UNIVERSIDAD AUTÓNOMA DE ASUNCIÓN
FACULTAD DE CIENCIAS POLÍTICAS, JURÍDICAS Y DE LA COMUNICACIÓN
DOCTORADO EN CIENCIAS DE LA EDUCACIÓN

ESTIMACIÓN DE LA PROBABILIDAD DE EGRESO DE LOS ESTUDIANTES DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL DE VILLARRICA DEL ESPÍRITU SANTO

Mario Damián Vázquez

Asunción, Paraguay

2020

Mario Damián Vázquez

ESTIMACIÓN DE LA PROBABILIDAD DE EGRESO DE LOS ESTUDIANTES DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL DE VILLARRICA DEL ESPÍRITU SANTO

Tesis preparada a la Universidad Autónoma de Asunción como requisito parcial para la obtención del título de Doctor en Ciencias de la Educación

Orientador:

Dr. Fernando Solís Laloux

Asunción, Paraguay

2020

Vázquez, M. 2020. **Estimación de la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo.** Mario Damián Vázquez. 111 Pág. Asunción, Paraguay.

Tutor: Dr. Fernando Solís Laloux

Disertación académica en Doctorado en Ciencias de la Educación .UAA, 2020.

Palabras Clave: Egreso, Universidad, Probabilidad, Regresión Logística, Odds Ratio.

Mario Damián Vázquez

ESTIMACIÓN DE LA PROBABILIDAD DE EGRESO DE LOS ESTUDIANTES DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL DE VILLARRICA DEL ESPÍRITU SANTO

Esta tesis fue evaluada y aprobada en fecha __/__/__ para la obtención del título de Doctor en Ciencias de la Educación por la Universidad Autónoma de Asunción

_____

_____

_____

_____

_____

Asunción, Paraguay

2020

Para mis padres, Antonia y Juan Ramón.

Al Prof. Dr. Fernando Solís Laloux por su invalorable colaboración con este trabajo.

A la Universidad Autónoma de Asunción por la oportunidad de realizar el Doctorado.

A la UNVES por haber facilitado sus datos.

Daría todo lo que sé por la mitad de lo que ignoro.

Descartes

RESUMEN

La gran brecha existente entre la matrícula y el egreso en la educación superior ha generado la necesidad de estimar la probabilidad de egreso de los estudiantes. Como en la Universidad Nacional de Villarrica del Espíritu Santo UNVES las carreras de ingeniería poseen esa mayor brecha, conlleva a analizar cuáles serían los factores que ayuden a identificar en el primer año de la carrera a los estudiantes con mayor o menor probabilidad de egreso de manera a implementar políticas educativas que ayuden a disminuir esa brecha existente entre la matrícula y el egreso.

Para el desarrollo de la investigación se plantea la siguiente pregunta de investigación ¿Cuál es la probabilidad de egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay?, cuyo objetivo general es: Estimar la probabilidad de egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.

De este objetivo general se deducen los siguientes objetivos específicos: Analizar de manera descriptiva el egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018. Determinar las variables académicas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018. Determinar las variables demográficas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018. Determinar el modelo estadístico utilizando la regresión logística con mejor bondad de ajuste que estime la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.

El enfoque utilizado es el cuantitativo, el diseño es no experimental, el alcance la investigación es correlacional. Para la realización de este estudio se ha utilizado la ficha académica del estudiante de ingeniería de la UNVES, contenida en la base de datos del Centro Tecnológico de Informática y Comunicaciones CETIC dependiente de la Dirección General Académica.

Entre los resultados más significativos se pueden mencionar que el sexo del estudiante, su estado civil, el tipo de ingeniería que estudia, el promedio al final del primer semestre de la carrera, la cantidad de materias aprobadas en el primer y segundo semestre de la carrera son las variables que estiman la probabilidad de egreso del estudiante de ingeniería, el modelo obtenido mediante la regresión logística de respuesta binaria tiene una excelente bondad de ajuste identificando a 8 de cada 10 estudiantes con posibles problemas de no egresar. Para la interpretación de los parámetros del modelo predictivo se utilizan los Odds Ratio (OR) de manera a cuantificar el aumento o la disminución de las chances de egreso.

Palabras clave: Egreso, Universidad, Probabilidad, Regresión Logística, Odds Ratio.

ABSTRACT

The large gap between enrollment and graduation in higher education has generated the need to estimate the probability of students leaving. As in the National University of Villarrica del Espiritu Santo UNVES, engineering careers have that greater gap, it leads to analyze what would be the factors that help to identify in the first year of the race students with a greater or lesser probability of graduating in a way to implement educational policies that help reduce the gap between enrollment and graduation.

For the development of the research, the following research question is asked: What is the probability of graduation based on academic and demographic variables of engineering students of the National University of Villarrica del Espiritu Santo of the Republic of Paraguay? general is: Estimate the probability of graduation based on academic and demographic variables of engineering students of the National University of Villarrica del Espiritu Santo of the Republic of Paraguay.

From this general objective the following specific objectives are deduced: To analyze in a descriptive way the progress based on academic and demographic variables of the engineering students of the National University of Villarrica del Espiritu Santo of the Republic of Paraguay cohort 2009-2018. To determine the academic variables that estimates the probability of progress of engineering students of the National University of Villarrica del Espiritu Santo of the Republic of Paraguay cohort 2009-2018. To determine the demographic variables that estimates the probability of progress of engineering students of the National University of Villarrica del Espiritu Santo of the Republic of Paraguay cohort 2009-2018. Determine the statistical model using the logistic regression with better goodness of fit that estimates the probability of progress of the engineering students of the National University of Villarrica del Espiritu Santo of the Republic of Paraguay.

The approach used is the quantitative one, the design is not experimental, and the scope of the research is correlational. For the realization of this study the academic record of the engineering student of the UNVES has been used, contained in the database of the Technological Center of Informatics and Communications CETIC dependent on the General Academic Directorate.

Among the most significant results can be mentioned that the student's gender, marital status, the type of engineering he studies, the average at the end of the first semester of the degree, the number of subjects approved in the first and second semester of the degree are the variables that estimate the probability of graduation of the engineering student, the model obtained through the logistic regression of binary response has an excellent goodness of fit identifying 8 out of 10 students with possible problems of not graduating. For the interpretation of the parameters of the predictive model, the Odds Ratio (OR) is used in order to quantify the increase or decrease of the chances of graduation.

Keywords: Graduate, University, Probability, Logistic Regression, Odds Ratio.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE GRÁFICOS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

AIC	Siglas en inglés de Akaike's Information Criterion.
AUC	Siglas en inglés de Area Under the ROC Curve.
CETIC	Centro Tecnológico de Informática y Comunicaciones de la UNVES.
EEUU	Estados Unidos
gl	Grados de libertad.
Glm	Siglas en inglés de Generalized linear model.
MEC	Ministerio de Educación y Ciencias.
MBA	Siglas en inglés de Master of Business Administration.
MV	Máxima Verosimilitud.
OR	Siglas en inglés de Odds Ratio, Cocientes de ventaja.
p_c	Punto de Corte.
Q_1	Primer cuartil.
Q_2	Segundo cuartil, mediana.
Q_3	Tercer cuartil.
ROC	Siglas en inglés de Receiver Operating Characteristic.
TICs	Tecnologías de la Información y Comunicación.
UNVES	Universidad Nacional de Villarrica del Espíritu Santo.
χ^2	Distribución Chi cuadrado.

INTRODUCCIÓN

El propósito de esta investigación es estimar la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo UNVES en base a variables académicas y variables demográficas utilizando la regresión logística como herramienta estadística para la predicción de la variable dependiente egreso. El egreso de un estudiante puede definirse como aquel estudiante que ha cumplido todos los requisitos académicos para la obtención del título de grado de la carrera de ingeniería. Las variables académicas pueden definirse como las relacionadas a su estado académico dentro de la carrera tales como: ingeniería que estudia, calificación promedio en el primer y segundo semestre y la cantidad de materias aprobadas en el primer curso. Las variables demográficas son las características asociadas a cada estudiante como el sexo, estado civil, ciudad donde reside y el departamento donde reside.

Para el desarrollo de la investigación se plantea la siguiente pregunta de investigación ¿Cuál es la probabilidad de egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay?

Pudiendo formularse el siguiente objetivo general: Estimar la probabilidad de egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.

Del mismo se generan cuatro objetivos específicos:

Analizar de manera descriptiva el egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

Determinar las variables académicas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

Determinar las variables demográficas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

Determinar el modelo estadístico utilizando la regresión logística con mejor bondad de ajuste que estime la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.

Existen varias investigaciones referentes a este tema:

Bobadilla de Almada y la Red Martínez (2017), utilizando tecnología de almacén de datos y minería de datos han evidenciado las características representativas de alumnos de la Facultad Politécnica de la Universidad Nacional del Este de Paraguay con rendimiento académico muy bueno, regular y reprobado. Observaron que el grado educacional de los padres, la actitud general hacia el estudio y la utilización de las TICs inciden en el rendimiento académico de los alumnos y que los promedios generales del segundo semestre correlacionan significativamente con los valores de la situación académica global de los alumnos de los cinco primeros semestres.

En la investigación de Wilson y Hardgrave (1995) con el fin de aumentar la tasa de éxito en un programa de postgrado MBA en EEUU se utilizaron como predictores del rendimiento académico distintos factores como: la nota media durante la carrera, la puntuación en el test de admisión al postgrado, las cartas de recomendación, experiencia profesional, etc. Encontraron que las técnicas de clasificación como el análisis discriminante o la regresión logística son más adecuadas que la regresión lineal múltiple a la hora de

predecir el éxito/fracaso académico, puesto que la regresión múltiple tiende a ignorar los casos extremos en rendimiento.

Baird y Elías (2014), encontraron una escasa relación entre los recursos escolares y el desempeño de los estudiantes. Por otra parte, se estima que el rendimiento promedio de los estudiantes no varía significativamente entre las escuelas, sobre todo después de controlar sus características. Por lo tanto, el estudio concluye que los principales factores asociados a los resultados académicos en el Paraguay están más allá del aula y la escuela. Si bien hace una o dos décadas, Paraguay ha avanzado en la mejora del acceso a la escuela y en la permanencia escolar, los resultados de este estudio ponen de relieve que las barreras para el aprendizaje no son fáciles de superar.

Rodríguez Fontes, Díaz Rodríguez, Moreno Lazo & Bacallao Gallestey (2000), tomando como muestra 114 ingresantes a la carrera de medicina (curso 1991-1992) en la Facultad de Ciencias Médicas Victoria de Girón (Cuba), analizaron la dicotomía éxito-fracaso académico, considerando el éxito como la obtención (durante el 1º año de carrera) de un promedio no inferior a 4 (sobre un máximo de 5). Concluyeron que el índice académico preuniversitario (calificaciones de los últimos ciclos de este nivel) es el predictor más relevante y que el puntaje obtenido en el Examen de Ingreso no es relevante a la hora de predecir el desempeño en el 1º año de carrera.

García Tinisaray (2015), utilizando datos provenientes de una de las universidades ecuatorianas con más número de estudiantes a nivel de educación superior a distancia en Latinoamérica realizó un análisis basado en una regresión logística bivariante binaria y ordinal, el objetivo es analizar el rendimiento académico universitario a través de dos variables de respuesta asociadas, el grado o calificación académica y los créditos universitarios acumulados con cuatro covariables (edad de ingreso, género, región de procedencia y participación en actividades en línea). La población objeto de estudio está

constituida por 410 estudiantes matriculados en una carrera de 5 años equivalente a 282 créditos, cuyo tiempo de estudio comprende el periodo abril 2009 abril 2014, es decir, se realiza un análisis del rendimiento académico al finalizar el período de estudio de una carrera universitaria en donde las variables género y región de procedencia del estudiante no resultan significativas en relación con el rendimiento académico.

Reyes Rocabado y Escobar Flores (2007), realizaron unas predicciones del éxito en el primer semestre de los estudiantes de la carrera de Ingeniería Plan Común, en una cohorte estudiantil de primer año de la Universidad de Antofagasta de Chile. Para realizar los análisis consideraron tres criterios de exigencia para clasificar como exitoso a un estudiante en el primer semestre de su carrera. Aplicando un modelo de regresión logística, los resultados fueron comparados con los del método de análisis discriminante, analizando además su concordancia e índice de predictibilidad.

García Jiménez, Alvarado Izquierdo y Jiménez Blanco (2000), realizaron un estudio sobre alumnos de primer año de Psicología de la Universidad Complutense de Madrid, España. Por un lado utilizan la técnica de Regresión Múltiple para analizar el rendimiento académico y por otro lado, la Regresión Logística para predecir el éxito / fracaso académico, entendido en este caso como la aprobación (o no) de una asignatura del 1º ciclo lectivo; en ambos casos concluyeron que son determinantes el promedio de calificaciones del nivel medio (bachillerato), la participación y asistencia a clases. También destacaron que la capacidad de predicción del Rendimiento Académico de la Regresión Logística es sustancialmente superior al de la Regresión Lineal.

Di Gresia y Porto (2004), enfocaron su análisis en los logros académicos de los estudiantes de la cohorte 2000 de la Facultad de Ciencias Económicas de la Universidad Nacional de La Plata; mediante un Modelo Logit, analizaron la probabilidad que tiene un estudiante de no aprobar ninguna materia luego de dos años de permanencia en el sistema, y

encontraron que dicha probabilidad es más elevada (alrededor del 76%) para un estudiante varón, casado, nacido y residente en La Plata y que trabaja 30 horas a la semana; mientras que dicho riesgo es menor (con probabilidad del 44%) para una mujer, soltera, no nacida en La Plata pero residente allí y que no trabaja.

Vélez van Meerbeke y Roa González (2005), realizaron un estudio sobre los ingresantes 2003 a la Facultad de Medicina de la Universidad del Rosario (privada) en Bogotá, Colombia. Definieron al Rendimiento Académico en términos de éxito/fracaso, este último entendido como la pérdida de materias o el abandono de los estudios; utilizaron la Técnica de Regresión Logística para predecir esta variable y concluyeron que el éxito (fracaso) está asociado principalmente al desempeño académico en el primer semestre de la carrera.

Debido a la gran brecha existente entre la matrícula y el egreso en la educación superior este trabajo de investigación realiza la estimación de la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo UNVES en base a variables académicas y variables demográficas a través de la regresión logística de respuesta binaria. Con los resultados obtenidos se pretende tener un panorama general de los estudiantes de ingeniería, estimar su probabilidad de egreso y así determinar de manera temprana aquellos estudiantes con una baja probabilidad de egreso y así poder implementar políticas educativas que mejoren la calidad académica de la Universidad para poder aumentar el número de egresados, manteniendo la calidad educativa, y así disminuir la brecha existente entre la matrícula y el egreso.

