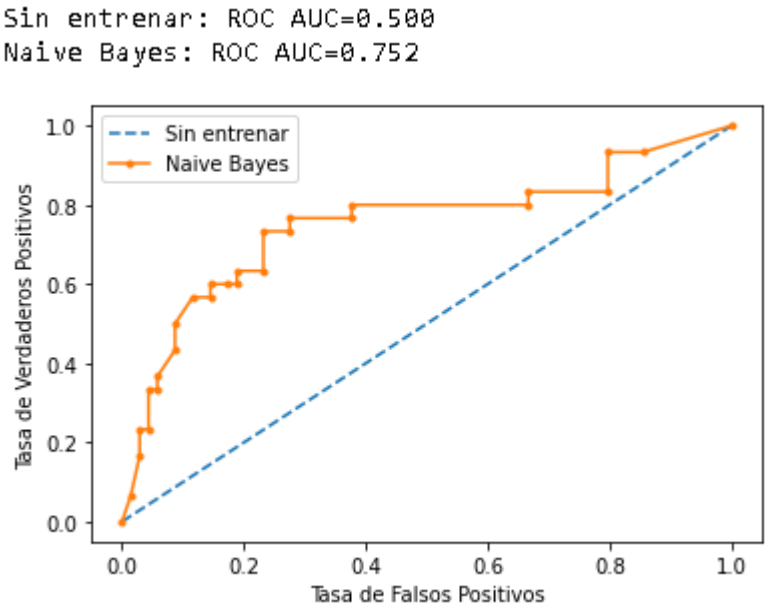


Figura 50. Curva ROC reducción a 24 características

Como se observa en las tablas 94 a 105 de métricas el clasificador Naive Bayes obtiene su mejor precisión cuando la reducción de características usando Random Forest se reduce de 93 a 9 se logra una Precisión de 75%, es decir, que 75% de los alumnos han sido clasificados correctamente, área de la curva ROC es de 0.821. Las características que corresponden al mejor modelo de Naive Bayes, en orden de importancia, son:

CICLO12, CICLO34, EDADN, 3 características en Datos generales.
COMP_HOGAR, TOTAL_INGR, CARGA_FAM, HIJOS_SUP, 4 características en el Aspecto Económico.
N_DORM, 1 variables en el aspecto de vivienda.
La Tabla 106. Muestra el mejor desempeño según la curva ROC para regresión logística con reducción a 12 características usando la técnica Random Forest y en todos los clasificadores.

Tabla 106

Resumen curva ROC, 40 características y reducciones

MÉTODO DE CLASIFICACIÓN	ROC_AUC			
	40	12	19	Prom.
RLC	0.883	0.907	0.897	0.896
DTC	0.787	0.809	0.744	0.78
MVSC	0.81	0.88	0.84	0.843
NBC	0.83	0.865	0.851	0.849

La Tabla 107. El mejor desempeño según la curva ROC se presenta en el clasificador regresión logística y para la reducción a 24 características con la técnica Featurewiz.

Tabla 107

Resumen curva ROC, 94 características y reducciones

MÉTODO DE CLASIFICACIÓN	ROC_AUC			
	94	9	24	Prom.
RLC	0.858	0.832	0.884	0.858
DTC	0.66	0.749	0.776	0.728
MVSC	0.82	0.84	0.85	0.835
NBC	0.738	0.821	0.752	0.770

4.4.2 Validación Cruzada K-Folds con k=5 y luego K=10)

Según Géron (2019), para evaluar de manera óptima se puede usar el recurso de validación cruzada de Scikit-Learn. El código ejecuta la validación cruzada K-Ford que divide aleatoriamente el conjunto de entrenamiento en 10 subconjuntos distintos llamados de partes (folds) entonces entrena y evalúa el modelo de clasificación 10 veces escogiendo una parte (fold) diferente a cada una de ellas para evaluación y entrenamiento en las otras 9 partes. El resultado es un array que contiene las 10 puntuaciones de la evaluación.

De la validación cruzada con el método K-Fold con K=10 para cada modelo clasificador, se obtiene: La precisión media y la precisión de la desviación estándar.

La tabla 108, muestra las validaciones cruzadas de las tres pruebas se contrastan y se observa que el clasificador más efectivo es Árboles de decisión en promedio (77%), seguido de Regresión logística en promedio (74%) y Máquina vector

soporte (70%) y finalizando Naive Bayes (55%). Con 329 datos, 39 variables explicativas y 1 variable dependiente. El mejor clasificador es Árboles de decisión y el peor es Naive bayes.

Tabla 108

Validación cruzada 40 características

Modelo de Clasificación	K=5	K=10
Regresión Logística	73%	75%
Árboles Decisión	76%	76%
Máquina vector soporte	69%	70%
Naive Bayes	55%	56%

La Figura 51 muestra el diagrama de cajas cuando K=5 folds, que la validación cruzada para el clasificador regresión logística los resultados se concentran en la parte superior y su asimetría es negativa o sesgada a la izquierda, mientras que para el clasificador Árboles de decisión los resultados se concentran en la parte inferior de la distribución con asimetría positiva y la dispersión de los resultados es baja, el clasificador Máquina vector soporte los resultados se concentran en la parte inferior de la distribución y tiene asimetría positiva, la dispersión es mínima es decir los datos están cerca entre sí, además de tener un valor atípico y para el clasificador Naive bayes la asimetría es negativa, los resultados se concentran en la parte superior de la distribución y los resultados de cada validación están mucho más dispersos.

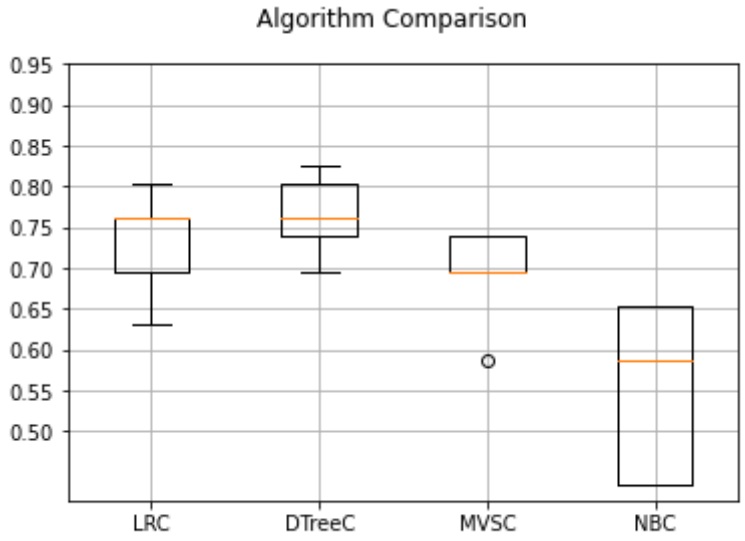


Figura 51. Diagrama de cajas y bigotes, K=5 folds.

La Figura 52, el diagrama de cajas cuando K=10 folds. La validación cruzada para el clasificador Regresión logística los resultados tienen una asimetría negativa y los resultados se encuentran concentrados en la parte superior de la distribución los resultados no están muy dispersos, para el clasificador Árboles de decisión los resultados tienen una asimetría negativa los datos se concentran en la parte superior los resultados están dispersos en relación con el anterior clasificador y tiene dos valores atípicos. Para el clasificador Máquina vector soporte los datos son asimétricos negativos los datos se encuentran concentrados en la parte superior y se encuentran más dispersos y finalmente el clasificador Naive bayes tiene una distribución asimétrica positiva los datos se concentran en la parte inferior y la dispersión es muy grande.

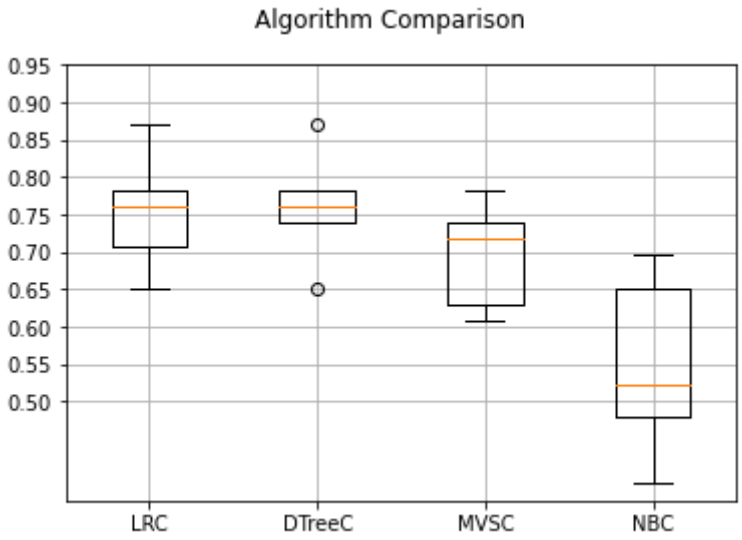


Figura 52. Diagrama cajas y bigotes, K=10 folds.

Con 329 datos, 11variables explicativas y 1 variable dependiente. (Random Forest)

En la Tabla 109 se tiene que Árboles de decisión tiene en promedio 81%, Naive bayes con 74% y finalmente Regresión logística y Máquina vector soporte tiene en promedio el mismo porcentaje 73%. El mejor clasificador es Árboles de decisión y los peores son Regresión logística y Máquina vector soporte.

Tabla 109

Validación cruzada con reducción a 12 características según RF.

Modelo de Clasificación	K=5	K=10
Regresión Logística	73%	72%
Árboles Decisión	80%	80%
Máquina vector soporte	73%	72%
Naive Bayes	73%	74%

En la Figura 53, los resultados de la validación cruzada para Regresión logística tienen una simetría negativa y se encuentran concentrados en la parte superior además su dispersión es grande, para Árboles de decisión los resultados también tienen una asimetría negativa y se encuentran concentrados en la parte superior y la dispersión de los mismos es menor en relación al anterior clasificador, el clasificador Maquina vector soporte los resultados también se distribuyen asimétricamente negativa los resultados se concentran en la parte superior y la dispersión de los mismos es mucho menor que los anteriores y finalmente tiene dos valores atípicos. Para el clasificador Naive bayes los resultados se distribuyen de manera simétrica alrededor de la media y los datos están menos dispersos y tiene un valor atípico. El mejor clasificador es Árboles de decisión y Máquina vector soporte tiene un menor rendimiento.

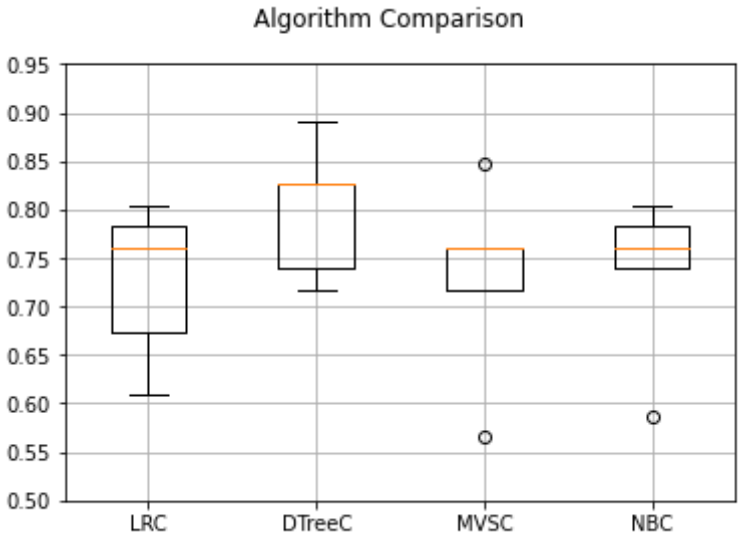


Figura 53. Diagrama cajas y bigotes, K=5 folds.

La Figura 54, los resultados de la validación cruzada cuando K=10 folds, el clasificador Regresión logística tiene una distribución simétrica y los resultados no

están muy dispersos alrededor de la media, para Árboles de decisión los datos están distribuidos asimétricamente positiva y se concentran en la parte inferior y la dispersión de los resultados no es muy grande, para Máquinas vector soporte los resultados se distribuyen asimétricamente negativa y la dispersión es grande finalmente para Naive bayes la distribución es asimétrica positiva y los resultados están mucho más dispersos.

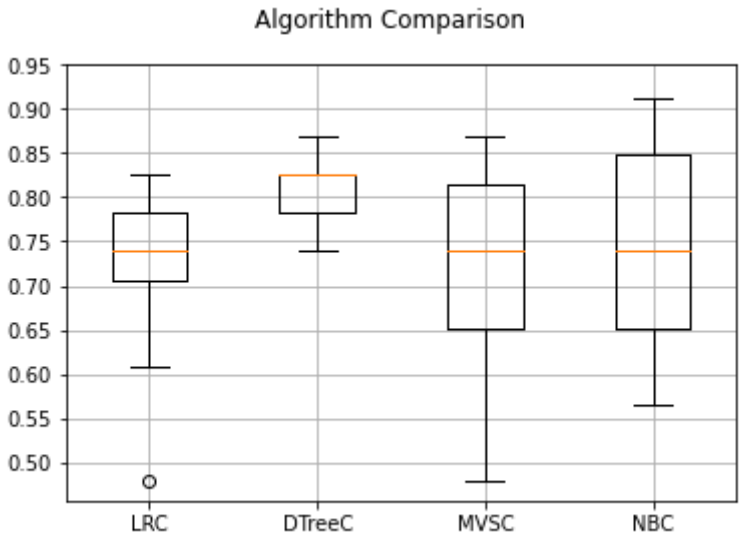


Figura 54. Diagrama de cajas y bigotes, K=10 folds.

Con 329 datos, 18 variables explicativas y 1 variable dependiente. (Featurewiz)

En la tabla 110, la mejor clasificación es Árboles de decisión con 79% promedio, seguido de Regresión logística 73% y Naive bayes con 72% en promedio y finalizando con Máquina vector soporte 69% en promedio. El mejor clasificador es Árboles de decisión y el peor Máquina vector soporte.

Tabla 110

Validación cruzada 19 características según Featurewiz.