Varios autores confirman que el mejor predictor del rendimiento académico futuro es el rendimiento anterior, como han puesto en evidencia múltiples estudios, Goberna y otros (1987), House, Hurst y Keely (1996), Jiménez (1987), Wilson y Hardgrave (1995).

La estructura de la tesis está organizada de la siguiente manera: en el apartado 1 se analizan los datos existentes a nivel nacional sobre la educación superior relacionadas a la matrícula y al egreso, en el apartado 2 se exponen las investigaciones relacionadas al egreso realizadas dentro y fuera del Paraguay, en el siguiente apartado se enuncian y explican las teoría estadísticas que sustentan la estimación de la probabilidad de egreso como su interpretación, su diagnosis y validación.

Los siguientes apartados corresponden a la Metodología, los Resultados, la Discusión Final y las Recomendaciones.

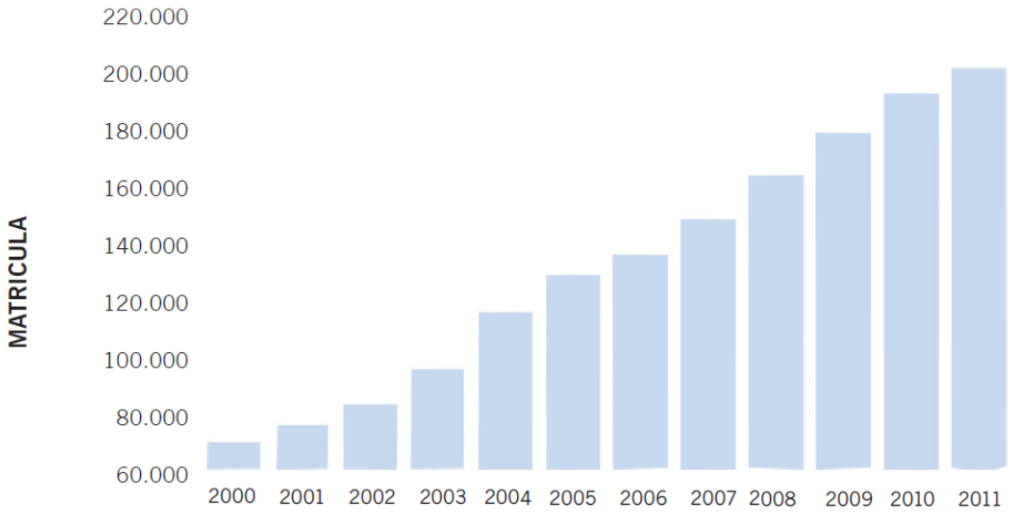
1. ANÁLISIS DE DATOS EN LA EDUCACIÓN SUPERIOR EN EL PARAGUAY

Según el MEC (2013) el Paraguay vive una época de muchas oportunidades de desarrollo, que abarca también a todo el subcontinente de América Latina y el Caribe. Menciona aprovechar esas oportunidades, y menciona que una de las premisas necesarias para el desarrollo económico y la inclusión social (necesarias y deseadas) constituye el fortalecimiento de la educación universitaria. A la vez el MEC recalca que la información disponible en es insuficiente sobre la calidad de las mismas, en lo que llamamos el interior de las universidades, su perfil y su oferta.

1.1. Estudiantes matriculados, incremento anual de la matrícula y distribución de la matrícula en el total de universidades

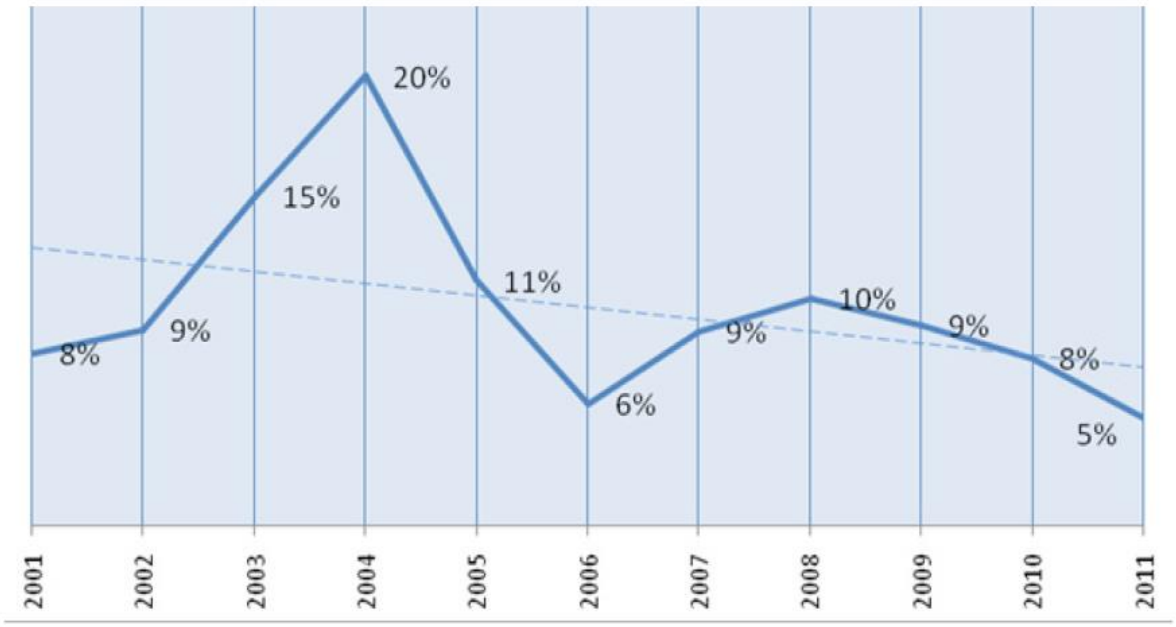
En relación a datos publicados en el 2012 por el MEC, un total de 70.205 estudiantes, hombres y mujeres, estaban matriculados en las universidades del Paraguay en el año 2000. En el año 2011 llegaron a 196.70414 los matriculados. Se dio un aumento acelerado de la matrícula. En el intervalo del 2000 al 2011, la matrícula se ha incrementado en un 180%. El ritmo de crecimiento anual de la matrícula en la década, ha sido cercano al 10%. Con una mayor aceleración entre los años 2000 - 2004, año en que el crecimiento llega a su máxima cifra, 20% anual. La velocidad de incremento desciende entre el 2004 y el 2006, hasta un 6%. Vuelve a acelerarse hasta el 10% en el 2008, y desciende hasta el 5% en al 2011.

Gráfico 1: Matrícula en universidades del Paraguay.



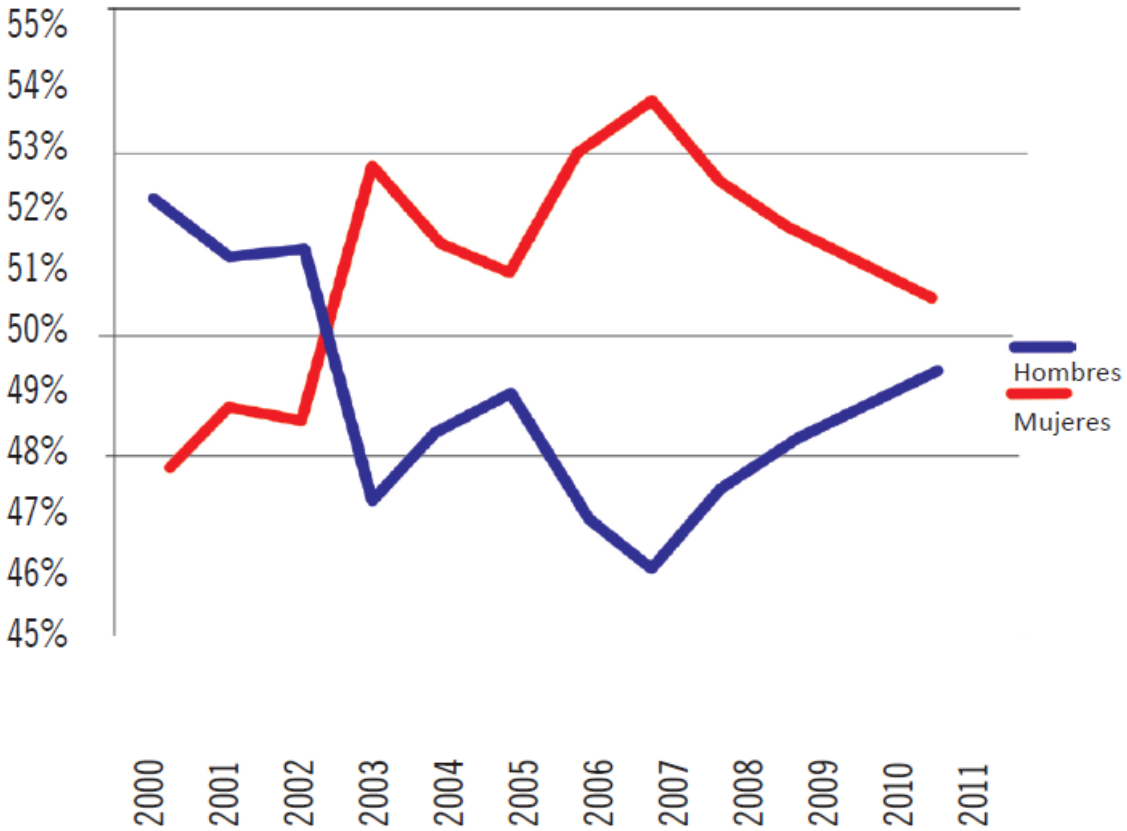
Fuente: Encuesta del Viceministerio de Educación Superior del MEC (mayo de 2012).

Gráfico 2: Incremento anual de la matrícula universitaria en el Paraguay.



Fuente: Encuesta del Viceministerio de Educación Superior del MEC (mayo de 2012).

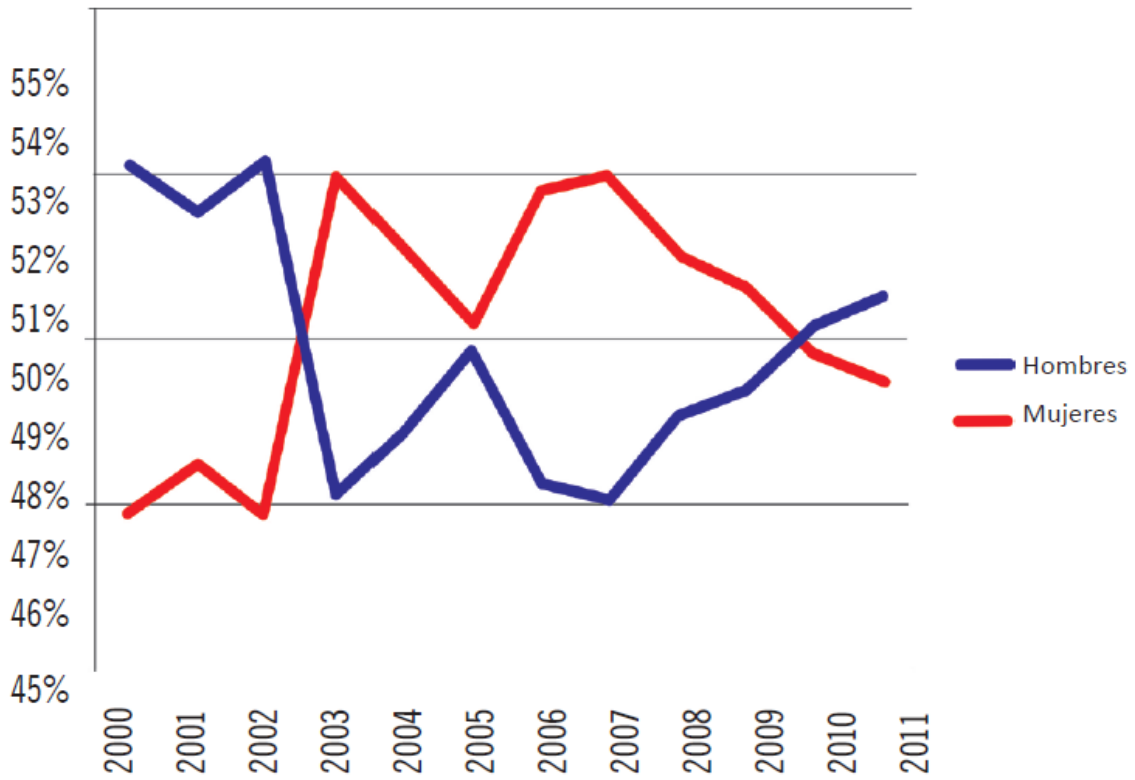
Gráfico 3: Matrícula femenina y masculina en el total de universidades públicas y privadas.



Fuente: Encuesta del Viceministerio de Educación Superior del MEC (mayo de 2012).

En el año 2000, los varones matriculados representaban el 52% de la población estudiantil; las mujeres el 48%. A partir del año 2003 empieza a subir la matrícula femenina y se mantiene como la más elevada hasta el año 2011. En comparación con el porcentaje de población masculina, en el año 2007 hubo mayor proporción de mujeres inscriptas (54% frente a 46% de los varones). El punto máximo alcanzado por la población estudiantil masculina corresponde al año 2000 y el más bajo se registra en el año 2007. En el año 2011 vuelven a acercarse en porcentaje ambas poblaciones (mujeres 51%, varones 49%).

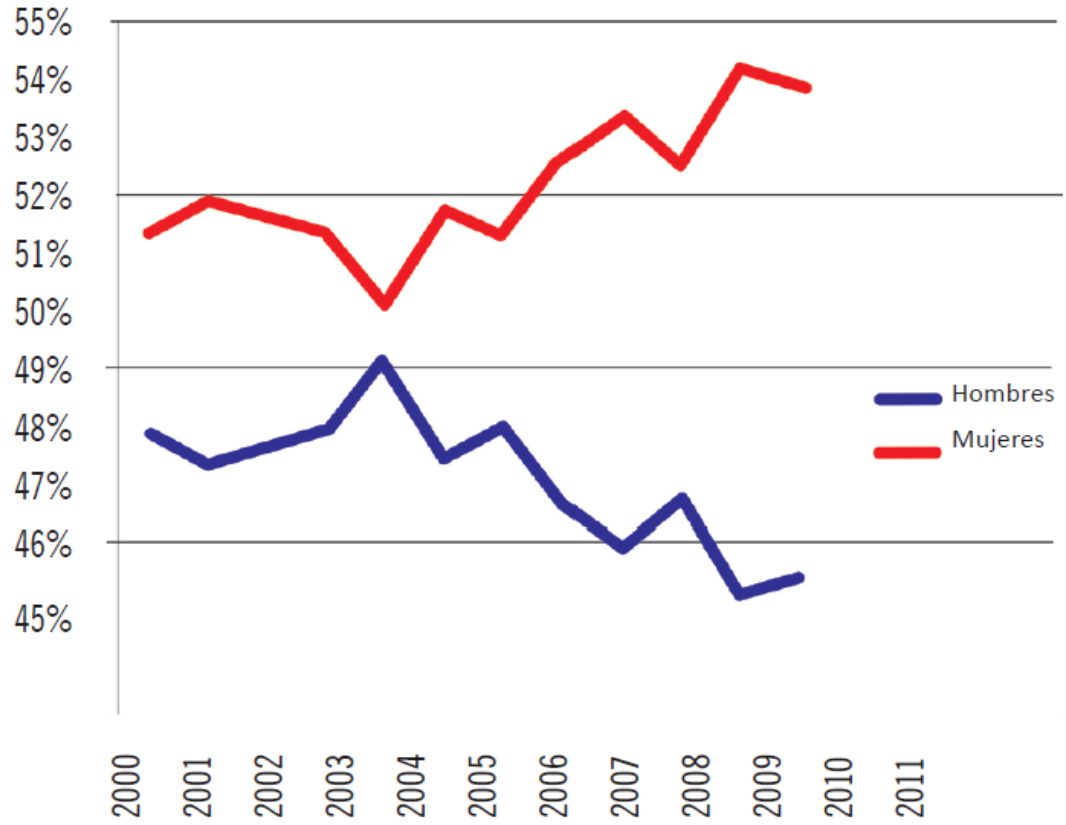
Gráfico 4: Matrícula femenina y masculina en universidades privadas.



Fuente: Encuesta del Viceministerio de Educación Superior del MEC (mayo de 2012).

La distribución varón-mujer de la matrícula en las universidades privadas es no lineal a lo largo del periodo observado: Del 2000 al 2002, la población masculina matriculada superaba en al menos un 10% a la femenina. En el 2003 la relación es exactamente la contraria. En el año 2005 se igualan ambas poblaciones y del 2006 al 2009 la población femenina supera al porcentaje de varones matriculados, alcanzando su punto máximo en el 2007 con el 55%. En el año 2010 ambos grupos vuelven a igualarse y en el 2011 la matrícula masculina vuelve a superar a la femenina, aunque solo en un 2%.

Gráfico 5: Matrícula femenina y masculina en universidades públicas.



Fuente: Encuesta del Viceministerio de Educación Superior del MEC (mayo de 2012).

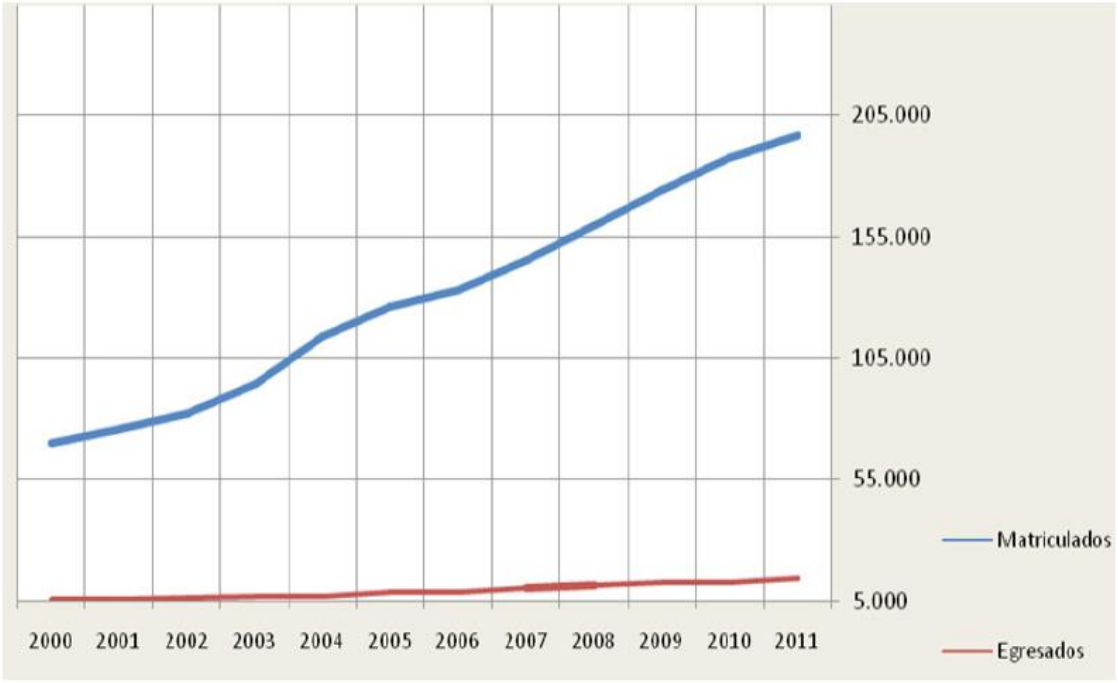
La distribución varón-mujeres en las universidades públicas es muy diferente a la de las privadas. En estas se da una constante de mayor población de mujeres matriculadas y la tendencia es que supere cada vez más a la población de hombres. Así se tiene que desde el año 2008, las mujeres inscriptas superan en un 6% a los varones (con excepción del año 2009, en donde la diferencia es siempre a favor de las mujeres, fue del 4%).

1.2. Relación matrícula - egreso en el total de universidades

En los datos publicados por el MEC (2012) no fue posible hacer una diferenciación entre número de ingresados por año y número de egresados. Se disponen datos acerca de la

matrícula total de las universidades entre los años 2000 y 2011, con lo cual es posible saber cuántos de los que ya están en las universidades egresan. Sería necesario conocer cuántos ingresan por año, cuántos abandonan la titulación y así se podría hablar de una relación entre la tasa de ingreso y egreso.

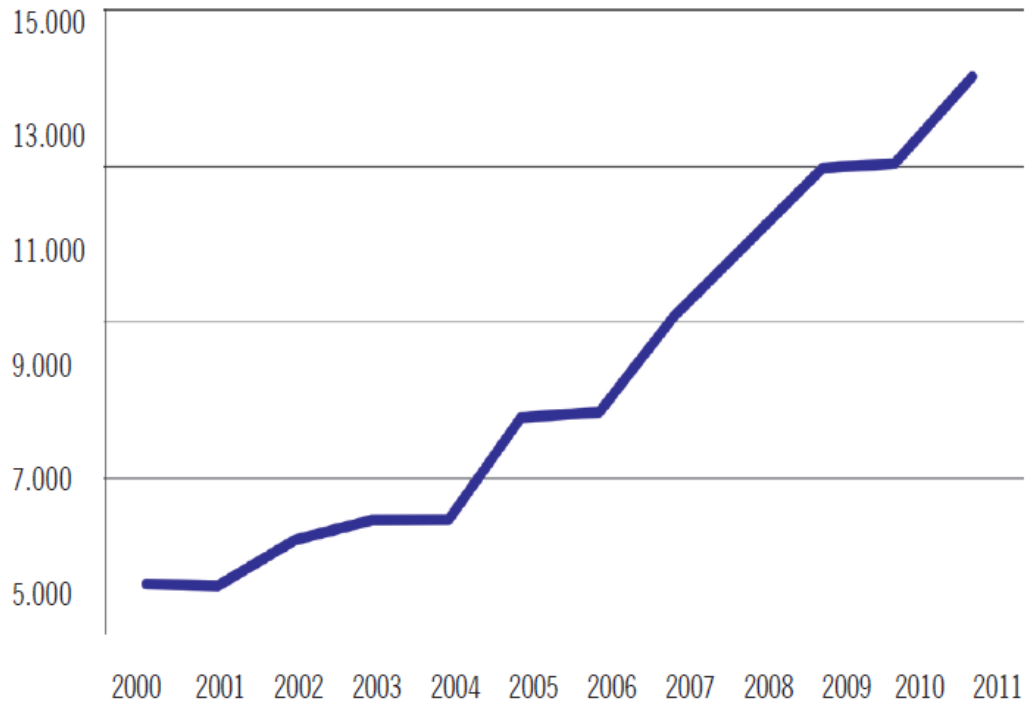
Gráfico 6: Matrícula - egreso en el total de universidades, públicas y privadas.



Fuente: Encuesta del Viceministerio de Educación Superior del MEC (mayo de 2012).

El MEC (2012) señala que la matrícula ascendió en un 180% entre el año 2000 y el 2011. El egreso sin embargo aumentó en un 146%. La diferencia entre el crecimiento de matrícula y de egreso, del año 2000 al año 2012, es del 34%. El MEC hace hincapié que uno de los parámetros de eficiencia de las universidades es la capacidad de retener y titular a sus estudiantes, y menciona que es particularmente importante orientar esfuerzos para reducir la diferencia entre matrícula y egreso.

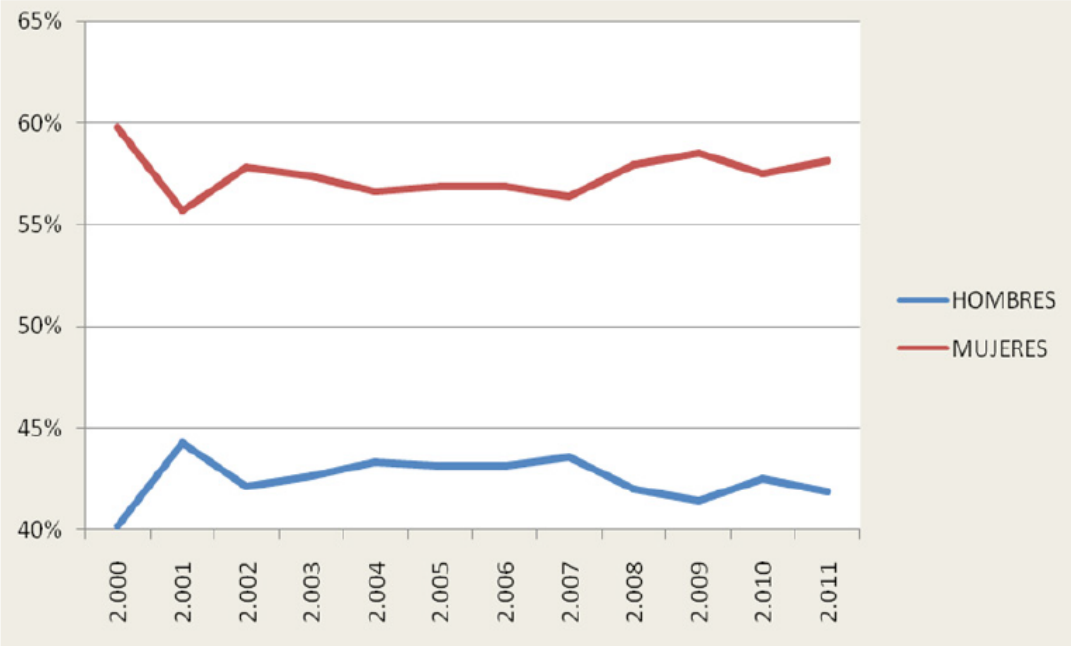
Gráfico 7: Total de estudiantes egresados, universidades públicas y privadas.



Fuente: Encuesta del Viceministerio de Educación Superior del MEC (mayo de 2012).

En el año 2000 eran 5.835 los estudiantes (varones y mujeres) egresados en el total de las universidades del Paraguay. En el año 2011 ascendieron a 14.334 como lo muestra el gráfico existe una tendencia de aumento en la tasa de egreso. De hecho, se constata que en el intervalo de tiempo que va del año 2000 al 2011, la tasa de egreso se ha incrementado en un 146%.

Gráfico 8: Egreso de hombres y mujeres, universidades públicas.



Fuente: Encuesta del Viceministerio de Educación Superior del MEC (mayo de 2012).

En las universidades públicas se tiene que en el año 2000 el porcentaje de mujeres superaba en mayor proporción al de varones egresados (60% vs. 40% respectivamente). Luego, a lo largo del periodo que va del 2001 al 2011, el porcentaje de varones titulados se ubica por debajo del 45% del total, mientras que el de las mujeres por encima del 55% del total de la población estudiantil egresada.

2. EGRESO DE LOS ESTUDIANTES UNIVERSITARIOS

2.1. Investigaciones realizadas en Paraguay

Bobadilla de Almada y la Red Martínez (2017), utilizando tecnología de almacén de datos y minería de datos han evidenciado las características representativas de alumnos de la Facultad Politécnica de la Universidad Nacional del Este de Paraguay con rendimiento académico muy bueno, regular y reprobado. Observaron que el grado educacional de los padres, la actitud general hacia el estudio y la utilización de las TICs inciden en el rendimiento académico de los alumnos y que los promedios generales del segundo semestre correlacionan significativamente con los valores de la situación académica global de los alumnos de los cinco primeros semestres. En este estudio se utilizaron técnicas estadísticas avanzadas y concluyeron que los resultados obtenidos con la aplicación de las técnicas de minería de datos de clústeres, árboles de decisión, asociación y clasificación han evidenciado las características de las clases representativas de alumnos con rendimiento académico regular muy bueno, regular y reprobado. Con el modelo de clústeres se identificó en las agrupaciones formadas las características de los alumnos de acuerdo a su situación final. Con la aplicación de la asociación se logró fijar las variables que consistentemente se asocian en función de las características de la situación final del alumno. Con el modelo de clasificación a través de árboles de decisión, se predijo las características de las clases formadas de acuerdo a la situación final del alumno. Con el modelo de clasificación con regresión se predijo cuál de los promedios de las notas de los cinco primeros semestres influyen en la situación final del alumno en la cual resultaron ser los dos primeros semestres, siendo el segundo semestres con una importancia del 25,97% y el primer semestre con una importancia del 22,13%. Con esta metodología se podría identificar en los primeros semestres del cursado de la carrera, a los posibles alumnos que podrían llegar a desertar de

sus carreras; con la aplicación a tiempo del programa de tutoría y otras medidas a los alumnos identificados se podrían evitar su mal desempeño académico y que llegaran a la deserción.

2.2. Investigaciones fuera de Paraguay

García Tinisaray (2015), utilizando datos provenientes de una de las universidades ecuatorianas con más número de estudiantes a nivel de educación superior a distancia en Latinoamérica realizó un análisis basado en una regresión logística bivariante binaria y ordinal, el objetivo es analizar el rendimiento académico universitario a través de dos variables de respuesta asociadas, el grado o calificación académica y los créditos universitarios acumulados con cuatro covariables (edad de ingreso, género, región de procedencia y participación en actividades en línea). La población objeto de estudio está constituida por 410 estudiantes matriculados en una carrera de 5 años equivalente a 282 créditos, cuyo tiempo de estudio comprende el periodo abril 2009 abril 2014, es decir, se realiza un análisis del rendimiento académico al finalizar el período de estudio de una carrera universitaria en donde las variables género y región de procedencia del estudiante no resultan significativas en relación con el rendimiento académico.