Modelo de Clasificación	K=5	K=10
Regresión Logística	73%	73%
Árboles Decisión	79%	79%
Máquina vector soporte	70%	69%
Naive Bayes	76%	73%

Al observar la Figura 55, para cuando K=5, el clasificador Regresión logística (RLC) la distribución de los datos alrededor de la media es asimétrica negativa

concentración de los datos en la parte superior y con una mínima dispersión, para los Árboles de decisión (DTreeC) los resultados tienen asimetría negativa concentrados en la parte superior y la dispersión es mayor en relación con el anterior, para el clasificador Máquina vector soporte (MVSC) la dispersión de los datos es pequeña y los datos tienen una asimetría negativa se concentran en la parte superior la dispersión de los mismos es relativamente baja, para el clasificador Naive Bayes (NBC) la dispersión de los datos alrededor de la media es grande y los datos tienen una asimetría positiva los datos se concentran en la parte inferior y la dispersión es pequeña.

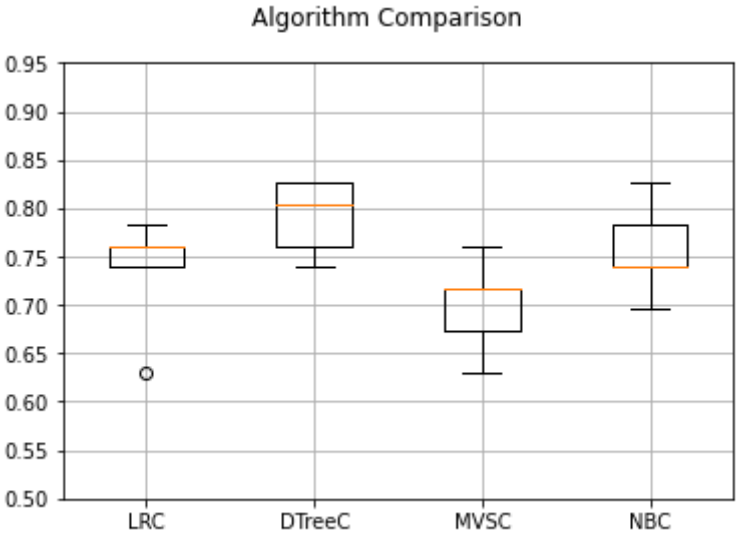


Figura 55. Diagrama de cajas y bigotes, K=5 folds.

La Figura 56, la validación cruzada con K=10 folds, el clasificador Regresión logística los resultados tienen una distribución asimétrica positiva y los datos se concentran en la parte inferior y la dispersión de los resultados es grande, para Árboles de decisión los resultados tienen una distribución asimétrica negativa los resultados se concentran en la parte superior y la dispersión es menor en relación a regresión logística, para Máquina vector soporte la distribución es asimétrica positiva los resultados están concentrados en la parte inferior y la dispersión de los resultados es menor y finalmente para Naive bayes la distribución es asimétrica positiva los resultados se concentran en la parte inferior y la dispersión es grande.

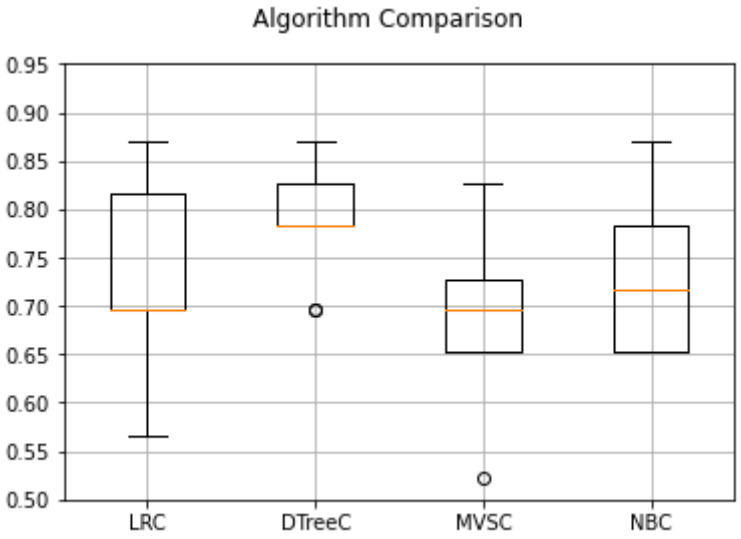


Figura 56. Diagrama de caja y bigotes, K=10 folds.

Con 329 datos, 93 variables explicativas y 1 variable dependiente.

Según la Tabla 111 la validación cruzada con mejor clasificación es Regresión logística 73% en promedio, Árboles de decisión 72% en promedio, luego Máquina vector soporte con 71%, finaliza con Naive Bayes con 56%. El mejor clasificador es Regresión logística y el peor es Naive bayes.

Tabla 111

Validación cruzada 94 características

Modelo de Clasificación	**K=5**	**K=10**
Regresión Logística	73%	72%
Árboles Decisión	71%	71%
Máquina vector soporte	70%	70%
Naive Bayes	55%	56%

Para la Figura 57, resultados de la validación cruzada k=5, los datos de la validación cruzada para Regresión logística (LRC) son asimétricos positivos y la dispersión de los datos es baja. Para los resultados de la validación cruzada de Árboles de decisión los datos son asimétricos negativos los datos están concentrados en la parte superior y la dispersión es similar en relación con el anterior clasificador. El clasificador Máquina vector soporte (MVSC) los datos son aparentemente simétricos la dispersión de los datos es muy baja en relación con el anterior clasificador y para el clasificador Naive Bayes (NBC) los resultados se distribuyen asimétricamente

negativos alrededor de la media y la dispersión de estos es grande con relación a todos los clasificadores.

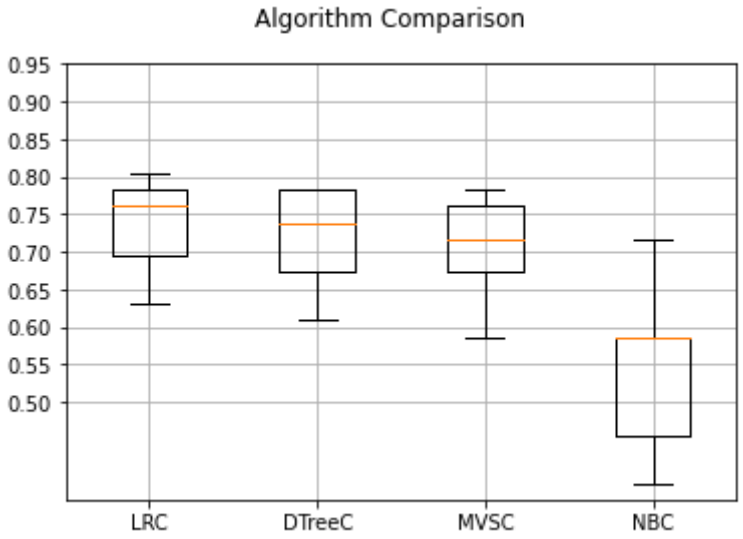


Figura 57. Diagrama cajas y bigotes, validación cruzada K=5 folds.

La Figura 58, os resultados de la validación cruzada K=10 folds, los resultados de la validación cruzada para la Regresión logística tienen una distribución asimétrica negativa y se encuentran concentrados en la parte superior además la dispersión es pequeña con tres valores atípicos, para Árboles de decisión los resultados son simétricos y se concentran alrededor de la media su dispersión es mayor, para Máquinas vector soporte se distribuyen asimétricamente negativa y la dispersión es menor en relación al anterior clasificador y finalmente Naive bayes tienen distribución simétrica alrededor de la media y la dispersión es mucho mayor que los anteriores clasificadores.

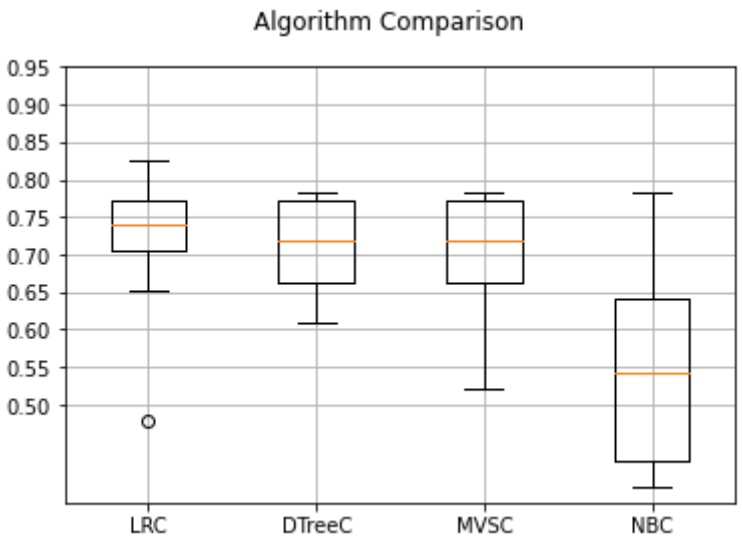


Figura 58. Diagrama de cajas y bigotes validación cruzada K=10 folds.

Con 329 datos, 8 variables explicativas y 1 variable dependiente. (Ramdon Forest)

Se observa en la Tabla 112, que el mejor clasificador es Árboles de decisión con 74% en promedio, seguidos de Regresión logística y Naive bayes con 73% en promedio y finalmente Máquina vector soporte con 72%.

Tabla 112

Validación cruzada reducción a 9 características según RF

Modelo de Clasificación	**K=5**	**K=10**
Regresión Logística	73%	72%
Árboles Decisión	73%	76%
Máquina vector soporte	72%	72%
Naive Bayes	73%	72%

En la Figura 59, los resultados del clasificador Regresión logística los datos se distribuyen de manera asimétrica negativa y la dispersión de los resultados es grande, con el clasificador Árboles de decisión (DTreeC) los datos se distribuyen de manera asimétrica positiva y la dispersión de los mismos es pequeña con dos valores atípicos, con el clasificador Máquina vector soporte (MVSC) los datos se distribuyen de manera asimétrica negativa los resultados se concentran en la parte superior y la dispersión de los datos aumenta con relación al anterior clasificador, para el clasificador Naive Bayes (NBC) los datos se distribuyen asimétricamente negativa y la dispersión de los datos es ligeramente menor .

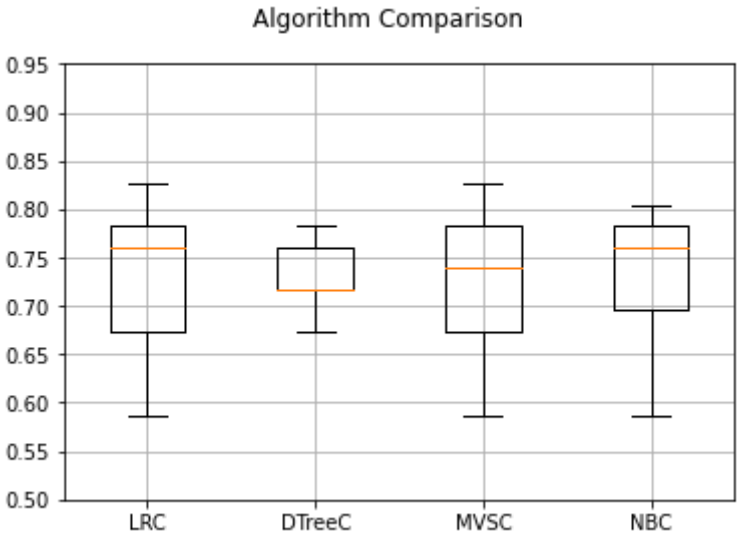


Figura 59. Diagrama de cajas y bigotes, validación cruzada K=5 folds.

La Figura 60, los resultados de la validación cruzada con K=10 folds, para el clasificador Regresión logística distribución asimétrica negativa los datos se concentran en la parte superior y la dispersión de los resultados es grande, para Árboles de decisión tienen una distribución asimétrica negativa y se concentran en la parte superior y la dispersión es pequeña además hay tres valores atípicos. Con Máquinas vector soporte los resultados se distribuyen simétricamente alrededor de la media y la dispersión es mayor con un valor atípico y finalmente los resultados de Naive bayes se distribuyen asimétricamente positiva los resultados se concentran en la parte inferior y la dispersión de los resultados es muy grande.

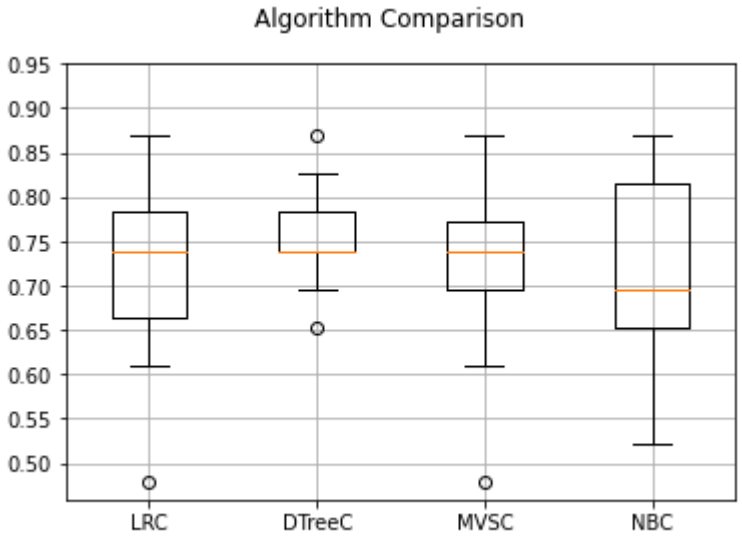


Figura 60. Diagrama de cajas y bigote validación cruzada K=10 folds.

Con 329 datos, 23 variables explicativas y 1 variable dependiente. (Featurewiz)

La siguiente Tabla 113 el mayor porcentaje es para el clasificador Árboles de decisión 78%, Máquina vector soporte 74%, Regresión logística 74%, y finalmente Naive Bayes el menor porcentaje 51%.

Tabla 113

Validación cruzada reducción a 24 características según Featurewiz

Modelo de Clasificación	**K=5**	**K=10**
Regresión Logística	73%	74%
Árboles Decisión	77%	78%
Máquina vector soporte	72%	72%
Naive Bayes	43%	43%

De la Figura 61 el clasificador Regresión logística tiene asimetría negativa concentración de los datos en la parte superior y la dispersión de los datos es alta con relación a los otros clasificadores, para el clasificador Árboles de decisión los resultados se distribuyen de manera asimétrica negativa y con una muy pequeña dispersión, para el clasificador Máquina vector soporte (MVSC) se distribuyen aproximadamente de manera asimétricamente negativa alrededor de la media y los datos tienen una mayor dispersión. Con el clasificador Naive Bayes (NBC) la distribución también es asimétrica positiva y la dispersión de los resultados no es muy grande y con un valor atípico. El mejor clasificador según la validación cruzada es Regresión logística y el peor es Naive bayes.