La discusión de resultados se realiza sobre la estimación de la Tabla 19, es el modelo que mejor se ajusta de acuerdo al estadístico desviación (Deviance) y al criterio de información de Akaike (AIC), por lo tanto se puede decir que:

En la parte de estimaciones de efectos fijos y aleatorios se observa que todos los predictores a nivel del estudiante y del aula son estadísticamente significativos, este resultado se obtiene al haber realizado un procedimiento “Stepwise” hacia adelante que permitió eliminar secuencialmente las variables que no tenían significación estadística.

Moral (2006); Acevedo & Rocha (2011) y Pantoja & Alcaide (2013) coinciden en concluir que no existen diferencias entre hombres y mujeres con respecto su rendimiento académico y posterior egreso.

Garzón, Rojas, Riesgo y Pinzón (2010) en una investigación sobre los actores que pueden influir en el rendimiento académico y egreso de estudiantes de Bioquímica que ingresan en el programa de Medicina de la Universidad del Rosario-Colombia mencionan que la región de procedencia del estudiante no es significativa en su relación con el rendimiento académico y egreso.

Porto & Di Gresia (2004) también mencionan que la región de procedencia del estudiante no es significativa sobre la variable respuesta.

En cuanto a las *becas* por nivel de ingresos o méritos académicos, a pesar de que Garzón et al. (2010) determinaron que existe una relación positiva y estadísticamente significativa entre las becas junto con el rendimiento académico y posterior egreso de los estudiantes.

Fonseca Grandón (2018) menciona en su investigación que en relación con las diferencias entre los estudiantes quienes permanecen y culminan su carrera universitaria y los que abandonan y por consiguiente no lo termina, el rendimiento académico marca la principal divergencia. En este sentido explica que cuando las calificaciones obtenidas por los estudiantes durante la primera parte de sus estudios no son las esperadas por ellos, ven fracasados los esfuerzos desplegados en el estudio de las asignaturas del currículum. Aquellos jóvenes que tienen una baja tolerancia a la frustración se ven afectados emocionalmente por los resultados obtenidos y no comprenden que dicha situación es parte del proceso de formación. Así comienzan a evidenciarse los primeros signos de desadaptación, si no se logra superar los conflictos que ello les genera y empieza a circular la intención de abandonar. Al margen de lo anterior, los estudiantes que abandonan no

necesariamente poseen únicamente bajo rendimiento, ya que los que permanecen eventualmente también pueden presentarlo, sin embargo, existen otras variables asociadas que permiten mejorar su desempeño y decidan permanecer.

Reyes Rocabado y Escobar Flores (2007), realizaron unas predicciones del éxito en el primer semestre de los estudiantes de la carrera de Ingeniería Plan Común, en una cohorte estudiantil de primer año de la Universidad de Antofagasta de Chile. Para realizar los análisis consideraron tres criterios de exigencia para clasificar como exitoso a un estudiante en el primer semestre de su carrera. Aplicando un modelo de regresión logística, los resultados fueron comparados con los del método de análisis discriminante, analizando además su concordancia e índice de predictibilidad.

García Jiménez, Alvarado Izquierdo y Jiménez Blanco (2000), realizaron un estudio sobre alumnos de primer año de Psicología de la Universidad Complutense de Madrid, España. Por un lado utilizan la técnica de Regresión Múltiple para analizar el rendimiento académico y por otro lado, la Regresión Logística para predecir el éxito / fracaso académico, entendido en este caso como la aprobación (o no) de una asignatura del 1º ciclo lectivo; en ambos casos concluyeron que son determinantes el promedio de calificaciones del nivel medio (bachillerato), la participación y asistencia a clases. También destacaron que la capacidad de predicción del Rendimiento Académico de la Regresión Logística es sustancialmente superior al de la Regresión Lineal.

Di Gresia y Porto (2004), enfocaron su análisis en los logros académicos de los estudiantes de la cohorte 2000 de la Facultad de Ciencias Económicas de la Universidad Nacional de La Plata; mediante un Modelo Logit, analizaron la probabilidad que tiene un estudiante de no aprobar ninguna materia luego de dos años de permanencia en el sistema, y encontraron que dicha probabilidad es más elevada (alrededor del 76%) para un estudiante varón, casado, nacido y residente en La Plata y que trabaja 30 horas a la semana; mientras

que dicho riesgo es menor (con probabilidad del 44%) para una mujer, soltera, no nacida en La Plata pero residente allí y que no trabaja.

Vélez van Meerbeke y Roa González (2005), realizaron un estudio sobre los ingresantes 2003 a la Facultad de Medicina de la Universidad del Rosario en Bogotá, Colombia. Definieron al Rendimiento Académico en términos de éxito/fracaso, este último entendido como la pérdida de materias o el abandono de los estudios; utilizaron la Técnica de Regresión Logística para predecir esta variable y concluyeron que el éxito (fracaso) está asociado principalmente al desempeño académico en el primer semestre de la carrera.

3. LA REGRESIÓN LOGÍSTICA

3.1. Modelo de Regresión Logística

El desarrollo teórico que se presenta a continuación parte de la base de los principales textos sobre Regresión Logística, como son Hosmer, D. y Lemeshow, S. (2000) y Agresti, A. (2002).

Considere una colección de p variables explicativas con n observaciones en cada una. Sea $x_i = (x_{i1},\ x_{i2},\ \dots\ , x_{ip})$ el vector que contiene las observaciones de cada variable para el i-ésimo individuo, con $i = 1, 2, \dots, n$. Sea X_{ki} la variable explicativa, con $k = 1, 2, \dots\ , p$ para el i–ésimo individuo. Sea Y_i la variable respuesta dicotómica, que toma el valor $y_i = 1$ cuando ocurre dicho suceso, mientras $y_i = 0$ cuando no ocurre el suceso en estudio.

El primer enfoque simple es formular el modelo de regresión:

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ki} + \varepsilon_i \quad ,\ i = 1, 2, \dots\ , n \qquad (1)$$

donde ε_i son los errores aleatorios. Los errores son variables aleatorias independientes y de esperanza cero. Tomando esperanzas matemáticas en la ecuación (1) para $X_{ki} = x_{ki}$, se tiene:

$$\mathrm{E}(\mathrm{Y_i}|x_i) \;=\; \beta_0 \;+ \sum_{k=1}^{p} \beta_k x_{ki} \quad , \; i \;=\; 1, 2, \ldots \;, n \qquad (2)$$

Como la variable Y_i toma los valores 0 o 1, ésta sigue una distribución binomial de parámetros 1 y p_i, que se denota por $Y_i \sim Bin(1, p_i)$, donde p_i se define como la probabilidad de que Y_i tome el valor 1 (ocurre el suceso en estudio) dependiente de cada valor de las p covariables, $p_i = P(Yi = 1|x_i)$.

De esta manera, la esperanza es:

$$\mathrm{E}(\mathrm{Y_i}|x_i) = P(Yi = 1|x_i) \times \; 1 \; + \; P(Yi = 0|x_i) \times \; 0 \; = \; p_i \qquad (3)$$

Por lo tanto, en respuestas binarias, un modelo análogo al de regresión lineal, combinando las ecuaciones (2) y (3), es:

$$p_i \;=\; \beta_0 \;+ \sum_{k=1}^{p} \beta_k x_{ki} \quad , \; i \;=\; 1, 2, \ldots \;, n \qquad (4)$$

Que se denomina Modelo de Probabilidad Lineal, ya que la probabilidad de ocurrencia del suceso cambia linealmente con respecto a los valores de las covariables X_{ki}. Este modelo tiene el problema de que aunque las probabilidades deben estar acotadas en el intervalo [0,1], generalmente, $p_i \in R$, por lo que ya no representaría una probabilidad.

Por lo general, las relaciones entre p_i y las covariables X_{ki} no son lineales, de modo que el cambio en X_{ki} tiene menor impacto cuando p_i está cerca de 0 o 1 que cuando p_i está más cerca de la mitad del rango. Por lo tanto, la relación no es definida en forma lineal.

Por otra parte, como los únicos valores posibles de Y_i son 0 o 1, los errores aleatorios ε_i no siguen una distribución normal y no son homocedásticos, y por tanto, la clasificación mediante la ecuación de relación lineal no es necesariamente óptima.

Todos los problemas presentados hacen que el modelo de probabilidad lineal no sea tan utilizado y, por lo tanto, para poder garantizar que los valores pronosticados estén en el intervalo [0,1], de tal manera que proporcione la probabilidad de ocurrencia del suceso en estudio, se recurre a la transformación de la variable de respuesta. Esto es,

$$p_i = \mathcal{F}\left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}\right) \tag{5}$$

donde $\mathcal{F}$ es una función de distribución, seleccionada debido a sus propiedades deseables: es una función no decreciente y está acotada entre cero y uno.

De esta manera, se toma como $\mathcal{F}$ la función matemática dada por:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki})}} \tag{6}$$

que se denomina función de distribución logística de la que derivan los Modelos de Regresión Logística.

Una transformación de p_i, que es fundamental en la Regresión Logística, es la transformación Logit. Esta transformación se define en término de p_i, como

$$g_i = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} \tag{7}$$

De modo que, al hacer la transformación, se tiene un modelo lineal que se denomina Logit. Así, el modelo de Regresión Logística también se llama Modelo Logit.

De la ecuación (6), se deduce que el modelo de Regresión Logística es, en principio, un modelo de regresión no lineal. Sin embargo, existe una relación lineal en escala logarítmica entre el cociente de probabilidades (la probabilidad de que ocurra un suceso divido por la probabilidad de que no ocurra) y las covariables. Por tanto, al ser una función lineal de las variables explicativas, facilita la estimación y la interpretación del modelo.

Existen diferentes transformaciones. La más extendida, es la transformación Logit, aunque existen otras como la Log–Log y la Probit.

En el modelo anterior, y teniendo en cuenta que la transformación Logit es monótona creciente, se tiene que:

Si X_k es binaria, β_k está asociado al cambio en Y_i debido al cambio en X_k desde su nivel de referencia al alternativo.

Si X_k es categórica con $c > 2$ categorías, en el modelo habrá $c - 1$ *dummies* (binarias auxiliares), una para cada nivel de X_k diferente al de referencia.

Si X_k es continua, β_k está asociado al cambio en Y_i debido a un incremento de una unidad en X_k.

En general, si $\beta_k = 0$, Y_i es independiente de X_k.

3.1.1. Estimación de los parámetros del modelo.

3.1.1.1. Estimación por Máxima Verosimilitud (MV).

Considere una Matriz de Datos de dimensión $(n \times (p + 1))$ definida por:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ n-1 \\ n \end{array} \begin{pmatrix} Y_1 & X_1 & X_2 & \cdots & X_p \\ 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 0 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & x_{n-1,1} & x_{n-1,2} & \cdots & x_{n-1,p} \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

La estimación de los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ de un Modelo de Regresión logística se realiza por medio del método de estimación de Máxima Verosimilitud (MV), que consiste en obtener los estimadores máximo verosímiles que hacen que se maximice la función de verosimilitud asociada a la muestra.

Al considerar una muestra de n observaciones independientes, como la variable Y_i toma los valores 0 o 1, ésta sigue una distribución binomial de parámetros 1 y p_i, que se denota por $Y_i \sim Bin(1, p_i)$.

La función de verosimilitud para una variable binomial, asociada a una muestra de tamaño n es:

$$\mathcal{L}\left(pi \setminus (x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n1}, \dots, x_{np}, y_n)\right) = \prod_{i=1}^{n} [pi]^{yi} (1 - pi)^{1-yi} \tag{8}$$

Dado que el máximo de una función coincide con el máximo de su logaritmo, el valor de los parámetros $\beta_0, \beta_1, \dots, \beta_p$se obtiene maximizando la ecuación (8) o equivalentemente, maximizando:

$$\log\left(\mathcal{L}\left(p_i(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n1}, \dots, x_{np}, y_n)\right)\right) \tag{9}$$

Así, desarrollando el logaritmo en la ecuación (8), se tiene la denominada función soporte, dada por:

$$\log\left(\mathcal{L}\left(p_i(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n1}, \dots, x_{np}, y_n)\right)\right)$$

$$= \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1 - p_i}\right) + \sum_{i=1}^{n} \log(1 - p_i) \tag{10}$$

donde $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ con $i = 1, 2, \dots, n$.

Teniendo en cuenta la ecuación (6) y (7), puede escribirse el segundo miembro de (10) como:

$$\sum_{i=1}^{n} y_i \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}\right) - \sum_{i=1}^{n} \log\left(1 + e^{\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}}\right) \tag{11}$$

Se observa que la ecuación (11) ya no depende de p_i sino de los parámetros de interés $\beta_0, \beta_1, \dots, \beta_p$, entonces puede denotarse:

$$L(\beta_0, \beta_1, .., \beta_p) = \sum_{i=1}^{n} y_i \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} \right) - \sum_{i=1}^{n} \log \left(1 + e^{\beta_0 + \sum_{k=1}^{p} \beta_k x_{ki}} \right) \quad (12)$$

Luego, los estimadores máximos verosímiles $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ para los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ se obtendrán resolviendo el siguiente sistema de $p + 1$ ecuaciones y $p + 1$ incógnitas.

$$\frac{\partial L(\beta_0, \beta_1, .., \beta_p)}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}}{1 + e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}} = 0 \quad (13)$$

$$\frac{\partial L(\beta_0, \beta_1, .., \beta_p)}{\partial \beta_1} = \sum_{i=1}^{n} y_i x_{i1} - \sum_{i=1}^{n} x_{i1} \frac{e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}}{1 + e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}} = 0 \quad (14)$$

$$\frac{\partial L(\beta_0, \beta_1, .., \beta_p)}{\partial \beta_p} = \sum_{i=1}^{n} y_i x_{ip} - \sum_{i=1}^{n} x_{ip} \frac{e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}}{1 + e^{\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}} = 0 \quad (15)$$

Las ecuaciones de verosimilitud no son lineales en los parámetros $\beta_0, \beta_1, \ldots, \beta_p$. Entonces, para hallar los estimadores de los parámetros, se hace uso de un método iterativo como es el de Newton–Raphson. Al utilizar dicho método iterativo, el algoritmo puede escribirse como:

$$\hat{\beta}_t + 1 = \hat{\beta}_t + \left(X^t \widehat{W}_t X \right)^{-1} X^t \left(Y - \widehat{P}_t \right) \quad (16)$$

donde:

- $\left(\widehat{\beta_t} = \hat{\beta}_0, \widehat{\beta_1}, .., \hat{\beta}_p \right)_t$ es la estimación en la *t*-ésima interacción de un vector de $p + 1$ componentes
- *X* es la matriz de diseño que contiene el conjunto de covariables, de orden $n \times (p + 1)$, dado que el modelo contiene una constante.
- $\widehat{W}$ es una matriz diagonal con términos $\hat{p}_i(1 - \hat{p}_i)$
- $\widehat{P}$ una matriz de orden $(n \times 1)$, cuya componente *i*-ésima es $\hat{p}_i$, siendo

$$\hat{p}_i = \frac{1}{1 + e^{-(x_i \widehat{\beta}_t)}} \tag{17}$$

por ende $\hat{p}_i$ se calcula con el valor $\widehat{\beta}_t$

Y es una matriz de orden $(n \times 1)$, cuyos elementos representan los valores de la variable respuesta.

De la ecuación (16) $X^t\left(Y - \widehat{P}_t\right)$ la forma matricial de la primera derivada de $L(\beta_0, \beta_1, \ldots, \beta_p)$ respecto a cada parámetro, conocida como Función de *Score*, mientras que, la segunda derivada, llamada Matriz Informativa o *Hessiana* es $X^t \widehat{W_T} X$.

El proceso iterativo empieza asignando un valor empírico a los parámetros de regresión, en general cero a todos ellos. Luego, en cada iteración $t + 1$ la matriz de nuevos parámetros experimentales resulta de sumar matricialmente un gradiente (∇) a la matriz de los parámetros experimentales del paso anterior. El gradiente es el resultado del cociente entre la forma matricial de la Función *Score* y la Matriz *Hessiana*, que se corresponden con la primera y segunda derivada de la función de verosimilitud, respectivamente.

Son varios los criterios de convergencia del método iterativo utilizado para es- timar los parámetros, pero en todos ellos la idea subyacente es que el algoritmo converge si $\hat{\beta}_{t+1} \cong \widehat{\beta}_t$. Una vez finalizada las iteraciones, la inversa de la Matriz *Hessiana*, $\left(X^t \widehat{W_T} X\right)^{-1}$, obtenida en la ultima iteración, ofrece los valores de varianzas y covarianzas de los parámetros estimados.

3.1.2. Pruebas de Hipótesis sobre los parámetros del modelo.

3.1.2.1. Prueba de razón de verosimilitudes para la significación de varios parámetros del modelo.

Si queremos decidir sobre la significación de diversas covariables en el modelo, podemos usar la prueba de razón de verosimilitudes.

Esta decisión se basa en la prueba de hipótesis

$$H_0: \forall k \in (1, \dots, p): \beta_k = 0$$

$$H_1: \exists k \in (1, \dots, p): \beta_k \neq 0$$

El estadístico de decisión es la diferencia de devianzas entre los modelos, dado por:

$$G = D_{Modelo\ bajo\ H_0} - D_{Modelo\ saturado} = -2 \log \frac{\mathcal{L}(\beta_{H_0}|Y, \boldsymbol{X})}{\mathcal{L}(\beta|Y, \boldsymbol{X})} \sim x_p^2 \qquad (18)$$

Las β_k testadas pueden corresponder a covariables continuas o categóricas.

El criterio de decisión es, si $G \geq \chi^2_{p,\alpha}$ donde α es el nivel de significación y p los grados de libertad de la distribución χ^2, se rechaza la hipótesis nula y se concluye que al menos uno de los coeficientes del modelo es distinto de cero y la variable explicativa correspondiente considerada en el modelo es significativa.

3.1.2.2. Prueba de Wald para la significación de un parámetro del modelo.

Bajo régimen asintótico, se puede usar la prueba de Wald, basada en la distribución normal, para decidir sobre la significación de la asociación entre la covariable X_k y la variable respuesta Y.

Esta decisión se basa en la prueba de hipótesis:

$$\begin{matrix} H_0: \beta_k = 0 \\ \\ H_1: \beta_k \neq 0 \end{matrix} \qquad para\ k = 0{,}1, \dots, p$$

En esta prueba, podemos usar equivalentemente cualquiera de los siguientes estadísticos:

$$z_w = \frac{\hat{\beta}_k}{\sqrt{\widehat{Var}(\beta_k)}} \sim \mathcal{N}(0{,}1) \quad o \quad \chi^2_w = \frac{\widehat{\beta_k^2}}{\widehat{Var}(\beta_k)} \sim \chi^2_1(0{,}1) \qquad (19)$$

siendo $\hat{\beta}_k$ las estimaciones de máxima verosimilitud de su β_k y $\sqrt{\widehat{Var}(\beta_k)}$ su correspondiente desviación estándar.

El criterio de decisión es, si $\chi_w^2 \geq \chi_{1,\alpha}^2$se rechaza la hipótesis nula y se concluye que la *k*-ésima variable explicativa es significativa.

3.2. Medidas de asociación entre variables categóricas.

Sean dos variables categóricas, X e Y , con r y c categorías respectivamente. Una forma útil para presentar los datos es a través de la Tabla de contingencia, que consiste en un cuadro bidimensional con r filas para la variable X y c columnas para la variable Y, donde las diferentes observaciones se asocian a diferentes celdas.

Considérese, primeramente, la Tabla XXX

Tabla 1: Tabla de Contingencia con probabilidades conjuntas π_{ij} y frecuencias observadas n_{ij} con $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, c$

Y / X	Columna 1	Columna 2	…	Columna c	Total
Fila 1	$\pi_{11}(n_{11})$	$\pi_{12}(n_{12})$	...	$\pi_{1c}(n_{1c})$	$\pi_{1\bullet}(n_{1\bullet})$
Fila 2	$\pi_{21}(n_{21})$	$\pi_{22}(n_{22})$	...	$\pi_{2c}(n_{2c})$	$\pi_{2\bullet}(n_{2\bullet})$
.	.	.	...	.	.
.	.	.	...	.	.
.	.	.	...	.	.
Fila r	$\pi_{r1}(n_{r1})$	$\pi_{r2}(n_{r2})$	...	$\pi_{rc}(n_{rc})$	$\pi_{r\bullet}(n_{r\bullet})$
Total	$\pi_{\bullet 1}(n_{\bullet 1})$	$\pi_{\bullet 2}(n_{\bullet 2})$	...	$\pi_{\bullet c}(n_{\bullet c})$	$1(n)$

donde:

π_{ij} : es la probabilidad conjunta en la i–ésima fila y la j–ésima columna

$\pi_{i\bullet}$: es la probabilidad marginal de X en la i–ésima fila

$\pi_{\bullet \mathrm{j}}$: es la probabilidad marginal de Y en la j–ésima columna

n_{ij}: es la frecuencia absoluta en la i–ésima fila y la j–ésima columna

$n_{i\bullet}$: es la sumatoria de las frecuencias absolutas de la i–ésima fila

$n_{\bullet \mathrm{j}}$: es la sumatoria de las frecuencias absolutas de la j–ésima columna

$$n = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{c} n_{\bullet \mathrm{j}}$$

Dado que el coeficiente de correlación de Pearson no puede aplicarse a variables ´ categóricas debido a la ausencia de una métrica asociada a las variables, a continuación se presentan dos medidas de asociación útiles para determinar el grado de asociación o dependencia entre variables categóricas.

3.2.1. Gamma de Goodman y Kruskal.

Goodman, L. y Kruskal, W. (1954) sugieren una medida de asociación para variables categóricas ordinales. Considere un par de individuos o de observaciones clasificados según dos variables cualitativas ordinales. En relación al orden de estos dos individuos en cada una de las variables, el par se puede clasificar como concordante o discordante.

Así, al comparar un par de observaciones, se dice que:

a) Es concordante cuando a mayor categoría en X se corresponde una mayor categoría en Y.
b) Es discordante cuando a mayor categoría en X se corresponde una menor categoría en Y.
c) Es empate cuando hay igualdad en X e Y.

Basándose en la tabla de contingencia (ver Tabla 1), las probabilidades de concordancia y de discordancia para esa pareja de variables son, respectivamente:

$$\Pi_C = 2\sum_{i=1}^{r}\sum_{j=1}^{c} \pi_{ij}\left(\sum_{h>i}\sum_{k>j} \pi_{hk}\right) \qquad (20)$$

$$\Pi_D = 2\sum_{i=1}^{r}\sum_{j=1}^{c}\pi_{ij}\left(\sum_{h>i}\sum_{k<j}\pi_{hk}\right) \tag{21}$$

La cantidad $\Pi_C - \Pi_D$ es una medida de asociación. Valores positivos indican una relación monótona creciente y valores negativos indican una relación monótona decreciente. Los estimadores de Π_C y Π_D son respectivamente:

$$\widehat{\Pi}_C = \sum_{i}\sum_{j} n_{ij}\left(\sum_{h>i}\sum_{k>j} n_{hk}\right) \tag{22}$$

$$\widehat{\Pi}_D = \sum_{i}\sum_{j} n_{ij}\left(\sum_{h>i}\sum_{k<j} n_{hk}\right) \tag{23}$$

Teniendo en cuenta que una pareja no está condicionada en ambas variables, se tiene que la probabilidad de concordancia es $\frac{\Pi_C}{H_C + H_D}$ y la probabilidad de discordancia es $\frac{\Pi_D}{H_C + H_D}$, la diferencias entre ambas probabilidades es:

$$\gamma = \frac{\Pi_C - \Pi_D}{H_C + H_D} \tag{24}$$

llamada Gamma de Goodman y Kruskal, siendo su estimador muestral:

$$\hat{\gamma} = \frac{\widehat{\Pi}_C - \widehat{\Pi}_D}{\widehat{\Pi}_C + \widehat{\Pi}_D} \tag{25}$$

$\gamma \in [-1,1]$ y su interpretación es similar al coeficiente de correlación de Pearson.

3.2.2. Tau de Goodman y Kruskal.

Los mismos autores, también sugieren una medida de asociación para variables categóricas nominales. Al no existir un orden natural entre las categorías, no tiene sentido intentar medir una relación de monotonía. Por esto, las medidas de asociación para variables nominales se basan en el concepto de varianza de Y explicada por X, en un paralelismo con el concepto de coeficiente de determinación, pero dada la imposibilidad de utilizar la

varianza por falta de métrica, se asume la existencia de cierta función $V(Y)$ como medida de variabilidad válida para variables categóricas, también se asume la misma función $V\ (Y \mid X = i)$ como medida de variabilidad de Y condicionada al i–ésimo valor de X, cuya esperanza se puede calcular como:

$$\mathbb{E}[V(Y|X)] = \sum_i \pi_i\, V(Y|X = i) \tag{26}$$

Así, se tiene una medida de reducción proporcional en la variabilidad de Y, definida por:

$$\Delta = \frac{V(Y) - \mathbb{E}[V(Y|X)]}{V(Y)} \tag{27}$$

y una medida de variación, definida por:

$$V(Y) = \sum_j \pi_{\bullet \mathrm{j}}\left(1 - \pi_{\bullet \mathrm{j}}\right) = 1 - \sum_{\mathrm{j}} \pi_{\bullet \mathrm{j}}^2 \tag{28}$$

luego, a partir de la ecuaciones (26) (27) y (28) se obtiene la Tau de Goodman y Kruskal, definida por:

$$\tau = \frac{\sum_i \sum_j \frac{\pi_{\mathrm{ij}}^2}{\pi_{\mathrm{i}\bullet}} - \sum_j \pi_{\bullet \mathrm{j}}^2}{1 - \sum_j \pi_{\bullet \mathrm{j}}^2} \tag{29}$$

siendo su estimador muestral:

$$\hat{\tau} = \frac{\sum_i \sum_j \frac{n_{ij}^2}{n_{i\bullet}} - \sum_j n_{\bullet j}^2}{1 - \sum_j n_{\bullet j}^2} \tag{30}$$

$\tau \in [0,1]$ y su interpretación es similar al coeficiente de determinación.

3.3. Selección de Variables.

3.3.1. Selección paso a paso (Stepwise).

Según Hosmer, D. y Lemeshow, S., (2000), los criterios para la inclusión de una variable en un modelo pueden variar de un problema a otro y de una disciplina científica a otra. Sin embargo, todos los criterios comparten el mismo enfoque para la construcción de modelos estadísticos, el cual implica buscar el modelo con parsimonia, consistente en un modelo que ajuste bien a los datos con el menor número de variables posibles, logrando de esta manera un equilibrio entre complejidad y precisión. Entre los principales procedimientos de selección de variables se tiene el método de selección paso a paso (Stepwise), el cual engloba una serie de procedimientos de selección automática de variables significativas, ya sea para la inclusión o exclusión de las mismas en el modelo de forma secuencial, basado únicamente en criterios estadísticos. Este método combina la selección hacia adelante (Forward) para incluir una nueva variable y la selección hacia atrás (Backward) para la eliminación de una variable. La selección o eliminación de variables a partir de un modelo se basa en un test estadístico que comprueba la importancia de la variable, y, o bien incluye la variable en cuestión, o bien la excluye sobre la base de una regla de decisión fija. La importancia de una variable se define en términos de una medida de significación del estadístico, donde el estadístico que se usa depende de los supuestos del modelo. En Regresión Logística los errores siguen una distribución binomial, y la importancia se evalúa a través de la prueba de razón de verosimilitudes.

De esta manera, en los pasos en los que se prueba seleccionar una variable, Salazar, A., (2008) propone realizar contrastes condicionales de razón de verosimilitudes con el modelo del paso anterior como hipótesis nula y cada uno de los nuevos modelos como hipótesis alternativa. Aquellos contrastes cuyos p-valores sean inferiores al nivel de

significación establecido determinarán las posibles variables a ser incluidas en el modelo en ese paso. Entre todas, se elegirá la que mejore el modelo, en el sentido de reducir la devianza, con un p-valor adecuado.

Con respecto a los pasos en los que se prueba eliminar una variable, los contrastes tienen como hipótesis nula el modelo que surgió en el último paso y como hipótesis alternativas los modelos resultantes de eliminar cada una de las variables introducidas en los pasos anteriores al último. En esta ocasión, las posibles variables de ser eliminadas serán aquellas cuyos modelos proporcionen un p-valor que supere el nivel de significación establecido, que por otra parte debe ser mayor que la nivel de significación fijado para la entrada de variables, de tal modo a evitar la posibilidad de incluir y eliminar la misma variable en pasos sucesivos. Hosmer, D. y Lemeshow, S., (2000) discuten que la elección del nivel de significancia igual 0,05 como criterio de inclusión de una variable es demasiado estricto, debido a que logra excluir variables importantes para el modelo, por lo que recomiendan la elección del nivel de significación entre 0,15 y 0,20.