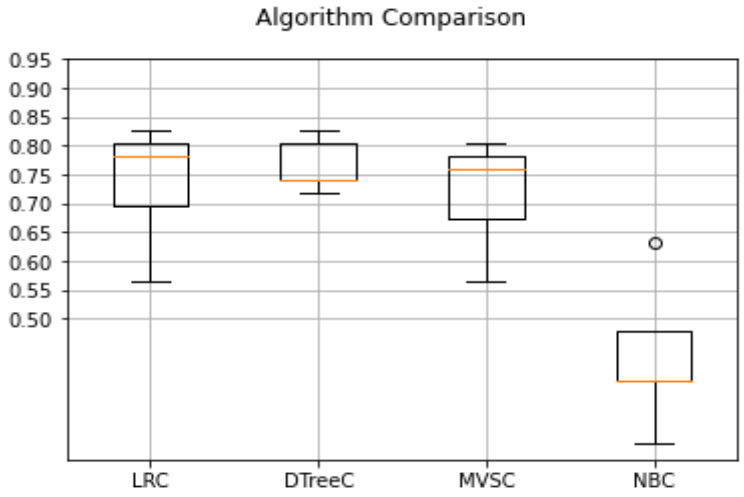


Figura 61. Diagramas de caja y bigotes, validación cruzada K=5 folds.

De la Figura 62 el clasificador Regresión logística tiene asimetría positiva y la dispersión de los datos es baja con relación a los otros clasificadores existe un valor atípico, para el clasificador Árboles de decisión los datos se distribuyen de manera simétrica y con una gran dispersión, para el clasificador Máquina vector soporte (MVSC) los datos se distribuyen aproximadamente de manera simétricamente alrededor de la media y los datos tienen baja dispersión. Con el clasificador Naive Bayes (NBC) la distribución también es asimétrica positiva y la dispersión de los datos es grande. El mejor clasificador según la validación cruzada es Árboles de decisión y el peor es Naive bayes.

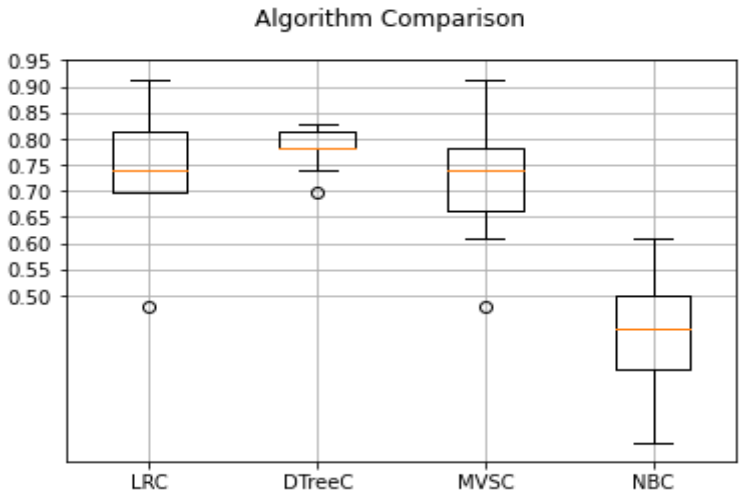


Figura 62. Diagrama de cajas y bigotes validación Cruzada K=10 folds.

En la siguiente Tabla 114 se integran los resultados obtenidos y se calculan los promedios de las clasificaciones según la cantidad de atributos y su reducción según las dos técnicas de random forest y featurewiz; observamos que en la línea de aciertos totales y según su desempeño el modelo Regresión logística 77% seguido de Árboles de decisión y el modelo Máquina vector soporte tiene 74% y finaliza con Naive bayes con 63.67% como sus aciertos promedio totales. También, el clasificador que menos errores en la clasificación tuvo es Regresión logística 22%, luego Árboles de decisión y Máquina vector soporte ambos 25%, la mayor proporción de fallos es 35.33% para Naive bayes. Los promedios de los aciertos son aceptables. El clasificador que tiene menos falla es Regresión logística 22%, seguido de Árboles de decisión y Máquina vector soporte ambos 25% y Naive bayes 35.33%. El mejor clasificador es Regresión logística y el peor es el clasificador Naive bayes.

Tabla 114

Matriz de Confusión 94 características, 9 y 24 características

	REGRESION LOGISTICA			**ARBOLES DE DECISIÓN**			**MÁQUINA VECTOR SOPORTE**			**NAÏVE BAYES**		
NÚMERO DE ATRIBUTOS	94	9	24	94	9	24	94	9	24	94	9	24
VERDADERO POSITIVO	24	29	28	19	22	21	23	30	29	35	31	37
VERDADERO NEGATIVO	52	46	52	50	55	55	47	45	48	31	43	14
TOTAL	76	75	80	69	77	76	70	75	77	66	74	51
ACIERTOS TOTALES	77			74			74			63.666667		
FALSO POSITIVO	17	23	17	19	14	14	14	16	13	30	18	47
FALSO NEGATIVO	6	1	2	11	8	9	15	8	9	3	7	1
TOTAL	23	24	19	30	22	23	29	24	22	33	25	48
FALLOS TOTALES	22			25			25			35.333333		

A continuación, la Tabla 115 se recopila las matrices de confusión calculando el promedio de aciertos totales y el promedio de fallos totales observando que el que mejor clasifica es Regresión logística 78.33%, luego Árboles de decisión 76.33%, Máquina vector soporte 68.67% y finalmente Naive bayes con el promedio más bajo de 68%. El clasificador que menos falla es Regresión logística 20.67% seguido de Árboles de decisión con 22.67%, a continuación, Máquina vector soporte con

30.33% y el que tiene mayor cantidad de fallos es Naive bayes 31%. El mejor clasificador según la Tabla 113, es Regresión logística y el que mejor clasifica correctamente o acierta en su clasificación y el que más falla es el clasificador Naive bayes.

Tabla 115

Matrices de confusión 40 características, 12 y 19 características.

	REGRESION LOGISTICA			**ARBOLES DE DECISIÓN**			**MÁQUINA VECTOR SOPORTE**			**NAÏVE BAYES**		
NÚMERO DE ATRIBUTOS	40	12	19	40	12	19	40	12	19	40	12	19
VERDADERO POSITIVO	29	28	29	21	20	23	21	27	21	26	27	25
VERDADERO NEGATIVO	48	52	49	57	56	52	47	46	44	35	48	43
TOTAL	77	80	78	78	76	75	68	73	65	61	75	68
ACIERTOS TOTALES	78.333333			76.33333333			68.66666667			68		
FALSO POSITIVO	21	17	20	12	13	15	14	15	17	26	13	18
FALSO NEGATIVO	1	2	1	9	10	9	17	11	17	12	11	13
TOTAL	22	19	21	21	23	24	31	26	34	38	24	31
FALLOS TOTALES	20.66666667			22.6666667			30.3333333			31		

Las matrices de confusión permiten evaluar métricas adicionales como precisión con la finalidad de obtener la probabilidad de clasificar un registro en su categoría según corresponda; y el puntaje F1, para el promedio ponderado de precisión y sensibilidad.

En la Tabla 116 se observan los clasificadores que obtuvieron mejores resultados cuando se tienen 40 atributos y sus reducciones a 12 y 19 características, en orden de importancia, son: Regresión logística con reducción a 12 características, Árboles de decisión con 40 atributos, Naive bayes con 12 características y Máquina vector soporte con 12 atributos.

Tabla 116

Resultados, métricas adicionales 40 atributos y sus reducciones

NÚMERO CARACTERISTICAS		PRECISIÓN			RECALL			F1		
		40	**12**	**19**	**40**	**12**	**19**	**40**	**12**	**19**
REGRESIÓN LOGÍSTICA	NO DESERTOR	98%	96%	98%	70%	75%	71%	81%	85%	82%
	DESERTOR	58%	62%	59%	97%	93%	97%	72%	75%	73%
	MP	85.9%	85.7%	86.2%	78.2%	80.5%	78.9%	78.3%	82%	79.3%
ÁRBOLES DECISIÓN	NO DESERTOR	86%	85%	85%	83%	81%	78%	84%	83%	81%
	DESERTOR	64%	61%	61%	70%	67%	72%	67%	63%	66%
	MP	85.9%	77.7%	77.2%	79.1%	76.8%	76.1%	78.8%	76.9%	76.2%
MACHINE VECTOR SUPPORT	NO DESERTOR	73%	81%	75%	77%	75%	72%	75%	78%	72%
	DESERTOR	60%	64%	57%	55%	71%	55%	58%	67%	55%
	MP	68%	74.5%	65.5%	68.6%	73.5%	65.5%	68.5%	73.8%	65.5%
NAÏVE BAYES	NO DESERTOR	74%	81%	77%	57%	79%	70%	65%	80%	74%
	DESERTOR	50%	68%	58%	68%	71%	66%	58%	69%	62%
	MP	64.8%	76%	69.2%	61.2%	75.9%	68.5%	62.3%	75.8%	69.4%

En la Tabla 117 se muestra que los clasificadores con mejores resultados cuando el set de atributos es de 94 y sus reducciones a 9 y 24 características, en orden descendente, son: Regresión logística con 24 características, Máquina vector soporte con 24 atributos, Árboles de decisión con 9 características, Máquina vector soporte con 24 características y finalmente Naive bayes con reducción al set de 9 atributos.

Tabla 117

Resultados, métricas adicionales 94 atributos y sus reducciones.

NÚMERO CARACTERISTICAS		PRECISIÓN			RECALL			F1		
		94	9	24	94	9	24	94	9	24
REGRESION LOGISTICA	NO DESERTOR	90%	98%	96%	75%	67%	75%	82%	79%	85%
	DESERTOR	59%	56%	62%	80%	97%	93%	68%	71%	75%
	MP	80.6%	85.3%	85.7%	76.5%	76.1%	80.5%	77.8%	76.6%	82%
ÁRBOLES DECISIÓN	NO DESERTOR	82%	87%	86%	72%	80%	80%	77%	83%	83%
	DESERTOR	50%	61%	60%	63%	73%	70%	56%	67%	65%
	MP	72.3%	79.1%	78.1%	69.3%	77.9%	77%	70.6%	78.2%	77.5%
MACHINE VECTOR SUPPORT	NO DESERTOR	76%	85%	84%	77%	74%	79%	76%	79%	81%
	DESERTOR	62%	65%	69%	61%	79%	76%	61%	71%	72%
	MP	70.6%	77.3%	78.2%	70.9%	75.9%	77.8%	70.2%	75.9%	77.5%
NAÏVE BAYES	NO DESERTOR	91%	86%	93%	51%	70%	23%	65%	77%	37%
	DESERTOR	54%	63%	44%	92%	82%	97%	68%	71%	61%
	MP	76.8%	77.2%	74.2%	66.7%	74.6%	51.4%	66.2%	74.7%	46.2%

Según la Tabla 118 se resume el porcentaje de los correctamente clasificados de la data codificada con 40 características y sus reducciones a 12 características por la técnica de Random Forest y 19 características por la técnica automática de Featurewiz y que el mejor desempeño lo tiene Regresión logística 79%, Árboles de decisión 77%, Máquina vector soporte con 70% y Naive bayes con 69%.

Tabla 118

Resultados Métricas set 40 atributos y sus reducciones.

NÚMERO DE ATRIBUTOS	CORRECTAMENTE CLASIFICADOS (Accuracy)			
	Regresión Logística	Árboles Decisión	Máquina Vector Soporte	Naive Bayes
Todas 40	78%	79%	69%	62%
12 RF	81%	77%	74%	76%
19 FW	79%	76%	66%	69%
Promedio	79%	77%	70%	69%

En la Tabla 119 se resume los correctamente clasificados de la data con características dummy y que resultan 94 características teniendo la mejor clasificación Regresión logística con un promedio aproximado de 78% seguido de Árboles de decisión y Máquina vector soporte ambos con 75% y Naive bayes con 65%.

Tabla 119

Resultados Métricas 94 atributos y sus reducciones.

NÚMERO DE ATRIBUTOS	**CORRECTAMENTE CLASIFICADOS (Accuracy)**			
	Regresión logística	Árboles decisión	Máquina Vector Soporte	Naive Bayes
Todas 94	77%	70%	71%	67%
9 RF	76%	78%	76%	75%
24 FW	81%	77%	78%	52%
Promedio	78%	75%	75%	65%

La Tabla 120 presenta los resultados de la validación cruzada con K = 5 y K= 10 Folds, considerando el desempeño de cada modelo según las características y su reducción usando las técnicas de Random Forest y Featurewiz. Los promedios del desempeño indican que el modelo Árboles de decisión tiene el mejor desempeño, le sigue Regresión Logística, luego Máquina vector soporte y finalmente Naive Bayes.

Tabla 120

Validación cruzada promedio K=5 y K=10.

	VALIDACIÓN CRUZADA K FOLDS			
	Regresión logística	**Árboles Decisión**	**Máquina vector soporte**	**Naïve Bayes**
Características	**K=5 K=10**	**K=5 K=10**	**K=5 K=10**	**K=5 K=10**
40	0.7413	0.7608	0.6957	0.5543
12	0.7239	0.8022	0.7261	0.7369
19	0.7347	0.7891	0.6957	0.7435
94	0.7260	0.7086	0.7043	0.5522
9	0.7239	0.7435	0.7217	0.7239
24	0.7391	0.7739	0.7195	0.4347
PROMEDIO	0.7315	0.7631	0.7105	0.6243

En la Tabla 121 se muestran los atributos que influyen en cada modelo de clasificación y que tienen mejor desempeño notándose que en los aspectos generales son ciclo y edad, en el aspecto económico el ingreso total, carga familiar y los hijos que estudian en superior, en el aspecto del estudiante dependencia económica y el riesgo familiar para el aspecto de la vivienda el tipo de construcción y número de dormitorios.