Por lo tanto, en cada paso se realizan varios contrastes, tanto de inclusión de variables como de eliminación y el proceso continúa hasta que los contrastes dejen de ser significativos, es decir, que no se incluyan más variables, ni se elimine ninguna de las que entraron.

3.3.2. Parsimonia.

Una posible medida de parsimonia (equilibrio entre complejidad y precisión) es el Criterio de Información Akaike, conocido como el índice AIC. El criterio que sigue el procedimiento Stepwise puede basarse en esta medida dada por:

$$\mathrm{AIC} = -2\left[\log\left(\mathcal{L}(b)\right) - p\right] \tag{31}$$

donde $\mathcal{L}$ es la verosimilitud y p es el número de parámetros en el modelo. Según este criterio, sería preferible el modelo con menor valor de AIC.

3.4. Interpretación de los parámetros del modelo.

En la formulación del modelo se tiene una serie de coeficientes que son los parámetros, a saber:

β_0: la ordenada en el origen

$(\beta_1, \beta_2 \ \ldots \ , \beta_p)$: las pendientes, donde p es el número de variables explicativas

A partir de estos parámetros pueden calcularse los denominados cocientes de ventaja (OR), que serán de mucha utilidad a la hora de interpretar el modelo.

Para ayudar a interpretar los coeficientes de Regresión Logística se define OR como el cociente de probabilidades entre que ocurra un suceso respecto a que no ocurra. En efecto, de la ecuación (7) se deduce que:

$$OR_i = \frac{pi}{1 - pi} = e^{\beta_0} \prod_{k=1}^{p} e^{\beta_k x_{ki}} \tag{32}$$

Suponga que se tienen dos individuos con valores iguales en todas las variables menos en una. Sean $x_i = (x_{i1}, \ldots, x_{ih}, \ldots, x_{ip})$ el vector que contiene los valores de cada variable para el *i*-ésimo individuo y $x_k = (x_{k1}, \ldots, x_{kh}, \ldots, x_{kp})$ el vector para el *k*-ésimo individuo, donde los valores de las variables son las mismas en ambos vectores, excepto en la variable h, donde $x_{ih} = x_{kh} + 1$. Entonces, el cociente de *OR* para estos dos individuos es:

$$\frac{OR_i}{OR_k} = e^{\beta_h} \tag{33}$$

3.4.1. Interpretación en términos de OR.

La interpretación de estos parámetros varía ligeramente según la naturaleza de la variable que le acompaña. En el caso trivial en que todos los coeficientes β_k del modelo fuesen cero, la variable Y_i sería independiente de las variables explicativas, siendo así la constante β_0 el valor del logaritmo de la ventaja de respuesta p_i para un individuo que tiene valor cero en todas las variables explicativas.

Una ventaja del enlace por transformación Logit respecto a otras transformaciones (por ejemplo, Log-Log o Probit) es que, bajo tal enlace, existe una estrecha relación entre el parámetro β_k del modelo y la OR como medida de asociación entre X_k e Y_i, como se observa a continuación.

a) Caso de una covariable continua**:** si X_k es una variable continua (por ejemplo, peso, estatura, nivel de exposición a rayos UV, etc.), estamos asumiendo que la variación de Y_i en la escala logarítmica de su OR es lineal respecto a X_k. Así, un incremento de r unidades en X_k, manteniendo fijas el resto de las covariables del modelo, implica una estimación de OR de Y_i igual a:

$$\widehat{OR}_{\Delta X_k=\bar{r}}(Y) = e^{\bar{r}\widehat{\beta}_k} \tag{34}$$

Obsérvese que el efecto anterior no depende del valor inicial de X_k, solo del tamaño de su variación r.

b) Caso de una covariable binaria: suponga que X_k es una variable explicativa binaria en el Modelo de Regresión Logística, entonces partiendo de la ecuación (7), se obtiene:

$$\widehat{OR}_{X_k}(Y) = e^{\widehat{\beta_k}} \tag{35}$$

c) Caso de una covariable categórica con más de 2 niveles: si X_k es una variable categórica con $c > 2$ niveles, en el modelo habrá $c - 1$ coeficientes asociados a esa variable, uno por cada nivel de la variable diferente al de referencia.

La interpretación de cada coeficiente se realiza exactamente igual que en el caso de una binaria, comparando en este caso el nivel asociado al coeficiente con el nivel de referencia.

d) Caso de dos covariables binarias: suponga que X_j y X_k son dos covariables explicativas binarias en el Modelo de Regresión Logística, entonces partiendo de la ecuación (7), se obtiene:

$$\widehat{OR}_{X_j,X_k}(Y) = e^{\beta_j+\beta_k} \tag{36}$$

El resultado en la ecuación (36) implica la multiplicidad de la OR, esto es, aplicando logaritmos a dicha ecuación se tiene:

$$\begin{aligned}\log\left[\widehat{OR}_{X_j,X_k}(Y)\right] &= \hat{\beta}_j + \hat{\beta}_K \\ &= \log\left\{\widehat{OR}_{X_j}(Y)\right\} + \log\left\{\widehat{OR}_{X_k}(Y)\right\} \\ &= \log\left\{\left[\widehat{OR}_{X_j}(Y)\right]\left[\widehat{OR}_{X_k}(Y)\right]\right\} \\ \widehat{OR}_{X_j,X_k}(Y) &= \left[\widehat{OR}_{X_j}(Y)\right]\left[\widehat{OR}_{X_k}(Y)\right] \qquad (37)\end{aligned}$$

e) Constante del modelo: fijando el vector de covariables $X = 0$ en la ecuación del modelo, se tiene:

$$P\left(Y|\widehat{X = 0}\right) = \frac{1}{1 + e^{-\widehat{\beta_0}}} \tag{38}$$

El valor proporcionado por la ecuación (38) se interpreta como la probabilidad de $Y = 1$ para un individuo para el cual, todas las covariables categóricas toman su valor de referencia.

3.5. Bondad de ajuste del modelo.

La bondad de ajuste del modelo surge de la necesidad de conocer cuan efectivamente el modelo ajustado describe la variable respuesta, es decir, de evaluar si los valores predichos por el modelo son una precisa representación de los valores observados. Así un modelo no presenta un buen ajuste si la variabilidad residual es grande. En caso de que el modelo presente un mal ajuste, no puede ser utilizado para efectuar predicciones ni extraer conclusiones. Hosmer, D. y Lemeshow, S., (2000) discuten algunas de las medidas de bondad de ajuste global más utilizadas en estudios de Regresión Logística, las cuales se describen a continuación:

3.5.1. Estadístico χ^2 de Pearson y la Devianza.

En el Modelo Logit los residuos estandarizados se definen por:

$$\hat{e}_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} \tag{39}$$

Si el modelo es adecuado, los residuos son variables de media cero y varianza uno que sirven para hacer el diagnóstico del modelo. Se puede realizar el contraste global de la bondad de ajuste del modelo mediante el estadístico de Pearson definida por:

$$\chi^2 = \sum_{i=1}^{n} \hat{e}_i^2 \tag{40}$$

Este estadístico minimiza la sumatoria de los cuadrados de los residuos, y, si el modelo es correcto se distribuye asintóticamente como una χ^2 con $\big(n-(p+1)\big)$ grados de libertad, siendo $(p+1)$ es el número de parámetros en el modelo.

Las desviaciones de las observaciones o pseudoresiduos, definidas por:

$$d_i = -2[y_i \log p_i + (1-y_i)\log(1-p_i)] \tag{41}$$

que aparecen en la maximización de la función de verosimilitud, son utilizadas frecuentemente, en lugar de los residuos de Pearson. Un contraste de razón de verosimilitudes global del modelo se puede realizar con las siguientes hipótesis:

H_0 : El modelo es adecuado

(es decir, las probabilidades pueden calcularse con $p+1$ parámetros).

H_1 : El modelo no es adecuado.

(esto es, las n-probabilidades son libres).

El contraste de la razón de verosimilitudes se reduce al estadístico desviación global:

$$D = -2\sum_{i=1}^{n}[y_i \log p_i + (1-y_i)\log(1-p_i)] \tag{42}$$

Este estadístico es llamado Devianza y, bajo la hipótesis nula, se distribuye asintóticamente como una χ^2 con $\big(n-(p+1)\big)$ grados de libertad.

La regla de decisión en ambos casos es, si $\chi^2 \geq \chi^2_{n-(p+1),\alpha}$ o $D \geq \chi^2_{n-(p+1),\alpha}$ se rechaza H_0 con un nivel α de significación y se concluye que el modelo no es el adecuado.

En general, para poder aplicar el test basado en la Devianza, como para el estadístico χ^2, tiene que verificarse que el número de observaciones para cada combinación de las variables explicativas sea grande.

3.5.2. Test de Hosmer-Lemeshow.

Para poder asumir la distribución χ^2, en los dos test anteriores, debe cumplirse que el 80% de las frecuencias estimadas bajo el modelo sean mayores que cinco y, a su vez, todas mayores que uno. Hosmer, D. y Lemeshow, S., (2000) proponen que cuando esto no es posible, se puede hacer uso de un estadístico que lleva sus nombres. Para la construcción de este estadístico, se agrupan las variables explicativas en g grupos o clases (los autores recomiendan 10 grupos basados en los deciles de las probabilidades estimadas $\hat{p}_i$). Sean n_j el número total de observaciones en el g-ésimo grupo, O_j el número de respuestas $Y = 1$ para el j-ésimo grupo. El estadístico está dado por:

$$\chi^2_{HL} = \sum_{j=1}^{g} \frac{\left(O_j - E_j\right)^2}{v_j} H_0 \sim \chi^2_{g-2} \tag{43}$$

Donde:

$E_j = n_j \hat{\pi}_j$

$v_j = n_j \hat{\pi}_j (1 - \hat{\pi}_j)$

$\hat{\pi}_j$: es el promedio de las probabilidades estimadas en el j-ésimo grupo, es decir la frecuencia esperada.

La hipótesis nula en este test es que el modelo propuesto ajusta al conjunto de datos observados, por lo tanto, cuanto mayor sea el valor de χ^2_{HL} peor será el ajuste del modelo.

3.5.3. Matriz de Confusión.

La Matriz de Confusión, también conocida como Tabla de Clasificación, se utiliza para evaluar la capacidad de discriminación del modelo ajustado como un indicador de bondad de ajuste. Esta Matriz resulta de la clasificación cruzada de la variable respuesta, Y_i, con una variable dicotómica cuyos valores se derivan de las probabilidades estimadas.

Considere la siguiente tabla, en él, se tiene:

Tabla 2: Matriz de Confusión

E_i / O_i	$Y = 1$	$Y = 0$	Total
$Y = 1$	n_{11}	n_{12}	$n_{1\bullet}$
$Y = 0$	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

n_{11}: es el número de observaciones de respuesta $Y = 1$ correctamente clasificadas.

n_{12}: es el número de observaciones de respuesta $Y = 1$ incorrectamente clasificadas.

n_{21}: es el número de observaciones de respuesta $Y = 0$ correctamente clasificadas.

n_{22}: es el número de observaciones de respuesta $Y = 0$ incorrectamente clasificadas.

Entonces, los índices para medir la bondad de ajuste son:

a) Tasa de aciertos o precisión: es el cociente entre las predicciones correctas y el total de predicciones:

$$a = \frac{n_{11} + n_{22}}{n} \tag{44}$$

b) Sensibilidad: es la razón entre los valores 1 correctos y el total de valores 1 observados. En otras palabras, es la probabilidad de clasificación correcta ($Y = 1$).

$$s = \frac{n_{11}}{n_{11} + n_{12}} \tag{45}$$

c) Especificidad: es la razón entre la frecuencia de valores 0 correctos y el total de valores 0 observados. En otras palabras, es la probabilidad de clasificación correcta ($Y = 1$)

$$e = \frac{n_{22}}{n_{21} + n_{22}} \tag{46}$$

d) Tasa de errores: es el cociente entre las predicciones incorrectas y el total de predicciones:

$$E = \frac{n_{21} + n_{12}}{n} \tag{47}$$

e) Tasa de falsos ceros: es la proporción entre la frecuencia de valores 0 incorrectos y el total de valores 0 observados:

$$E_O = \frac{n_{21}}{n_{21} + n_{22}} \tag{48}$$

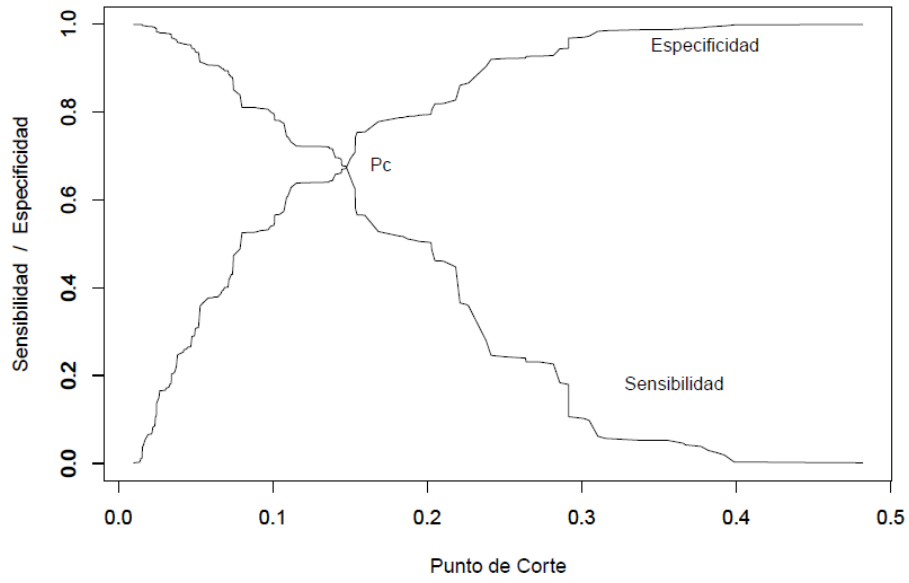
f) Tasa de falsos unos: es la razón entre los valores 1 incorrectos y el total de valores 1 observados:

$$E_C = \frac{n_{12}}{n_{11} + n_{12}} \tag{49}$$

Para clasificar a los individuos se fija un punto de corte (p_c) tal que si la probabilidad estimada por el modelo para un individuo es mayor, se clasifica como $Y = 1$, en caso contrario se clasifica como $Y = 0$. Aunque muchas veces el punto de corte se toma como 0,5. Hosmer, D. y Lemeshow, S., (2000), sugieren que si el objetivo es elegir un punto de corte óptimo para los fines de clasificación, se puede seleccionar un punto de corte que maximiza tanto la sensibilidad como la especificidad. Esta elección se facilita a través de un

gráfico como el que se muestra en la Figura 1 donde se observa la opción óptima para un punto de corte donde aproximadamente la sensibilidad y especificidad se intersecan. Una vez seleccionado el nivel de umbral, se obtiene los valores esperados según las probabilidades estimadas (E_i) y dado que los valores reales de Y_i son conocidos (O_i) basta con mirar si la bondad del ajuste (contabilizar el porcentaje de aciertos) es elevada o no. Como cualquier tipo de regla predictiva está sujeta a errores, habrá ceros que se clasifiquen incorrectamente como unos y viceversa.

Figura 1: Ilustración de Sensibilidad y Especificidad versus Punto de Corte de Hosmer–Lemeshow.



3.5.4. La curva ROC.

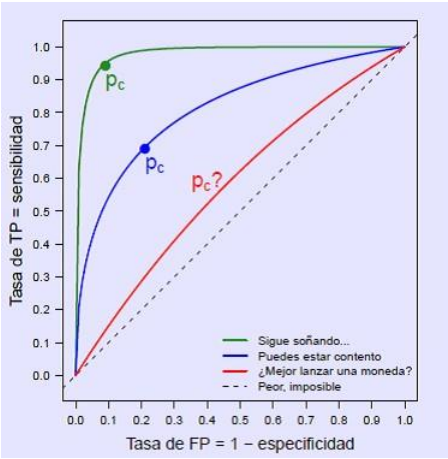
ROC es el acrónimo de Receiver Operating Characteristic, cuyo origen son los estudios de imágenes de radar después de segunda Guerra Mundial. La curva ROC consiste en una representación gráfica de (s) frente $(1-e)$, con s y e definidas por las ecuaciones (45) y (46) para tareas de detección con solo dos resultados posibles (presente/ausente) o (SÍ/NO).

3.5.4.1. Relación entre el punto de corte y las medidas de sensibilidad y especificidad

El modelo será más satisfactorio cuanto menores sean los valores de n_{21} y n_{12} (idealmente, 0) y, por tanto, mayores sean los valores de s y e (idealmente, 1).

Los valores de las celdas de la Matriz de Confusión, (n_{11}, n_{22}, n_{21}, y n_{12}), dependen del punto de corte, p_c , por encima del cual se clasifica un individuo en el grupo con $Y = 1$. Por tanto, dado el modelo ajustado, podemos considerar a, s y e como funciones de p_c.

Figura 2: Curva ROC.



Fuente: Apuntes de clase. Barrera, J. (2013).

El mejor método posible de predicción se sitúa en un punto en la esquina superior izquierda, o coordenada (0,1) del espacio ROC en la Figura 2 representando un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún falso positivo). Por el contrario, una clasificación totalmente aleatoria (o adivinación aleatoria) daría un punto a lo largo de la línea diagonal, que se llama también línea de no discriminación, desde el extremo inferior izquierdo hasta la esquina superior derecha (independientemente de los tipos de base positivas y negativas). Un ejemplo típico de adivinación aleatoria sería decidir a partir de los resultados de lanzar una moneda al aire.

3.5.4.2. Área bajo la curva ROC

La prueba de Hosmer-Lemeshow por ejemplo, evalúa un aspecto de la validez del modelo: la calibración (grado en que la probabilidad predicha coincide con la observada). El otro aspecto es la discriminación (grado en que el modelo distingue entre individuos en los que ocurre el evento y los que no).

El área bajo la curva ROC, denominado AUC, varía entre 0,5 (no hay discriminación, se elige al azar) y 1 (discriminación perfecta). Esta área representa la probabilidad de que un individuo con respuesta 1 tenga un valor en la escala de medida considerada mayor que un individuo con respuesta 0. Por tanto, lo deseable es que esta medida sea lo más alta posible, según Hosmer, D. y Lemeshow, S., (2000) el modelo es preciso y tiene alta capacidad de discriminación cuando $AUC \geq 0{,}7$.

3.6. Diagnosis y Validación del modelo.

3.6.1. Análisis de los residuos.

Los residuos son muy importantes en el análisis de regresión, pues, informan sobre el grado de exactitud de los pronósticos, esto quiere decir que, cuanto más pequeño es el error típico de los residuos, mejores son los pronósticos, o lo que es lo mismo, mejor se ajusta la línea de regresión a la nube de puntos. El análisis de las características de los casos con residuos grandes (sean positivos o negativos; es decir, grandes en valor absoluto) puede ayudarnos a detectar casos atípicos y, consecuentemente, a perfeccionar la ecuación de regresión a través de un estudio detallado de los mismos. Si algunos de los residuos dados por las ecuaciones (39) o (41) resultan ser significativos, esto es, residuos superiores a ± 2, debe estudiarse su influencia sobre el ajuste del modelo. Una medida para estudiar la influencia de los residuos significativos es conocida como Distancia de Cook.

3.6.2. Distancia de Cook.

Cook, R. (1977) introduce una estadística para indicar la influencia de una observación con respecto a un modelo particular. Para una única observación, esta estadística proporciona también información sobre si dicha observación es un outlier. La distancia de Cook consiste en buscar una medida que indique la separación entre los parámetros estimados, caso incluyan la observación i-ésima y caso no la incluyan, y queda definida por:

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}_i\right)' X' X \left(\hat{\beta} - \hat{\beta}_i\right)}{pS^2} \tag{50}$$

donde:

$\hat{\beta} = \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ es un vector de $k = p + 1$ componentes,

$\hat{\beta}_i$: es la estimación de $\hat{\beta}$ sin la i-ésima observación,

$S^2 = \frac{SC_{Residual}}{n-k}$: es el estimador insesgado de la varianza residual,

k : es el número de parámetros,

p : es el número de variables explicativas,

n : es el tamaño de la muestra.

Cook sugiere que cada D_i sea comparada con el percentil de una F con k y $n - k$ grados de libertad; en otras palabras, grandes valores de D_i indican que la observación es influyente, y se utiliza como criterio de decisión. En la práctica, se sugiere que cuando los D_i son mayores a 1, las observaciones serán consideradas influyentes en el modelo.

4. METODOLOGÍA

Ésta investigación es de carácter cuantitativo, no experimental, de alcance correlacional, en la misma se analizó la probabilidad de egreso de los estudiantes de ingeniería de la UNVES y las variables que inciden en dicho egreso.

Para el desarrollo de la investigación se formuló la siguiente pregunta de investigación:

¿Cuál es la probabilidad de egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay?

Como objetivo general:

Estimar la probabilidad de egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.

Como objetivos específicos:

Analizar de manera descriptiva el egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

Determinar las variables académicas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

Determinar las variables demográficas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

Determinar el modelo estadístico utilizando la regresión logística con mejor bondad de ajuste que estime la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.

4.1. Población

La población está conformada por los estudiantes de las distintas carreras de ingenierías ofertadas de la UNVES compuesta por un total de 1250 estudiantes.

En esta investigación se trabaja con los 1250 estudiantes de las distintas ingenierías de la UNVES cohorte 2009-2018.

4.1.1. Participantes o sujetos

Los estudiantes de las distintas carreras de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo UNVES.

4.1.2. Descripción del lugar de estudio

La investigación ha sido realizada en la Universidad Nacional de Villarrica del Espíritu Santo UNVES ubicada en la ciudad de Villarrica, departamento del Guairá, de la república del Paraguay. En dicha casa de estudios se imparten carreras de grado como licenciaturas e ingenierías; así como también carreras de posgrado como diplomados, especializaciones y maestrías en las distintas áreas del saber.

4.2. Diseño de investigación

El diseño de la investigación es no experimental, ya que no se realiza la manipulación deliberada de variables y sólo se observan los fenómenos sucedidos en su ambiente natural para analizarlos, Campoy (2016), Kerlinger y Lee (2001) y Hernández,

Fernández y Baptista (2013). El alcance de la investigación es correlacional de tipo predictivo, ya que se realiza la estimación de la probabilidad de egreso (variable respuesta) de los estudiantes de las carreras de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo cohorte 2009-2018 a través de variables predictoras, Campoy (2016), Bisquerra (2009) y Hernández, Fernández y Baptista (2013), y el enfoque es cuantitativo ya que utiliza exclusivamente el procesamiento estadístico de las variables para la obtención de un modelo matemático que estima la probabilidad de la variable respuesta. Campoy (2016), Hernández, Fernández y Baptista (2013), Fernández y Pértegas (2002), Hueso y Cascant (2012)

Hipótesis: Las variables demográficas y las variables académicas estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo, cohorte 2009-2018.

Las variables analizadas son aquellas relacionadas con la estimación de egreso de los estudiantes de ingeniería de la UNVES tales como:

Calificación promedio por semestre hasta el primer curso.

Cantidad de materias aprobadas por semestre hasta el primer curso.

Ingeniería que estudia.

Sexo de los estudiantes.

Estado civil de los estudiantes

Ciudad de residencia del estudiante

Departamento de residencia del estudiante

4.3. Técnica de Recolección de datos

4.3.1. Herramientas

Para la realización de este esta investigación se ha utilizado la ficha académica del estudiante de ingeniería obrante en la base de datos del Centro Tecnológico de Informática y Comunicaciones CETIC, dependiente de la Dirección General Académica de la Universidad Nacional de Villarrica del Espíritu Santo UNVES, encargada de todo el historial demográfico y académico de los estudiantes de la UNVES. Los datos demográficos en la ficha académica del estudiante de ingeniería son llenados por el mismo estudiante al momento de su matriculación en cada semestre de la carrera en su respectiva unidad académica. Los datos académicos en la ficha académica del estudiante de ingeniería son llenados por los funcionarios de la coordinación académica de cada carrera de ingeniería en base a las actas de exámenes de cada asignatura correspondiente a la carrera.

Se pueden enumerar las siguientes las razones que conllevaron a utilizar los datos del CETIC para la realización de este estudio: CETIC es la encargada directa del control y depuración constante de la base de datos del historial académico y demográfico de los estudiantes de la UNVES, todos los certificados de estudios solicitados por los estudiantes en la UNVES son elaborados en base a los datos del CETIC, por eso la confiabilidad y validez de los datos en esta investigación.

4.3.2. Procedimiento

El procedimiento para la extracción de los datos que permitieron la realización del estudio se ha hecho en forma directa con la dirección del CETIC y los ingenieros desarrolladores del sistema informático propio de la UNVES, consistente en planillas electrónicas en soporte magnético. A partir de esta base de datos se utilizó el registro

académico y demográfico de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo, cohorte 2009-2018.

4.3.3. La Regresión Logística en la modelación de las variables

En este estudio ha sido necesario aplicar diferentes técnicas estadísticas para determinar las variables que han permitido estimar la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo, cohorte 2009-2018.

Dentro de las técnicas estadísticas utilizadas se han utilizado tanto la estadística descriptiva y la estadística inferencial.

En la Tabla de Operacionalización de Variables se presentan las variables e indicadores que se estudiaron para cada uno de los objetivos específicos.

Tabla 3: Operacionalización de variables.

Objetivos específicos	Variables	Indicadores	Unidades de análisis
Analizar de manera descriptiva el egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.	Egreso de las carreras de ingeniería	Número de egresados de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.	Ficha académica del estudiante en la base de datos del CETIC de la UNVES
Determinar las variables académicas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.	Variables explicativas Académicas	Calificación promedio por semestre hasta el primer curso.	
		Cantidad de materias aprobadas por semestre hasta el primer curso	
		Ingeniería que estudia	

Objetivos específicos	Variables	Indicadores	Unidades de análisis
Determinar las variables demográficas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.	Variables explicativas Demográficas	Sexo de los estudiantes	´
		Estado civil de los estudiantes	
		Ciudad de residencia los estudiantes	
		Departamento de residencia del estudiante	Ficha académica del estudiante en la base de datos del CETIC de la UNVES
Determinar el modelo estadístico utilizando la regresión logística con mejor bondad de ajuste que estime la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.	Probabilidad de egreso	Valor 0 para el estudiante no egresado Valor 1 para el estudiantes egresado	

4.4. Técnica de análisis de datos

Para el análisis de los datos se ha utilizado la estadística inferencial, en el modelado estadístico de las variables de estudio, se hace uso de la regresión logística con respuesta binaria, es un proceso estadístico para estimar las relaciones entre variables e incluye técnicas para el modelado y análisis de diversas variables, centrando la atención en la relación entre una variable dependiente y una o más variables independientes; dichas técnicas son explicadas de manera analítica con los fundamentos matemáticos y estadísticos en el marco teórico, para el procesamiento y modelado estadístico se ha utilizado el software R en su versión 3.5.3, R es un software libre y es uno de los lenguajes de programación más utilizados en investigación científica, principalmente en modelado estadístico. En el software R se utilizaron los siguientes paquetes : MASS, ROCR, ResourceSelection y ggplot2. Un paquete (package) es una colección de funciones, datos y código R que se almacenan en una carpeta conforme a una estructura bien definida, fácilmente accesible desde el Software R. Se ha utilizado también como herramienta para la visualización y ordenamiento de la base de datos el software LibreOffice en su versión 6.2.3, LibreOffice es un software libre de ofimática el cual incluye la planilla de cálculo y procesador de texto.

5. RESULTADOS

Para la presentación de los resultados de este estudio se consideró analizar los datos referente a los estudiantes de las carreras de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018 proporcionada por su Dirección General Académica a través del CETIC donde se encuentra el historial académico de los estudiantes, en dicho análisis se ha utilizado la técnica denominada regresión logística con respuesta binaria, la regresión logística es un proceso estadístico para estimar las relaciones entre variables e incluye técnicas para el modelado y análisis de diversas variables, centrando la atención en la relación entre una variable dependiente y una o más variables independientes.

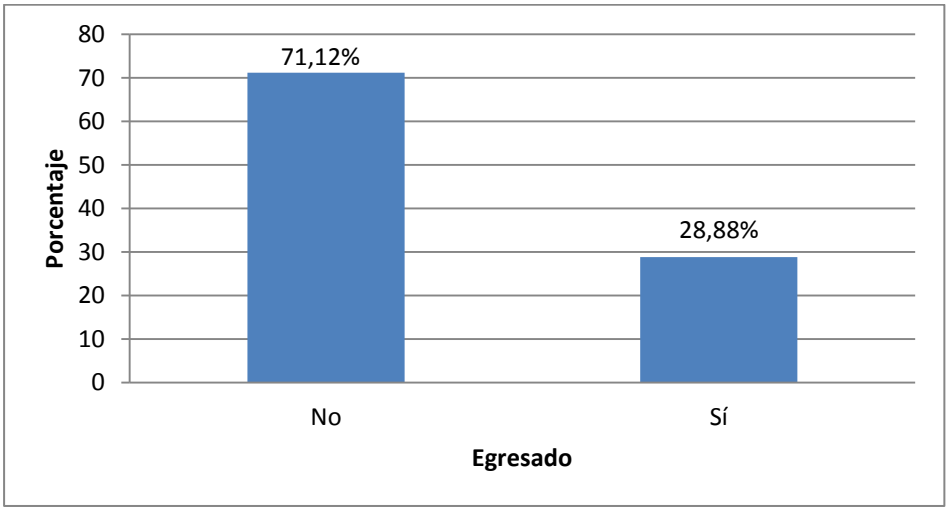
Los resultados están ordenados de acuerdo al logro de los objetivos específicos de la investigación que se detallan a continuación, para ello se logró:

a) La descripción el egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.
b) La determinación de las variables académicas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.
c) La determinación de las variables demográficas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.
d) La determinación del modelo estadístico utilizando la regresión logística con mejor bondad de ajuste que estime la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.