Tabla 121

Resumen atributos que influyen

ASPECTO	12 ATRIBUTOS	9 ATRIBUTOS	24 ATRIBUTOS
DATOS GENERALES	CICLO, EDADN, DIR_ZONA	CICLO12, CICLO34, EDADN	CICLO12, EP_MINAS
ASPECTO ECONOMICO	TOTAL_INGRE, CARGA_FAM, COMP_HOGAR, HIJOS_SUP, VIVIENDA_EST, SOST_HOGAR	TOTAL_INGRE, CARGA_FAM, COMP_HOGAR, HIJOS_SUP.	CARGA_FAM, HIJOS_SUP, VIV_CUID, VIV_ALOJ
ASPECTO ESTUDIANTE			DEPEC_SIMIS, HUERF_MAD, HUERF_PAD.
ASPECTO DE LA VIVIENDA	TIPO_CONSTR, N_DORM.	N_DORM	TEVIV_PRO, TEVIV_INV, TPC_NOBLE, TPC_MIX, TPC_PREC, TIV_DEPA, TIV_CONV, TIV-OTR, N_BANO, TV_COLOR, CELULAR, INTERNET, COMPUTADORA, BIBLIO_PERS.

4.5 Discusión de los resultados

La deserción estudiantil en los estudiantes de la universidad nacional de Moquegua se evidencia que se presentan en los primeros ciclos y las variables que influyen son las condiciones socioeconómicas y aspectos vinculados con la familia como lo afirma (Gamarra *et al*., 2018).

Respecto a la reducción de características, cuando se comparan los modelos construidos a partir de datos completos con los modelos a los cuales se aplicaron técnicas de reducción los modelos construidos con datos completos funcionan mejor que los segundos.(Silva y Roman, 2021)

Los enfoques o técnicas empleados para evaluar el desempeño de los clasificadores son las métricas son: accuracy, recall, F1 y el área debajo de las curvas ROC tal y como se mencionan en la investigación de (Silva y Roman, 2021) y también (Del Bonifro et al., 2020c)

Para la selección del mejor modelo se usan las métricas de precisión, matriz de confusión y métricas adicionales ajustando los parámetros usando la validación cruzada con K=5 y 10 Folds. (Ayala-yaguara y Valenzuela-sabogal, 2020)

Los resultados obtenidos a partir de datos que no tienen valor pedagógico o didáctico son una herramienta para poder reducir la deserción.(Del Bonifro et al., 2020c)

Estudios hechos sobre deserción universitaria en universidades latinoamericanas coinciden en que la mayoría de estudiantes desertan en los primeros ciclos y que las variables que más influyen en la deserción las condiciones socioeconómicas del estudiante en nuestro caso coincidimos en dependencia económica del estudiante, composición familiar e ingreso familiar además del tipo de vivienda que, para esta investigación, corresponden a los datos generales, aspecto económico y aspecto de vivienda.(Gamarra et al., 2018)

CONCLUSIONES

La presente investigación tuvo como objetivo determinar el mejor modelo de Machine Learning para analizar y predecir la deserción estudiantil de los alumnos de la Universidad Nacional de Moquegua sede Mariscal Nieto. Concluyendo lo siguiente:

Primera: La hipótesis que se planteó al iniciar esta investigación sobre el mejor modelo de clasificación supervisada para predecir la deserción estudiantil es regresión logística para los tipos de variables identificados, no se verifica, según la validación cruzada, el mejor clasificador es Árboles de decisión su mejor precisión se obtiene cuando se hace la reducción de características usando Random Forest de 40 a 12 características logrando una Precisión de 80%. En la reducción de características de 94 a 24 usando Featurewiz dos veces, Árboles de decisión logra una precisión de 77% en promedio. Según las métricas de evaluación se ha demostrado que el mejor modelo para estudiar la deserción estudiantil de los alumnos de la Universidad Nacional de Moquegua es Árboles de decisión (ver la Tabla 120)

Segunda: En el proceso de selección de características que determinan la deserción estudiantil se detectaron como importantes, las características que corresponden al mejor modelo en su mejor desempeño 80% de precisión y 12 características. el CICLO, EDADN, DIR_ZONA en la parte de datos generales y TOTAL_INGRE, CARGA_FAM, COMP_HOGAR, HIJOS_SUP, VIVIENDA_EST, SOS_HOGAR en el aspecto económico, también, TIPO_CONSTR, N_DORM en el aspecto de vivienda. Este estudio ha identificado las características que influyen en la deserción estudiantil de los alumnos de la Universidad Nacional de Moquegua. (Ver Tabla 121)

Tercera: Según la validación cruzada el clasificador Regresión logística obtiene su mejor precisión con 40 características se logra una Precisión de 74%, es decir, que 74% de los alumnos han sido clasificados correctamente como desertores. El clasificador Máquina vector soporte obtiene su mejor precisión cuando la reducción de características usando Random Forest

se reduce de 40 a 12 se logra una Precisión de 73%, es decir, que 73% de los alumnos han sido clasificados correctamente como desertores. El clasificador Naive Bayes obtiene su mejor precisión cuando la reducción de características usando Featurewiz se reduce de 40 a 12 se logra una Precisión de 74%, es decir, que 74% de los alumnos han sido clasificados correctamente. (Ver Tabla 120)

Cuarta: Según la validación cruzada, y los promedios de su desempeño tenemos que el modelo de Árbol de decisión tiene el mejor desempeño con 77% aproximadamente, le sigue Regresión Logística con 73%, luego Máquina vector soporte con 71% y finalmente Naive Bayes con 62%. Ver Tabla 120. La validación cruzada, determina y garantiza que el rendimiento del clasificador es independiente de la partición debido a que esa selección es completamente aleatoria entre datos de entrenamiento y prueba por lo que el mayor rendimiento corresponde al clasificador Árboles de decisión. (Ver Tabla 120)

Quinta: La introducción de variables dummy mejora el desempeño de los cuatro modelos de clasificación, esta mejora disminuye en los dos primeros clasificadores: Regresión logística y Árboles de decisión, pero esta se compensa cuando las características disminuyen a 24 en Regresión logística y en Árboles de decisión. En los otros dos últimos clasificadores (Maquina vector soporte y Naive Bayes) la mejora aumenta, se incrementa en la reducción a 9 características en el caso de Máquina vector soporte y 9 características en Naive bayes. (ver Tabla 120)

RECOMENDACIONES

Primera: Al trabajar con datos reales se debe tener cuidado con el uso y la confidencialidad e identidad de estos. El modelo de clasificación debe ser integrado como una herramienta de predicción los otros servicios de la Universidad con el objetivo de monitorear la deserción universitaria.

Segunda: La selección de características es un tema amplio y no existe una sola técnica para hacer la selección. Desde el momento que obtenemos la data se hace reducción de características ya sea por falta o ausencia de datos por alta correlación entre características o por muy baja varianza de sus datos. Se debe buscar que las características tengan la mayor varianza. Se debe hacer un estudio exhaustivo de esas técnicas.

Tercera: Muy a pesar de que los métodos de clasificación tienen parámetros por defecto siempre es necesario revisar otras opciones o configuración de sus parámetros. Como es el caso en regresión logística que tiene establecido por defecto un máximo de 100 iteraciones y el método que usa para minimizar el problema es por defecto lbfgs (método cuasi newton para minimizar funciones) y no esta activada la opción para equilibrar las muestras Class_weight, en nuestro caso 220 no desertan y 109 desertan no son muestras equilibradas, que fue considerado usando la opción 'balanced' que está disponible en los tres primeros clasificadores Regresión lineal, Árboles de decisión y Máquina vector soporte, pero no disponible en Naive bayes y no permitió que las métricas relacionadas a la matriz de confusión y sus métricas no generalizara bien el clasificador. Por lo que se debe tener cuidado con las diferentes configuraciones.

Cuarta: Un trabajo importante y motivo de otros trabajos similares es el desarrollo de técnicas de reducción de dimensiones para datos mixtos, es decir factores cuantitativos y cualitativos para cantidades de grandes características.

Quinta: Como todo modelo, para analizar algún problema de clasificación, se sugiere una revisión periódica, ya que el comportamiento de la modelo

varia al incrementarse o modificarse las variables. En el estudio de la deserción estudiantil los datos siempre serán desequilibrados debido a que la cantidad de estudiantes que desertan no igualan a la cantidad de estudiantes que permanecen por lo que se recomienda considerar alguna de las técnicas para datos desequilibrados.

Sexta: Se recomienda la actualización de la ficha socioeconómica. A la fecha la universidad nacional de Moquegua ha incrementado su oferta académica con nuevas carreras y dado este incremento seria motivo de otros estudios comparar los niveles de deserción y predicción de estos a nivel de las dos provincias Mariscal nieto e Ilo puesto que la universidad tiene carreras profesionales diferenciadas en cada provincia y podría hacerse seguimiento a la implementación del modelo de predicción.

BIBLIOGRAFÍA

Alban Taipe, M. S. (2019). Contribuciones a la Predicción de la Deserción Universitaria a través de Minería de Datos. In *Universidad Nacional Mayor de San Marcos*. https://doi.org/htpp://orcid.org/0000-0003-1519-4023 2.

Ayala-yaguara, H. Y., y Valenzuela-sabogal, G. M. (2020). *Revista Ontare, 7, 133-150. 7*. https://journal.universidadean.edu.co/index.php/Revistao/article/view/2676

Benitez, R. A. (2012, December 7). *Autor del Proyecto | Proyecto R.A.M.O.N.* https://proyectoramon.wordpress.com/about/

Buabeng-andoh, C. (2022). *Using Machine Learning Techniques to Predict Seasonal Rainfall-New*. *2022*.

Burkov, A. (2019). The Hundred-Page Machine Learning Book. En *Syria Studies*. https://www.researchgate.net/publication/269107473_What_is_governance/link/548173090cf22525dcb61443/download%0Ahttp://www.econ.upf.edu/~reynal/Civil wars_12December2010.pdf%0Ahttps://think-asia.org/handle/11540/8282%0Ahttps://www.jstor.org/stable/41857625

Carmona, E. (2016). *Abstract Support Vector Machine 1 Introducción*. *November*, 1–27. https://www.researchgate.net/publication/263817587_Tutorial_sobre_Maquinas_de_Vectores_Soporte_SVM

Centro de microdatos Universidad de Chile. (2008). *Causas-Desercion-Universitaria-Chile*. https://doi.org/10.1007/BF03186740

Cevallos, F. (2014). Desercion, calidad y reforma universitaria. In *Serie: Cuadernos del Contrato Social por la Educación*.

De, A. C., Gomes, S., Santos, D., De, T., Menezes, P., Rego, H., Da Hora, M., Sugiyarti, E., Jasmi, K. A., Basiron, B., Huda, M., Shankar, K., Maseleno, A., Nacional, P., Edital, E., Tempore, P., Simplificado, P. S., Disposi, D. A. S., Preliminares, E. S., … Gaudencio Do Rêgo, T. (2018). Modelo para Classificação do Risco de Abandono Escolar em Cursos de Engenharia com Base em Métodos de Academic Analytics. *Anais Dos Workshops Do VI Congresso Brasileiro de Informática Na Educação (CBIE 2017)*, *1*(Cbie).

Del Bonifro, F., Gabbrielli, M., Lisanti, G., y Zingaro, S. P. (2020). Student Dropout Prediction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12163 LNAI*, 129–140. https://doi.org/10.1007/978-3-030-52237-7_11

Del Bonifro, F., Gabbrielli, M., Lisanti, G., y Zingaro, S. P. (2020). Student dropout prediction. En Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12163 LNAI. Springer International Publishing. https://doi.org/10.1007/978-3-030-52237-7_11

Diego, M. P. (2019). Modelo de mineria de datos basado en factores asociados para la predicción de deserción estudiantil universitaria. [Universidad Nacional de Moquegua]. In *Universidad Nacional De Moquegua*. https://www.mendeley.com/viewer/?fileId=7fdce2cc-93fb-c5ad-7830-62f46ce48877&documentId=38b50c2e-8b88-3cd6-b075-4055739ed36a

Farquad, M. A. H., Ravi, V., & Bapi, R. S. (2009). Support vector machine-based hybrid classifiers and rule extraction thereof: Application to bankruptcy prediction in banks. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. https://doi.org/10.4018/978-1-60566-766-9.ch019

Fiegehen, L. E. G., & Díaz, O. E. (2016). Deserción En Educación Superior En América Latina Y El Caribe. Paideia, Revista De Educación, 0(45), 33–46. https://www.researchgate.net/publication/275275484_Desercion_en_educacion_superior_en_America_Latina_y_el_Caribe_2008-16.

Gamarra, D., Matos, R., & Yupanqui, M. (2018). Detección de patrones de éxito en estudios universitarios de la Universidad Continental. *Apuntes de Ciencia & Sociedad*, *08*(01). https://doi.org/10.18259/acs.2018005

Gamarra, D., Matos, R., & Yupanqui, M. (2018). Detección de patrones de éxito en estudios universitarios de la Universidad Continental. Apuntes de Ciencia & Sociedad, 08(01). https://doi.org/10.18259/acs.2018005

Garavaglia, S., Sharma, A., & Hill, M. (1998). a Smart Guide To Dummy Variables: Four Applications and a Macro. Entropy. http://www.ats.ucla.edu/stat/sas/library/nesug98/p046.pdf

Géron, A. (2019). Mãos à Obra Aprendizado de Máquina com Scikit-Learn & TensorFlow: Conceitos, Ferramentas e Técnicas para a Construção de Sistemas Inteligentes.

Hernández Sampieri, R., Fernández Collado, C., & del Pilar Baptista Lucio, M. (2010). *Metodología de la investigación, 5ta Ed*. www.FreeLibros.com

Isphording, I. E., & Raabe, T. (2019). *RESEARCH REPORT SERIES Early Identification of College Dropouts Using Machine-Learning*. *89*. https://doi.org/10.5157/NEPS

Iván, J., & Echeverry, M. (2017). Propuesta de un modelo estadístico para caracterizar y predecir la deserción estudiantil Universitaria. http://www.bdigital.unal.edu.co/58059/1/71787491.2017.pdf

Kiesling, H. J. (1971). *Multivariate analysis of schools and educational policy*. *1*, ix, 47 p.

Lantz, B. (2019). *Machine Learning with R*.

Lottering, R., Hans, R., & Lall, M. (2020). A machine learning approach to identifying students at risk of dropout: a case study. *International Journal of Advanced*

Computer Science and Applications, *11*(10), 417–422. https://doi.org/10.14569/IJACSA.2020.0111052

Marrugat, A. S., & Ginebra, J. (2020). *Màster Universitari en Enginyeria d ' Organització Comparación de algoritmos de clasificación supervisada Escola Tècnica Superior d ' Enginyeria Industrial de Barcelona*.

Merlino, A., Ayllón, S., Escanés, G., Electrónica, R., Investigativas, A., & Rica, U. D. C. (2011). *Variables That Influence First Year University Students' Dropout Rates. Construction of Dropout Risk Indexes*. http://www.redalyc.org/articulo.oa?id=44720020005

Oviedo Bayas, B., & Zambrano-Vega, C. (2019). *Nuevo clasificador bayesiano simple para el análisis de datos educativos | Universidad y Sociedad*. *11*. https://rus.ucf.edu.cu/index.php/rus/article/view/1191

Parra Rodríguez, F. J. (2017). Métodos de clasificación | Estadística y Machine Learning con R. In *Estadística y Machine Learning con R.* (p. 288).

Peralta Castro, R., Rodríguez Mora, J., & Jiménez Serrato, S. (2016). Variables asociadas a la deserción estudiantil: Estudio de caso en la Fundación Universitaria. *Escenarios*, *14*(141), 117–129. https://doi.org/10.15665/esc.v14i1.883

Pérez Hoyos, S. (1996). *Introducción a la regresión logística*. *7*, 1–11.

Plasencia, R. (2018). Más de 200 millones de dólares se perdieron en dos años por deserción universitaria. https://logrosperu.com/blog/actualidad/mas-de-200-millones-de-dolares-se-perdieron-en-dos-anos-por-desercion-universitaria-707

Quintela Dávila, G. E., & Hoy, S. (2013). Deserción universitaria, una aproximación sociológica al proceso de toma de decisiones de los estudiantes. *Sociedad Hoy*, *24*, 83–106. http://www.ashe.ws/

Raschka, S., Liu, Y. (Hayden)., Mirjalili, V., & Dzhulgakov, D. (2022). Machine Learning with Pytorch and Scikit-Learn Develop Machine Learning and Deep Learning Models with Python.

Revista Dinero. (2017). ¿Por qué enfrentamos una tasa tan alta de deserción en la educación superior? Http://Www.Dinero.Com/Pais/Articulo/Desercion-y-Abandono-de-La-Educacion-Universitaria-En-Colombia/247068. http://www.dinero.com/pais/articulo/desercion-y-abandono-de-la-educacion-universitaria-en-colombia/247068

Ríos Ramírez Roger Ricardo. (2017). *ROGER RICARDO RIOS RAMIREZ Metodología para la investigación y redacción*.

Roa Sánchez, N. (2017). Fundamentos de investigación. En *Fundamentos de investigación*. https://doi.org/10.33132/9789585459670

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, *8*, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

Sanchez Nina, A. (2017). *Modelo Estadístico para Determinar la Deserción Estudiantil de las Escuelas Profesionales de la UNA - PUNO, 2017*. 111. http://repositorio.unap.edu.pe/handle/UNAP/6297

Sánchez-Hernández, G., Barboza-Palomino, M., & Castilla-Cabello, H. (2017). Análisis de la deserción y los factores asociados a la permanencia estudiantil en una universidad peruana. *Actualidades Pedagógicas*, *69*, 169–191. https://doi.org/10.19052/ap.4075

Silva, J. J. da, & Roman, N. T. (2021). *Predicting Dropout in Higher Education: a Systematic Review*. *Cbie*, 1107–1117. https://doi.org/10.5753/sbie.2021.217437

Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology*, *9*(4), 1–5. https://doi.org/10.17485/ijst/2016/v9i4/87032

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to Predict Dropout in University Students with Machine Learning. *2018 IEEE International Work Conference on Bioinspired Intelligence, IWOBI 2018 - Proceedings*, 1–6. https://doi.org/10.1109/IWOBI.2018.8464191

Soto, G. (2015). *Uso de técnicas de machine learning para predecir el rendimiento académico de los estudiantes de la Carrera de Ingeniería Civil en Informática de la Universidad del Bío-Bío, Chillán*. http://repobib.ubiobio.cl/jspui/handle/123456789/2610

Tan, M., & Shao, P. (2015). Prediction of student dropout in E-learning program using machine learning method. *International Journal of Emerging Technologies in Learning*, *10*(1), 11–17. https://doi.org/10.3991/ijet.v10i1.4189

Tinto, V. (1982). *Definir la deserción: una cuestión de perspectiva*. 1–9. http://publicaciones.anuies.mx/

Valderrama Mendoza, S. (2018). *Pasos para elaborar proyectos de investigación científica*.

Vereau, E. V. B. (2012). Análisis Comparativo de modelos de clasificación en el estudio de la deserción universitaria. *Interfases*, 45–82.

Wlodarczak, P. (2019). Machine Learning and its Applications. In *Machine Learning and its Applications*. https://doi.org/10.1201/9780429448782

Zarate Valderrama, A. J. (2019). Universidad nacional de san agustín de arequipa facultad de ingeniería de producción y servicios. In *Universidad Nacional de San Agustín de Arequipa*. Universidad Nacional de San Agustín de Arequipa. http://repositorio.unsa.edu.pe/handle/UNSA/9419

Zea, L. D. F., Reina, Y. F. P., & Molano, J. I. R. (2019). Machine Learning for the Identification of Students at Risk of Academic Desertion. *Communications in Computer and Information Science*, *1011*, 462–473. https://doi.org/10.1007/978-3-030-20798-4_40

ANEXOS

Anexo 1: Ficha Socioeconómica

UNAM

Oficina de Bienestar Universitario
Unidad de Asistencia Social

FICHA SOCIO ECONOMICA DEL ESTUDIANTE

CARRERA PROFESIONAL: .....CÓDIGO Nº: ..... CICLO: _____

DATOS GENERALES DEL ESTUDIANTE:

1.1. IDENTIFICACION CORREO ELECTRONICO: .....

APELLIDOS Y NOMBRES	SEXO	FECHA DE NACIMIENTO	EDAD	ESTADO CIVIL	LUGAR NACIMIENTO DPTO/PROV/DIST.
	F ( ) M ( )	... / ... / ...			

3.1. DIRECCION DOMICILIARIA

CALLE/AVENIDA/PASAJE/JIRON/COMITÉ	Nº	MZ	LT	URB/PJ/CPM/ASOC.	LOCALIDAD DPTO/PROV/DIST.	Fono: Fijo..... Cel.....

3.1. DIRECCION DOMICILIARIA DEL LUGAR DE PROCEDENCIA **(PARA ALUMNOS FORANEOS)**

CALLE/AVENIDA/PASAJE/JIRON/COMITÉ	Nº	MZ	LT	URB/PJ/CPM/ASOC.	LOCALIDAD DPTO/PROV/DIST.	TELEFONO

ANTECEDENTES ACADEMICOS:

2.1. NOMBRE DE LA I. E. DE PROCEDENCIA: ..... Lugar .....

2.2.TIPO			
ESTATAL		PRIVADO/ PARROQUIAL	
Primaria		Primaria	
Secundaria		Secundaria	

2.3.PREPARACION UNIVERSITARIA	
Profesor Particular	
Academia Particular	
CEPRE	
Por su Cuenta	

2.4. MODALIDAD DE INGRESO A LA UNIVERSIDAD	
Admisión Ordinaria	
Admisión Extraordinaria (CEPRE Ley Deporte, Primeros puestos Instituciones Educativas , Traslado Interno Externo.....)	

ASPECTO ECONOMICO:

3.1 DE LA FAMILIA

3.1.1. COMPOSICION FAMILIAR

Nº	Apellidos y Nombres	Edad	Parentesco	Estado Civil	Grado de Instrucción	Titulado /Maestría	Ocupación	Centro Laboral y/o Estudios	Localidad
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									

3.1.2.QUIEN SOSTIENE EL HOGAR	
Padre Madre	
Padre	
Madre	
Familiar Tutor	
El mismo Alumno	

3.1.3.OCUPACION DE QUIEN SOSTIENE EL HOGAR			
A. DEPENDIENTE		B.INDEPENDIENTE	
Sector	Público	Comercio Formal	
	Privado	Comercio Ambulatorio	
Condición Laboral	Nombrado	Por Servicios	
	Contratado	Otro	
Especifique Actividad:		Especifique Actividad:	
C .DESOCUPADO			

3.1.4.MODALIDAD DE INGRESO ECONOMICO	
Mensual	
Quincenal	
Semanal	
Diario	

Usted Trabaja: Si ( ) No ( )

3.1.5. INGRESO ECONOMICO FAMILIAR	
Remuneración del Padre	S/
Remuneración de la Madre	S/
Ingreso del Alumno	S/
Otros Ingresos	S/
TOTAL INGRESO FAMILIAR	S/

¿Dónde?

3.1.6. VIVIENDA DEL ESTUDIANTE	
Con sus Padres	
Con uno de sus Padres	
Como alojado	
Como Cuidante	
Cuarto Alquilado	
¿Cuánto Paga? S/	

3.1.7 CARGA FAMILIAR DE QUIEN SOSTIENE EL HOGAR (CONSIDERAR PADRES, HIJOS Y FAMILIARES MAYORES DE 5 AÑOS)	
01 Persona	
02 Personas	
03 Personal	
04 Personas	
05 Personas	
06 Personas	

3.1.8 N° DE HIJOS QUE CURSAN ESTUDIOS	
NIVEL	CANTIDAD
Primario	
Secundario	
Superior	

ASPÉCTO ALIMENTICIO:

TOMA SUS ALIMENTOS:	
Hogar	
Parientes	
Pensión Costo:.....	
Otros (especifique)	

3.2.DEL ESTUDIANTE

3.2.1 DEPENDENCIA ECONOMICA	
Depende de ambos Padres	
Sólo del Padre	
Sólo de la Madre	
De un familiar	
De si mismo	

3.2.2 RIESGO FAMILIAR	
Hijo de Padres Vivos	
Huérfano de Madre	
Huérfano de Padre	
Huérfano de Padre y Madre	
Vive sólo	

IV. ASPECTO DE VIVIENDA

4.1 DE LA FAMILIA

4.1.1 TENENCIA		4.1.2 TIPO DE CONSTRUCCION		4.1.3 TIPO DE VIVIENDA		4.1.4 SERVICIOS	
Propia		Noble		Independiente		Agua	
Alquilada		Mixto		Edificio Dpto.		Desagüe	
Invasión		Rustico		Conventillo		Luz	
Otro		Precario		Otros		Teléfono	

4.1.5.ESTRUCTURA DE VIVIENDA (Precisar Cantidad)	
N° de Pisos	
N° de Dormitorios	
Cocina	
Baño	
Sala	
Comedor	

4.1.6. EQUIPAMIENTO			
T.V. Color		T.V. B/N	
Radio		Teléfono	
Equipo de Sonido		Ropero	
Plancha		Refrigerador	
Celular		Internet	
Laptop		Biblioteca Personal	
Cable		Computadora	

4.2 DEL ESTUDIANTE FORANEO

Departamento: ..... **Provincia** ..... **Distrito** .....

2.1 TENENCIA		4.2.2 TIPO DE CONSTRUCCION		4.2.3 TIPO DE VIVIENDA		4.2.4 SERVICIOS	
Propia		Noble		Independiente		Agua	
Alquilada		Mixto		Edificio Dpto.		Desagüe	
Invasión		Rustico		Conventillo		Luz	
Otro		Precario		Otros		Teléfono	

4.2.5.ESTRUCTURA DE VIVIENDA (Precisar Cantidad)	
N° de Pisos	
N° de Dormitorios	
Cocina	
Baño	
Sala	
Comedor	

4.2.6. EQUIPAMIENTO			
T.V. Color		T.V. B/N	
Radio		Teléfono	
Equipo de Sonido		Ropero	
Plancha		Refrigerador	
Celular		Internet	
Laptop		Biblioteca Personal	
Cable		Computadora	

V. ASPECTO DE SALUD

5.1 DE LA FAMILIA		5.2 DEL ESTUDIANTE	
Tiene Familiar Enfermo	SI ( ) NO ( )	¿Padece de alguna Enfermedad?	SI ( ) NO ( )
¿Quién es?		Tipo de Enfermedad	
Tipo de Enfermedad		Tiene Seguro Si ( ) No ( )	¿Cuál?

VI. DIAGNÓSTICO SOCIAL: .....

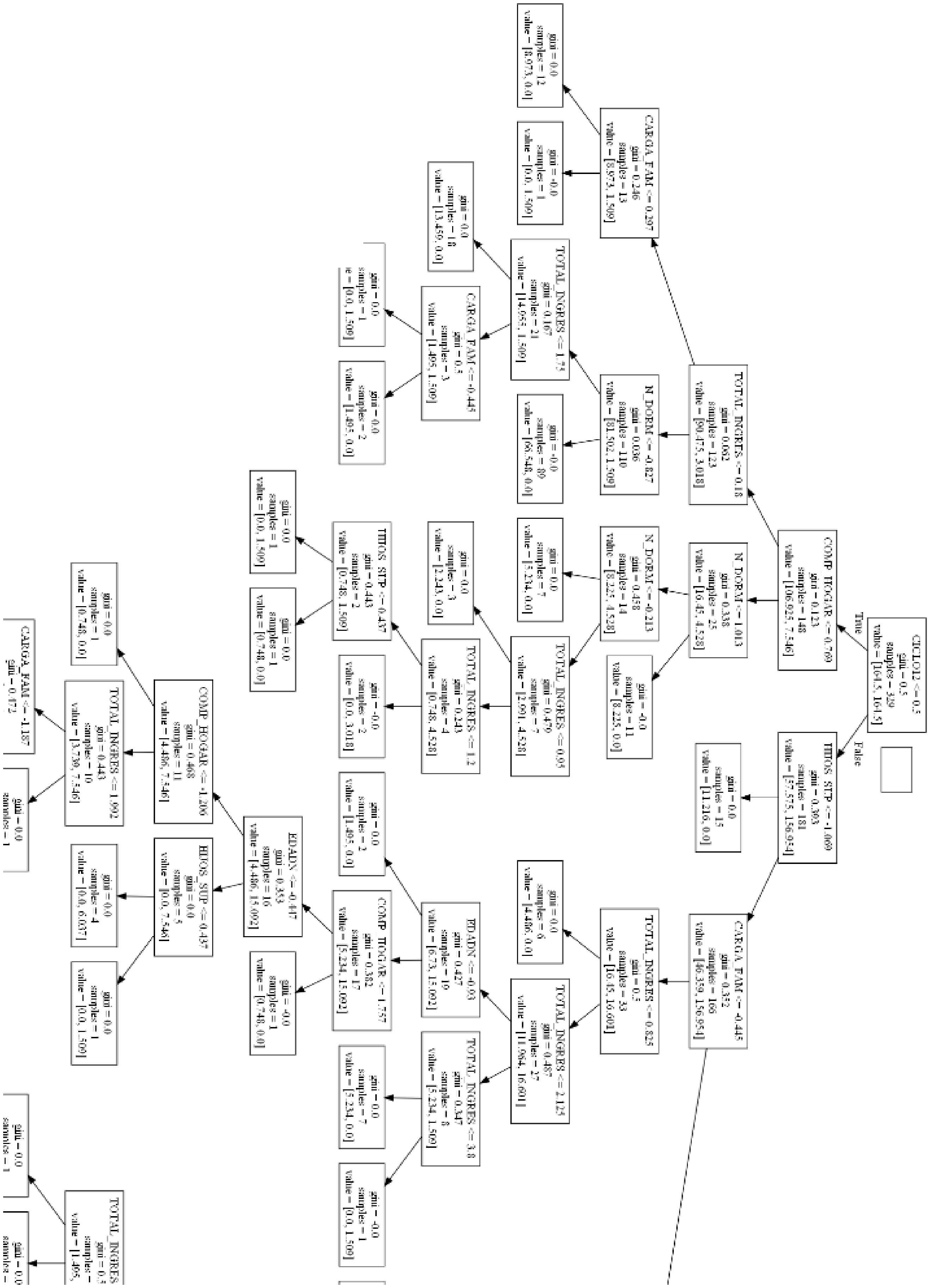
.....

.....

Firma del Estudiante: ..... A.S.....

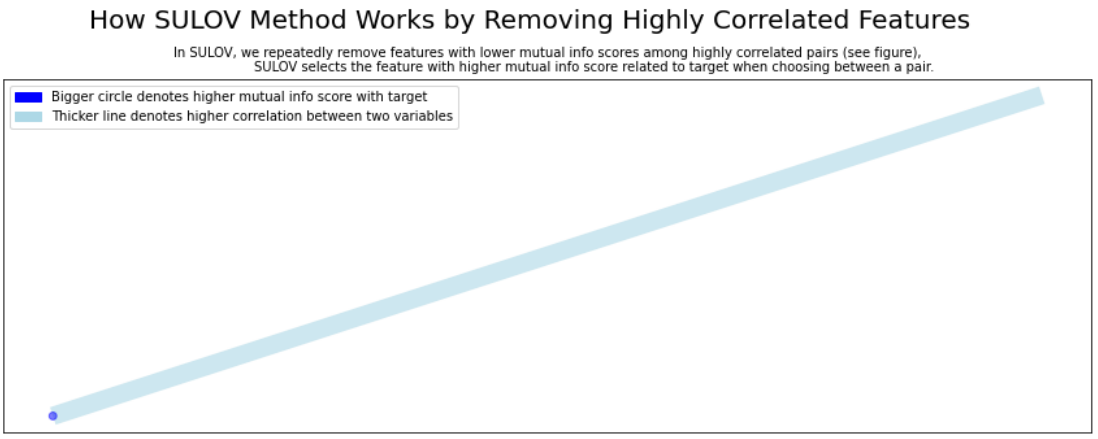
Documento de Identidad N°..... ..... de ..... del 2.....

Anexo 2: Árbol de decisión 9 características



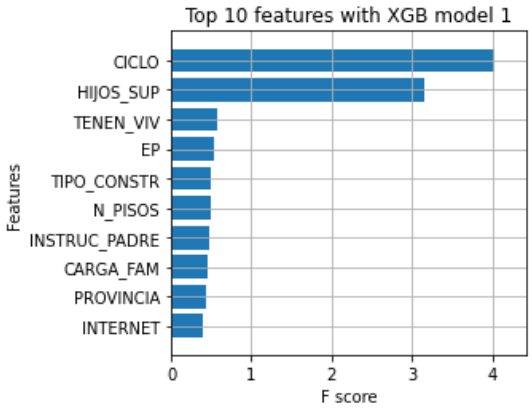
Anexo 3: Reducción por Featurewiz aplicado a 40 características

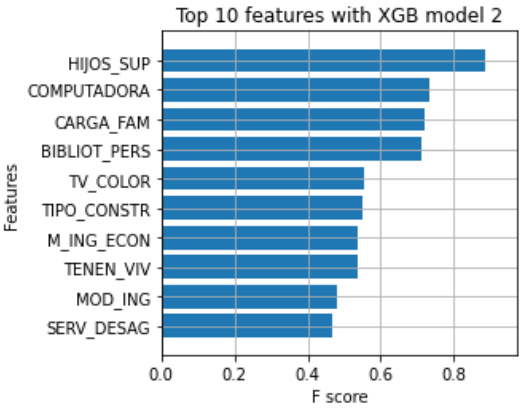
```
#######################################################################################################
############       F A S T   F E A T U R E  E N G G   A N D   S E L E C T I O N ! ########
# Be judicious with featurewiz. Don't use it to create too many un-interpretable features! #
#######################################################################################################
Skipping feature engineering since no feature_engg input...
Skipping category encoding since no category encoders specified in input...
**INFO: featurewiz can now read feather formatted files. Loading train data...
    Shape of your Data Set loaded: (329, 40)
    Loaded train data. Shape = (329, 40)
    Some column names had special characters which were removed...
No test data filename given...
#######################################################################################################
######################### C L A S S I F Y I N G  V A R I A B L E S  ####################
#######################################################################################################
Classifying variables in data set...
    39 Predictors classified...
        No variables were removed since no ID or low-information variables found in data set
No GPU active on this device
    Tuning XGBoost using CPU hyper-parameters. This will take time...
    After removing redundant variables from further processing, features left = 39
No interactions created for categorical vars since feature engg does not specify it
#### Single_Label Binary_Classification problem ####
#######################################################################################################
#####  Searching for Uncorrelated List Of Variables (SULOV) in 39 features #############
#######################################################################################################
    there are no null values in dataset...
    Removing (1) highly correlated variables:
    ['SOST_HOGAR']
```

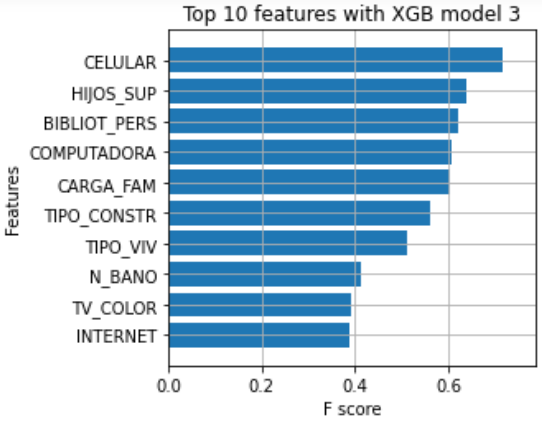


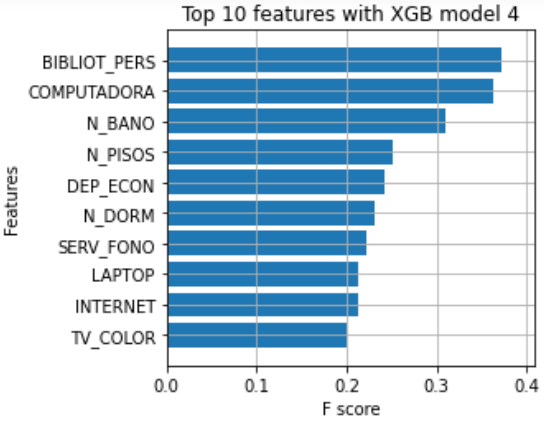

```
Time taken for SULOV method = 6 seconds
    Adding 0 categorical variables to reduced numeric variables  of 38
Final list of selected vars after SULOV = 38
Readying dataset for Recursive XGBoost by converting all features to numeric...
#######################################################################################
#####    R E C U R S I V E   X G B O O S T : F E A T U R E   S E L E C T I O N  #######
#######################################################################################
```

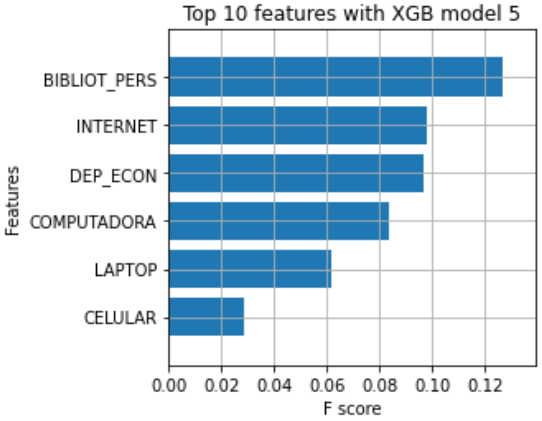
```
    using regular XGBoost
Train and Test loaded into Dask dataframes successfully after feature_engg completed
Current number of predictors = 38
    XGBoost version using 1.6.1 as tree method: hist
Number of booster rounds = 100
        using 38 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 30 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 22 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 14 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 6 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
```

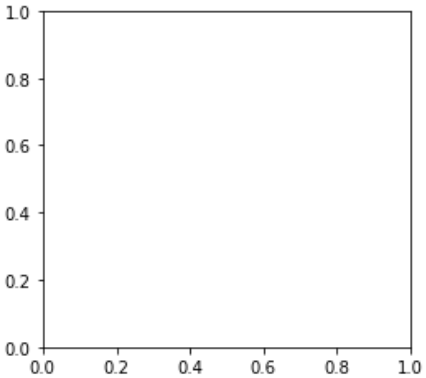




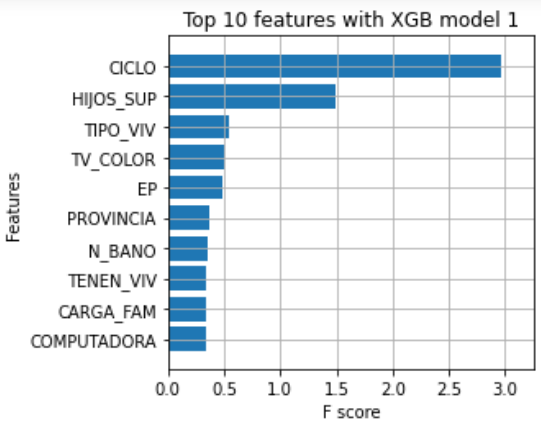


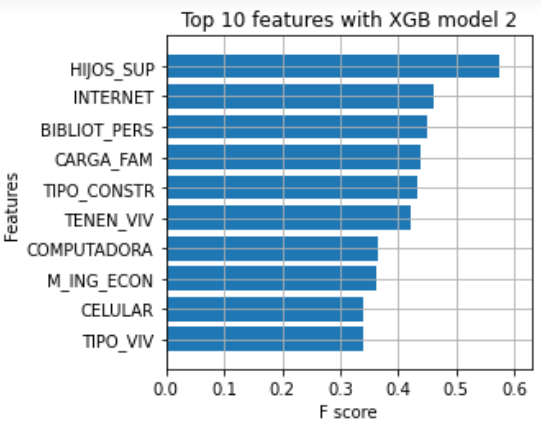


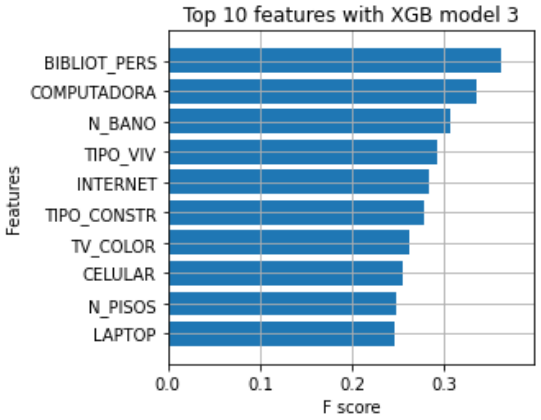


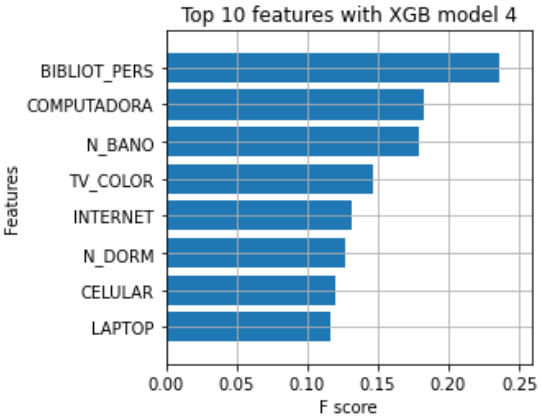


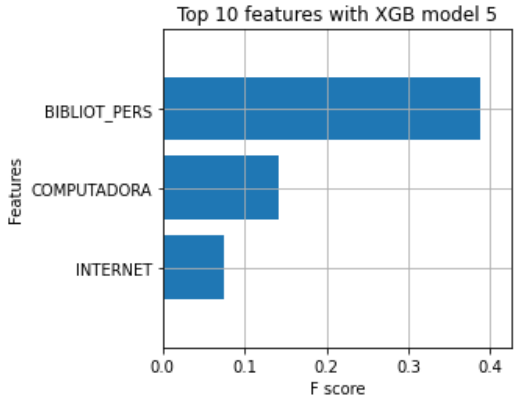
```
           Total time taken for XGBoost feature selection = 3 seconds
###############################################################################
#####         F E A T U R E   S E L E C T I O N   C O M P L E T E D          #######
###############################################################################
Selected 23 important features:
['CICLO', 'HIJOS_SUP', 'TENEN_VIV', 'EP', 'TIPO_CONSTR', 'N_PISOS', 'INSTRUC_PADRE', 'CARGA_FAM', 'PROVINCIA', 'INTERNET',
'COMPUTADORA', 'BIBLIOT_PERS', 'TV_COLOR', 'M_ING_ECON', 'MOD_ING', 'SERV_DESAG', 'CELULAR', 'TIPO_VIV', 'N_BANO', 'DEP_EC
ON', 'N_DORM', 'SERV_FONO', 'LAPTOP']

    Time taken for feature selection = 10 seconds
Returning 2 dataframes: dataname and test_data with 23 important features.

#########################################################################################################
#############       F A S T   F E A T U R E   E N G G   A N D   S E L E C T I O N !  ########
# Be judicious with featurewiz. Don't use it to create too many un-interpretable features! #
#########################################################################################################
Skipping feature engineering since no feature_engg input...
Skipping category encoding since no category encoders specified in input...
**INFO: featurewiz can now read feather formatted files. Loading train data...
    Shape of your Data Set loaded: (329, 24)
    Loaded train data. Shape = (329, 24)
    Some column names had special characters which were removed...
No test data filename given...
#########################################################################################
########################## C L A S S I F Y I N G  V A R I A B L E S  ####################
#########################################################################################
Classifying variables in data set...
    23 Predictors classified...
        No variables were removed since no ID or low-information variables found in data set
No GPU active on this device
    Tuning XGBoost using CPU hyper-parameters. This will take time...
    After removing redundant variables from further processing, features left = 23
No interactions created for categorical vars since feature engg does not specify it
#### Single_Label Binary_Classification problem ####
#########################################################################################
#####  Searching for Uncorrelated List Of Variables (SULOV) in 23 features #############
#########################################################################################
Selecting all (23) variables since none of numeric vars are highly correlated...
Time taken for SULOV method = 0 seconds
    Adding 0 categorical variables to reduced numeric variables  of 23
Final list of selected vars after SULOV = 23
Readying dataset for Recursive XGBoost by converting all features to numeric...
#########################################################################################################
#####    R E C U R S I V E   X G B O O S T : F E A T U R E   S E L E C T I O N  #######
#########################################################################################################
    using regular XGBoost
Train and Test loaded into Dask dataframes successfully after feature_engg completed
Current number of predictors = 23
    XGBoost version using 1.6.1 as tree method: hist
Number of booster rounds = 100
        using 23 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 18 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 13 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 8 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 3 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
```

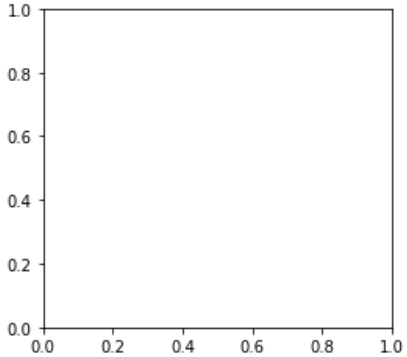










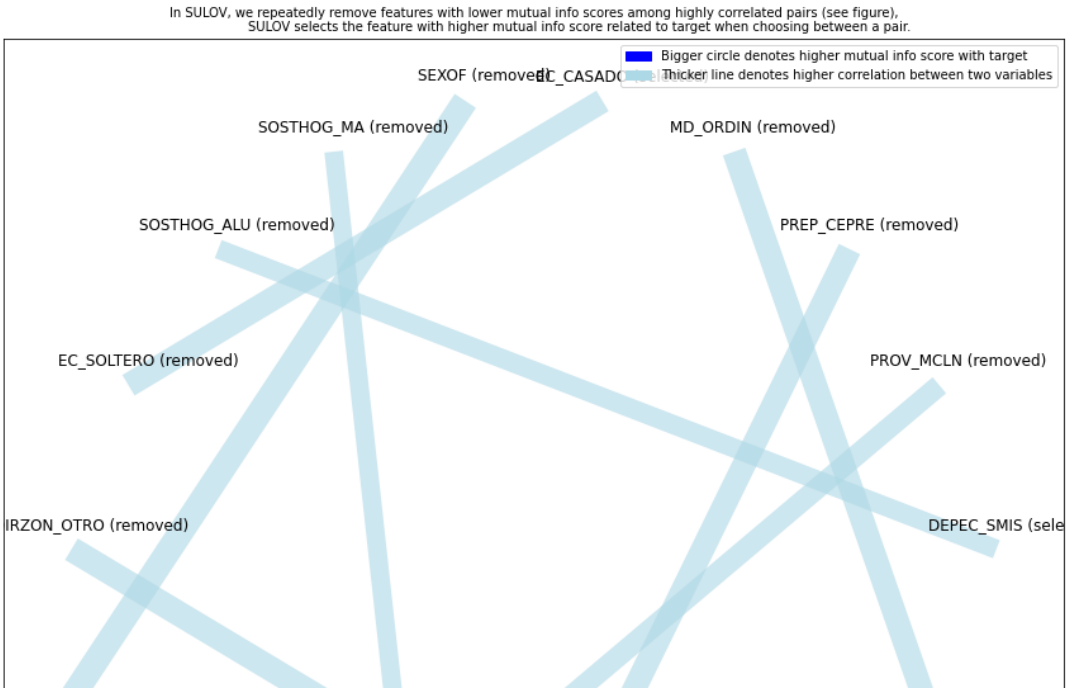


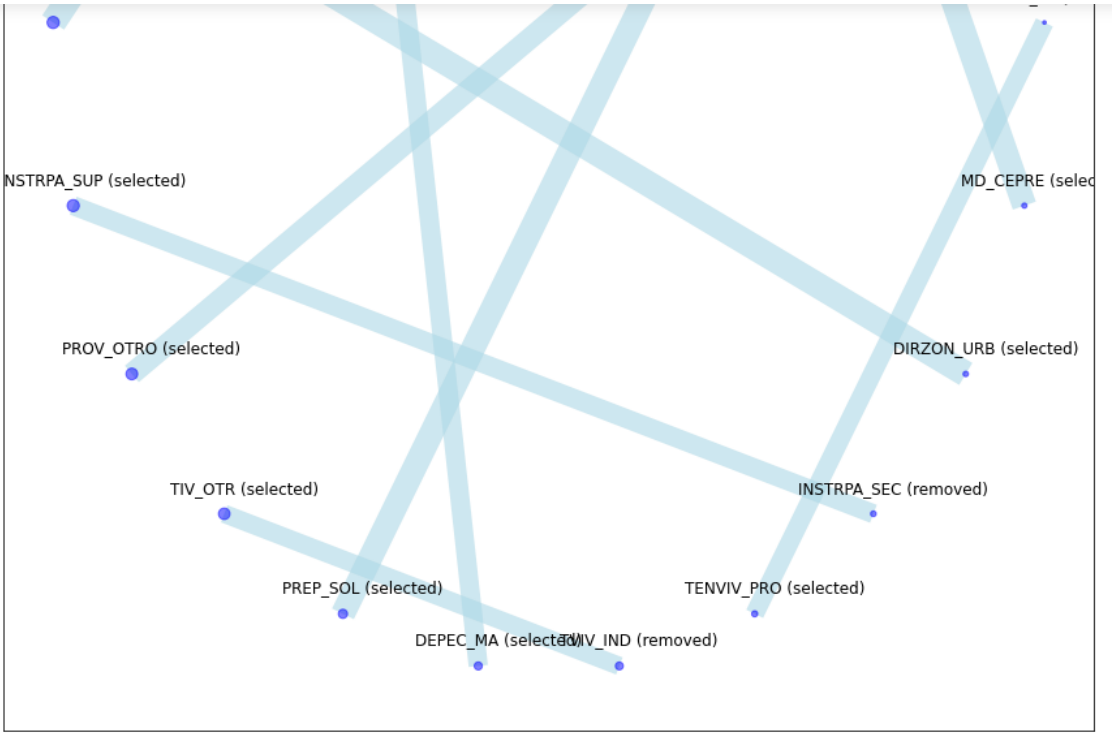

```
Total time taken for XGBoost feature selection = 2 seconds
##############################################################################################
#####          F E A T U R E   S E L E C T I O N   C O M P L E T E D            ########
##############################################################################################

Selected 18 important features:
['CICLO', 'HIJOS_SUP', 'TIPO_VIV', 'TV_COLOR', 'EP', 'PROVINCIA', 'N_BANO', 'TENEN_VIV', 'CARGA_FAM', 'COMPUTADORA', 'INTE
RNET', 'BIBLIOT_PERS', 'TIPO_CONSTR', 'M_ING_ECON', 'CELULAR', 'N_PISOS', 'LAPTOP', 'N_DORM']

    Time taken for feature selection = 3 seconds
Returning 2 dataframes: dataname and test_data with 18 important features.
```

Anexo 4: Salida de Featurewiz aplicado a 94 caracteristicas

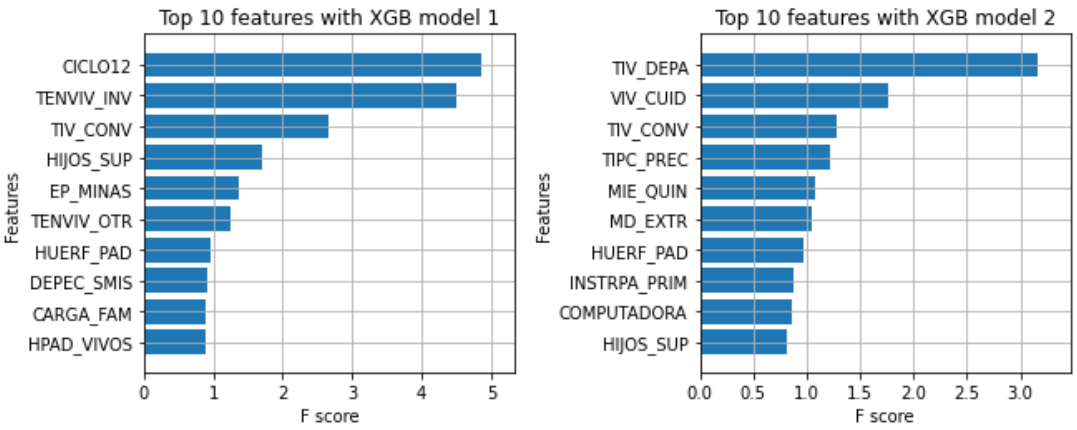
```
#######################################################################################################
#############       F A S T   F E A T U R E  E N G G    A N D    S E L E C T I O N !  ########
# Be judicious with featurewiz. Don't use it to create too many un-interpretable features! #
#######################################################################################################
Skipping feature engineering since no feature_engg input...
Skipping category encoding since no category encoders specified in input...
**INFO: featurewiz can now read feather formatted files. Loading train data...
    Shape of your Data Set loaded: (329, 94)
    Loaded train data. Shape = (329, 94)
    Some column names had special characters which were removed...
No test data filename given...
#######################################################################################################
########################## C L A S S I F Y I N G  V A R I A B L E S  ####################
#######################################################################################################
Classifying variables in data set...
    93 Predictors classified...
        4 variable(s) to be removed since ID or low-information variables
        variables removed = ['EC_SEP_DIV', 'PREP_PP', 'INSTRPA_SIN', 'INSTRMA_SIN']
train data shape before dropping 4 columns = (329, 94)
        train data shape after dropping columns = (329, 90)
    Converted pandas dataframe into a Dask dataframe ...
No GPU active on this device
    Tuning XGBoost using CPU hyper-parameters. This will take time...
    After removing redundant variables from further processing, features left = 89
No interactions created for categorical vars since feature engg does not specify it
#### Single_Label Binary_Classification problem ####
#######################################################################################################
#####  Searching for Uncorrelated List Of Variables (SULOV) in 89 features #############
#######################################################################################################
   there are no null values in dataset...
   Removing (11) highly correlated variables:
   ['SEXOF', 'EC_SOLTERO', 'DIRZON_OTRO', 'PROV_MCLN', 'PREP_CEPRE', 'MD_ORDIN', 'INSTRPA_SEC', 'SOSTHOG_MA', 'SOSTHOG_ALU', 'TENVIV_ALQ', 'TVIV_IND']
```

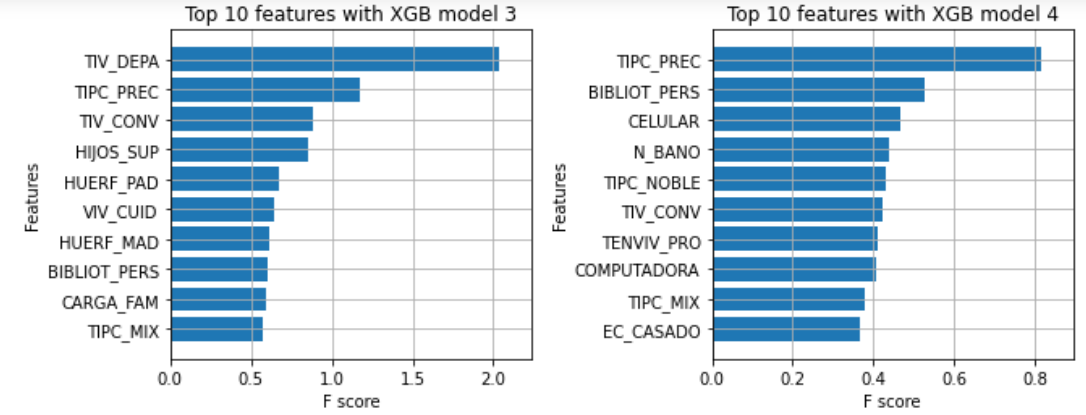
How SULOV Method Works by Removing Highly Correlated Features

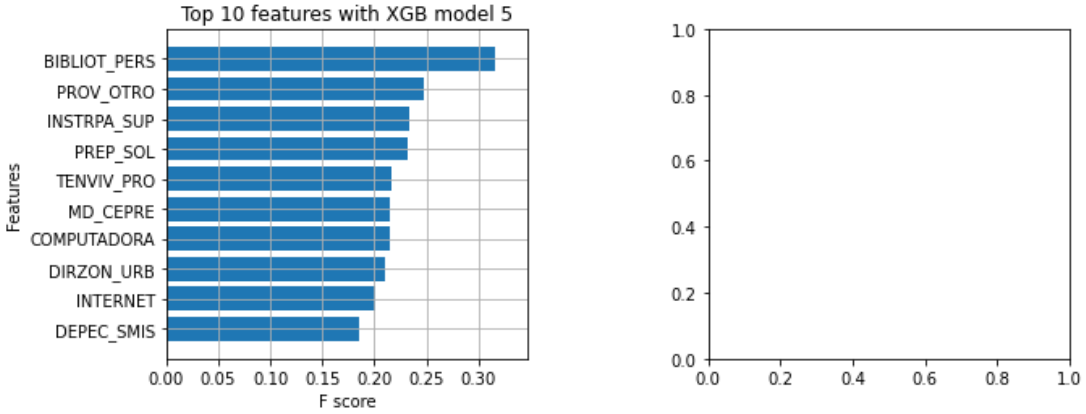





```
Time taken for SULOV method = 1 seconds
    Adding 0 categorical variables to reduced numeric variables  of 78
Final list of selected vars after SULOV = 78
Readying dataset for Recursive XGBoost by converting all features to numeric...
#################################################################################################
#####    R E C U R S I V E   X G B O O S T : F E A T U R E   S E L E C T I O N  #######
#################################################################################################
    using regular XGBoost
Train and Test loaded into Dask dataframes successfully after feature_engg completed
Current number of predictors = 78
    XGBoost version using 1.6.1 as tree method: hist
Number of booster rounds = 100
        using 78 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 62 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 46 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 30 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 14 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
```







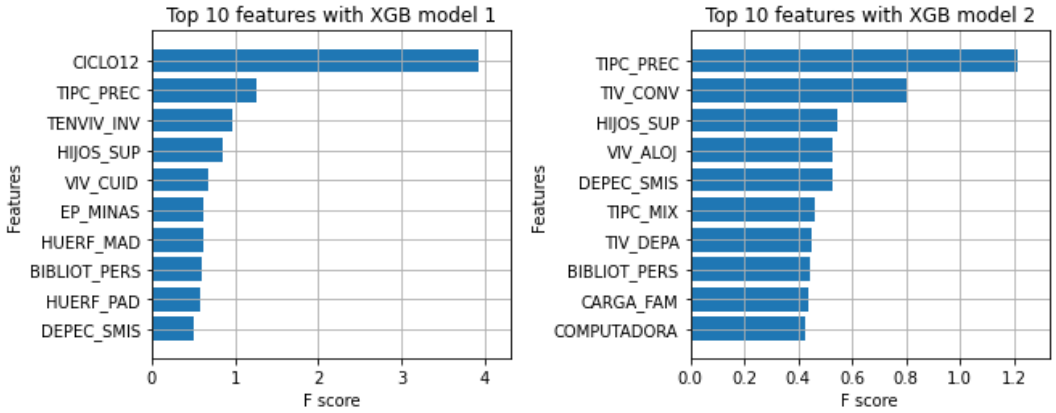

```
            Total time taken for XGBoost feature selection = 3 seconds
################################################################################################
#####          F E A T U R E   S E L E C T I O N   C O M P L E T E D            ########
################################################################################################
Selected 37 important features. Too many to print...

    Time taken for feature selection = 5 seconds
Returning 2 dataframes: dataname and test_data with 37 important features.
```

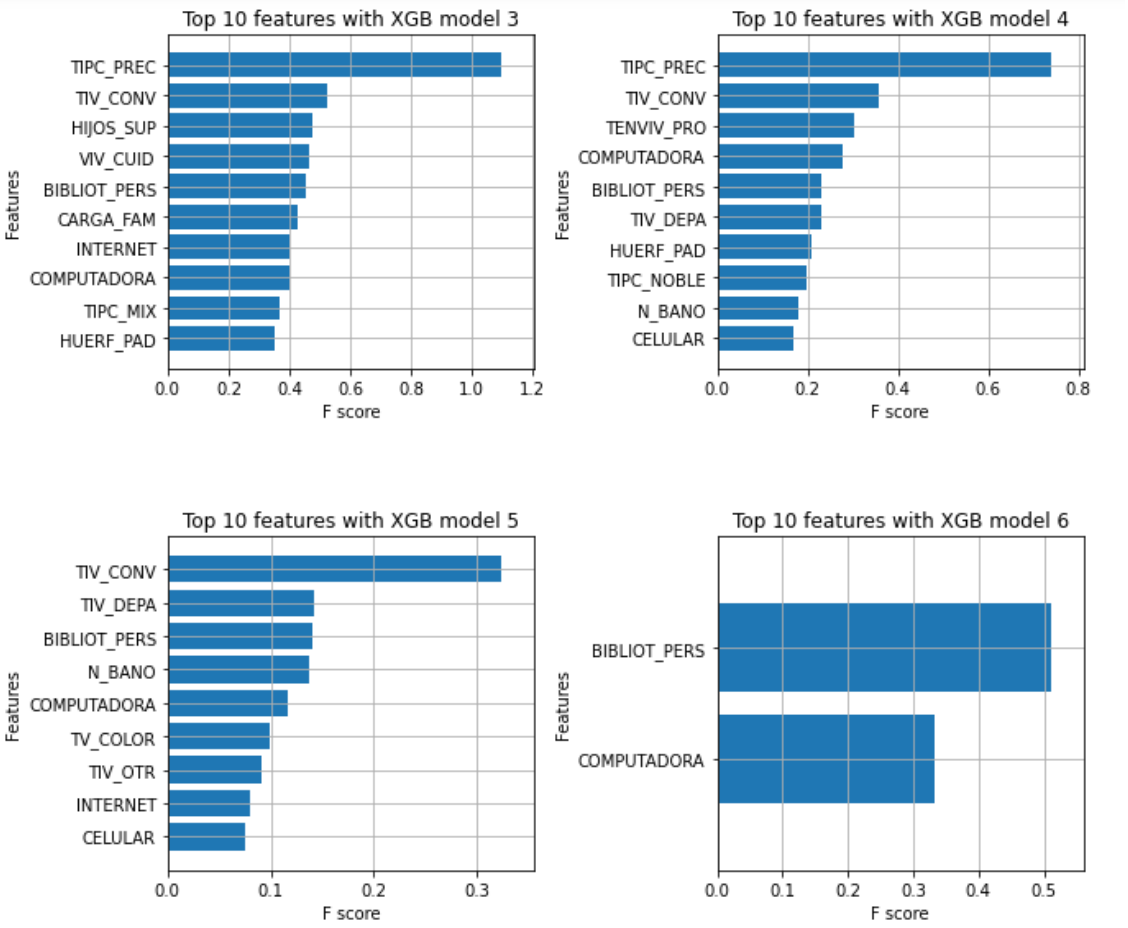
```
: print(features)
```

```
['CICLO12', 'TENVIV_INV', 'TIV_CONV', 'HIJOS_SUP', 'EP_MINAS', 'TENVIV_OTR', 'HUERF_PAD', 'DEPEC_SMIS', 'CARGA_FAM', 'HPAD
_VIVOS', 'PROV_OTRO', 'MIE_SEM', 'BIBLIOT_PERS', 'TIV_DEPA', 'VIV_CUID', 'TIPC_PREC', 'MIE_QUIN', 'MD_EXTR', 'INSTRPA_PRI
M', 'COMPUTADORA', 'VIV_ALOJ', 'TIPC_MIX', 'HUERF_MAD', 'CELULAR', 'TIPC_NOBLE', 'N_BANO', 'TENVIV_PRO', 'EC_CASADO', 'PRE
P_SOL', 'TV_COLOR', 'INSTRPA_SUP', 'MD_CEPRE', 'DIRZON_URB', 'INTERNET', 'TIV_OTR', 'SEXOM', 'DEPEC_MA']
```

Anexo 5: Salida de Featurewiz aplicado a 37 caracteristicas

```
#######################################################################################################
#############       F A S T   F E A T U R E  E N G G    A N D    S E L E C T I O N ! ########
# Be judicious with featurewiz. Don't use it to create too many un-interpretable features! #
#######################################################################################################
Skipping feature engineering since no feature_engg input...
Skipping category encoding since no category encoders specified in input...
**INFO: featurewiz can now read feather formatted files. Loading train data...
    Shape of your Data Set loaded: (329, 38)
    Loaded train data. Shape = (329, 38)
    Some column names had special characters which were removed...
No test data filename given...
#######################################################################################################
######################### C L A S S I F Y I N G  V A R I A B L E S  ####################
#######################################################################################################
Classifying variables in data set...
    37 Predictors classified...
        No variables were removed since no ID or low-information variables found in data set
No GPU active on this device
    Tuning XGBoost using CPU hyper-parameters. This will take time...
    After removing redundant variables from further processing, features left = 37
No interactions created for categorical vars since feature engg does not specify it
#### Single_Label Binary_Classification problem ####
#######################################################################################################
#####  Searching for Uncorrelated List Of Variables (SULOV) in 37 features #############
#######################################################################################################
Selecting all (37) variables since none of numeric vars are highly correlated...
Time taken for SULOV method = 0 seconds
    Adding 0 categorical variables to reduced numeric variables  of 37
Final list of selected vars after SULOV = 37
Readying dataset for Recursive XGBoost by converting all features to numeric...
#######################################################################################################
#####    R E C U R S I V E   X G B O O S T : F E A T U R E   S E L E C T I O N  #######
#######################################################################################################
    using regular XGBoost
Train and Test loaded into Dask dataframes successfully after feature_engg completed
Current number of predictors = 37
    XGBoost version using 1.6.1 as tree method: hist
Number of booster rounds = 100
        using 37 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 30 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 23 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 16 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 9 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 2 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
```

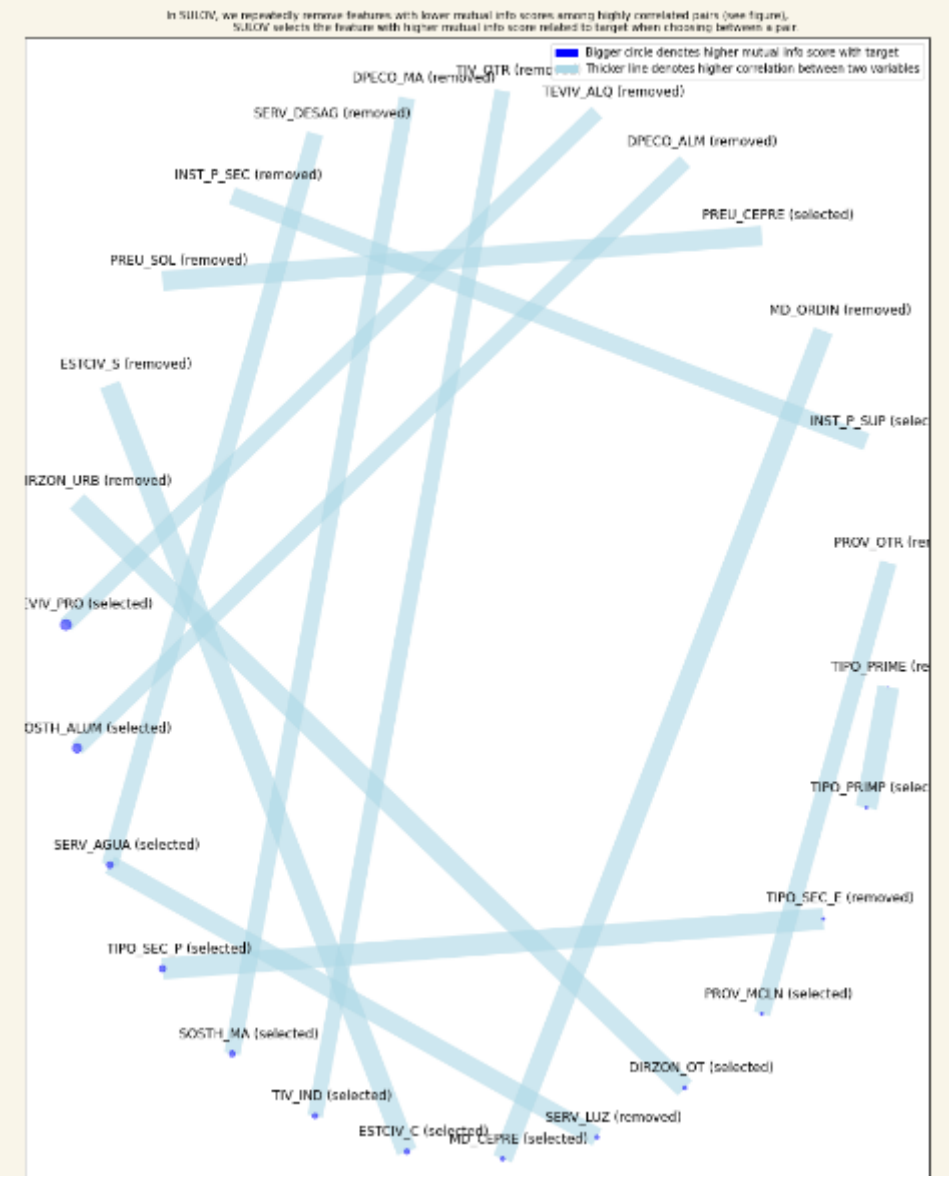





```
           Total time taken for XGBoost feature selection = 2 seconds
#############################################################################################
#####          F E A T U R E   S E L E C T I O N   C O M P L E T E D            #######
#############################################################################################
Selected 23 important features:
['CICLO12', 'TIPC_PREC', 'TENVIV_INV', 'HIJOS_SUP', 'VIV_CUID', 'EP_MINAS', 'HUERF_MAD', 'BIBLIOT_PERS', 'HUERF_PAD', 'DEP
EC_SMIS', 'TIV_CONV', 'VIV_ALOJ', 'TIPC_MIX', 'TIV_DEPA', 'CARGA_FAM', 'COMPUTADORA', 'INTERNET', 'TENVIV_PRO', 'TIPC_NOBL
E', 'N_BANO', 'CELULAR', 'TV_COLOR', 'TIV_OTR']

    Time taken for feature selection = 3 seconds
Returning 2 dataframes: dataname and test_data with 23 important features.
```

```
    there are no null values in dataset...

    Removing (14) highly correlated variables:

    ['ESTCIV_S', 'DIRZON_URB', 'PROV_OTR', 'TIPO_PRIME', 'TIPO_SEC_E', 'PREU_SOL', 'MD_ORDIN', 'INST_P_SEC',
'DPECO_ALM', 'TENVIV_ALQ', 'TIV_OTR', 'SERV_DESAG', 'SERV_LUZ']
```

How SULOV Method Works by Removing Highly Correlated Features



```
Time taken for SULOV method = 1 seconds
    Adding 0 categorical variables to reduced numeric variables  of 90
Final list of selected vars after SULOV = 90
Readying dataset for Recursive XGBoost by converting all features to numeric...
#######################################################################################
#####    R E C U R S I V E   X G B O O S T : F E A T U R E   S E L E C T I O N  #######
#######################################################################################
    using regular XGBoost
Train and Test loaded into Dask dataframes successfully after feature_engg completed
Current number of predictors = 90
    XGBoost version using 1.6.1 as tree method: hist
Number of booster rounds = 100
        using 90 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 72 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 54 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 36 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
        using 18 variables...
            Time taken for regular XGBoost feature selection = 0 seconds
```