5.1. Datos demográficos

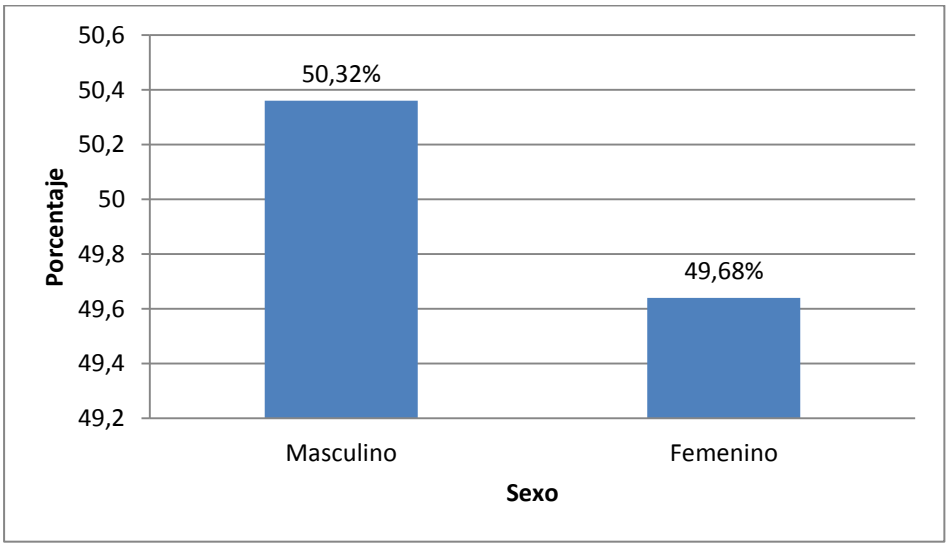
En primer término se presenta la variable dependiente sobre un total de 1250 estudiantes de la UNVES cohorte 2009-2018.

Gráfico 9: Porcentaje de egresados en la carreras de Ingeniería de la UNVES



De acuerdo al análisis de la información referente a los 1250 estudiantes se obtuvo que el 71,12% no culminaron la ingeniería y tan sólo el 28,88% de los estudiantes culminaron la carrera de ingeniería. Esto es aproximadamente 3 de cada 10 estudiantes culminaron la carrera de ingeniería.

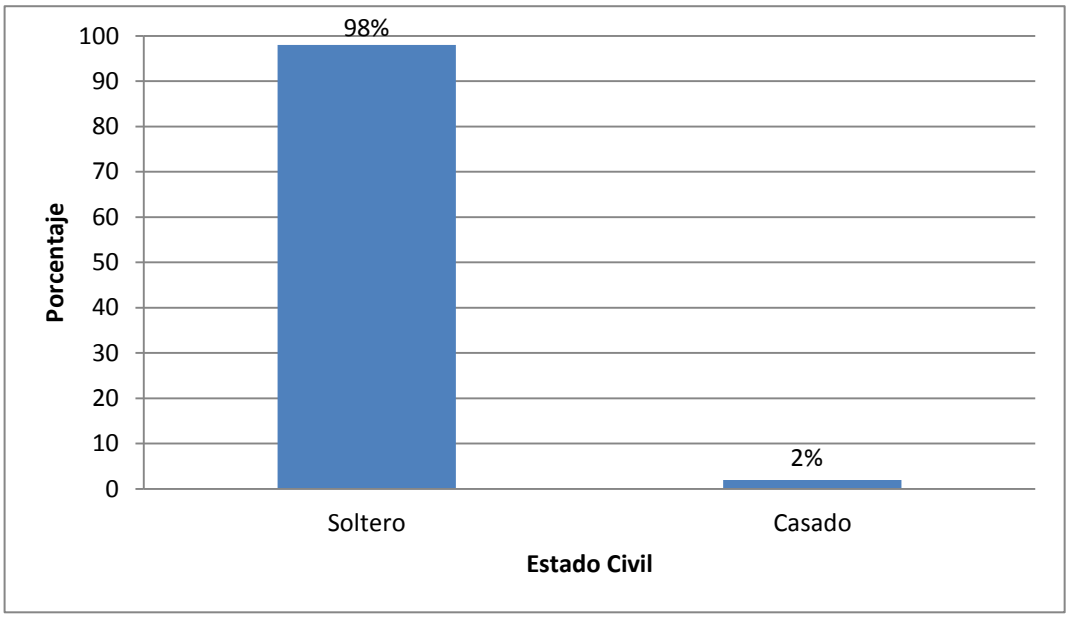
Gráfico 10: Porcentaje de estudiantes en la carreras de Ingeniería de la UNVES, según sexo.



Realizando el análisis descriptivo de los datos de la UNVES se observa que existe una distribución casi equitativa de los estudiantes de ingeniería con respecto al sexo. Sobre el total de 1250 estudiantes, el 50,32% corresponde al sexo masculino y el 49,68% corresponde al sexo femenino.

En relación al estado civil de los estudiantes de ingeniería de la UNVES, la mayor parte un 98% de los 1250 estudiantes son de estado civil soltero, mientras que tan sólo el 2% son de estado civil casado.

Gráfico 11: Porcentaje de estudiantes en la carreras de Ingeniería de la UNVES, según estado civil.



5.2. Análisis descriptivo el egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

En la Tabla 4 se observa y que el egreso condicionado por el sexo se encontró un mayor porcentaje muy superior de egresados en la categoría Femenino con un *40,47%* con respecto a la categoría Masculino con sólo el *17,34%* de egresados, esto equivale a una diferencia del *23,13%*, los estudiantes de sexo Masculino tienen menor proporción de

egreso que las de sexo Femenino. En el test χ^2 resultó que existe dependencia entre el Sexo y el Egreso, con una significancia inferior a $2{,}2 \times 10^{-16}$, es decir, el egreso o no egreso del estudiante depende del sexo.

Tabla 4: Condición de Egresado, según sexo.

	¿Es egresado?		Total
Sexo	No	Sí	
Masculino	*82,66%*	*17,34%*	*50,32%*
Femenino	*59,53%*	*40,47%*	*49,68%*
Total	*71,12%*	*28,88%*	*100%*

$\chi^2 = 3131.7, gl = 1, p\ valor < 2{,}2 \times 10^{-16}$

En la Tabla 5 se observa que el porcentaje de egresados de los estudiantes de estado civil Soltero es *28,49%* lo cual es muy similar al total general *28,88%*, en cambio el porcentaje de egresados de los estudiantes con estado civil Casado es *45,22%*.

En el test χ^2 resultó que existe dependencia entre el Estado Civil y el Egreso, con una significancia inferior a $2{,}2 \times 10^{-16}$, es decir, el egreso o no del estudiante depende de su estado civil.

Tabla 5: Condición de Egresado, según estado civil.

	¿Es egresado?		Total
Estado Civil	No	Sí	
Soltero	*71,51%*	*28,49%*	*98%*
Casado	*54,78%*	*45,22%*	*2%*
Total	*71,12%*	*28,88%*	*100%*

$\chi^2 = 127.05, gl = 1, p\ valor < 2{,}2 \times 10^{-16}$

En la Tabla 6 referente al promedio del estudiante al final de primer semestre resultó que ninguno de los estudiantes con promedio menor a 0,5 egresó, este porcentaje va aumentando a medida que va aumentado el promedio. En el test χ^2 resultó que existe dependencia entre el promedio al final del primer semestre y el Egreso, con una significancia

inferior a $2{,}2 \times 10^{-16}$, es decir, la variable egreso es dependiente del promedio del primer semestre del estudiante.

Tabla 6: Condición de Egresado, según promedio al final del primer semestre – primer curso.

Promedio	¿Es egresado? No	Sí	Total
[0 ; 0,5)	*100%*	*0%*	*12,32%*
[0,5 ; 1,5)	*86,54%*	*13,46%*	*5,04%*
[1,5 ; 2,5)	*82,01%*	*17,99%*	*24,48%*
[2,5 ; 3,5)	*68,69%*	*31,31%*	*30,8%*
[3,5 ; 4,5)	*53,75%*	*46,25%*	*21,68%*
[4,5 ; 5]	*35,37%*	*64,63%*	*5,68%*
Total	*71,12%*	*28,88%*	*100%*

$\chi^2 = 5950.6, gl = 5, p\ valor < 2{,}2 \times 10^{-16}$

En relación a la Tabla 7 sobre el promedio del estudiante al final de segundo semestre resultó que ninguno de estudiantes con promedio menor a 0,5 egresó, este porcentaje va aumentando a medida que va aumentado el promedio.

En el test χ^2 resultó que existe dependencia entre el promedio al final del segundo semestre y el Egreso, con una significancia inferior a $2{,}2 \times 10^{-16}$, es decir, la variable egreso es dependiente del promedio del segundo semestre del estudiante.

Tabla 7: Condición de Egresado, según promedio al final del segundo semestre – primer curso.

Promedio	¿Es egresado? No	Sí	Total
[0 ; 0,5)	*100%*	*0%*	*12,64%*
[0,5 ; 1,5)	*84,03%*	*15,97%*	*1,76%*
[1,5 ; 2,5)	*74,77%*	*25,23%*	*22,32%*
[2,5 ; 3,5)	*68,92%*	*31,08%*	*34,32%*
[3,5 ; 4,5)	*66,4%*	*33,6%*	*23,04%*
[4,5 ; 5]	*53,13%*	*46,87%*	*5,92%*
Total	*71,12%*	*28,88%*	*100%*

$\chi^2 = 1509.7, gl = 5, p\ valor < 2{,}2 \times 10^{-16}$

En relación a la Tabla 8 referente a la cantidad de materias aprobadas por parte del estudiante al final de primer semestre se observó que ninguno de los estudiantes que reprobaron todas las materias en el primer semestre egresó, el porcentaje de egresados va aumentando a medida que va aumentado la cantidad de materias aprobadas en el primer semestre. En el test χ^2 se observó que existe dependencia entre la cantidad de materias aprobadas en el primer semestre y el Egreso, con una significancia inferior a $2{,}2 \times 10^{-16}$, es decir, la variable egreso es dependiente cantidad de materias aprobadas en el primer semestre del primer curso.

Tabla 8: Condición de Egresado, según cantidad de materias aprobadas en el primer semestre - primer curso.

Cantidad de Materias Aprobadas	¿Es egresado? No	Sí	Total
0	*100%*	*0%*	*3,2%*
1	*91,58%*	*8,42%*	*38,8%*
2	*87,03%*	*12,97%*	*1,84%*
3	*85,89%*	*14,11%*	*2,64%*
4	*81,31%*	*18,69%*	*9,04%*
5	*54,58%*	*45,42%*	*5,6%*
6	*47,69%*	*52,31%*	*38,88%*
Total	*71,12%*	*28,88%*	*100%*

$\chi^2 = 9845.2, gl = 6, p\ valor < 2{,}2 \times 10^{-16}$

Con respecto a la Tabla 9 sobre a la cantidad de materias aprobadas por parte del estudiante al final de segundo semestre resultó que sólo el *8,89%* de los estudiantes que reprobaron todas las materias en el segundo semestre egresó, el porcentaje de egresados va aumentando a medida que va aumentado la cantidad de materias aprobadas en el segundo semestre. En el test χ^2 resultó que existe dependencia entre la cantidad de materias aprobadas en el segundo semestre y el Egreso, con una significancia inferior a $2{,}2 \times 10^{-16}$,

es decir, la variable egreso es dependiente cantidad de materias aprobadas en el segundo semestre del primer curso.

Tabla 9: Condición de Egresado, según cantidad de materias aprobadas en el segundo semestre - primer curso.

Cantidad de Materias Aprobadas	¿Es egresado? No	Sí	Total
0	*91,11%*	*8,89%*	*4,4%*
1	*91,11%*	*8,89%*	*5,8%*
2	*84,47%*	*15,53%*	*6%*
3	*78,01%*	*21,99%*	*4,08%*
4	*76,83%*	*23,17%*	*7,84%*
5	*64,98%*	*35,02%*	*38,56%*
6	*43,32%*	*56,68%*	*33,92%*
Total	*71,12%*	*28,88%*	*100%*

$\chi^2 = 10241, gl = 6, p\ valor < 2{,}2 \times 10^{-16}$

En la Tabla 10 resultó que el porcentaje de egresados de los estudiantes de la carrera de Ingeniería Ambiental es *86,82%* cual es el mayor porcentaje, en contrapartida Informática es la carrera con menor porcentaje de egresados, tan solo es *5,06%*. En el test χ^2se puede observar que existe dependencia entre la ingeniería que el estudiante optó por estudiar y el Egreso, con una significancia inferior a $2{,}2 \times 10^{-16}$, es decir, la variable egreso es dependiente de la ingeniería que el estudiante optó por estudiar.

Tabla 10: Condición de Egresado, según tipo de ingeniería.

Ingeniería	¿Es egresado? No	Sí	Total
Informática	*94,94%*	*5,06%*	*44,48%*
Zootecnia	*88,13%*	*11,87%*	*5,76%*
Química	*50,73%*	*49,27%*	*45,36%*
Azúcares	*25,51%*	*74,49%*	*2,24%*
Ambiental	*13,18%*	*86,82%*	*2,16%*
Total	*71,12%*	*28,88%*	*100%*

$\chi^2 = 13498, gl = 4, p\ valor < 2{,}2 \times 10^{-16}$

En relación a la condición de egresado según ciudad y departamento donde reside el estudiante, resultaron valores χ^2 muy pequeños y no significativos con p valores muy superiores al 0,05. La variable egreso no depende de la ciudad y departamento donde reside el estudiante.

5.3. Determinación de las variables académicas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

Las variables académicas estudiadas en esta investigación se citan a continuación:

Calificación promedio en el primer semestre.

Calificación promedio en el segundo semestre.

Cantidad de materias aprobadas en el primer semestre.

Cantidad de materias aprobadas en el segundo semestre.

Ingeniería que estudia.

En base a los test χ^2 aplicados a las variables académicas, todas las citadas resultaron con una relación dependencia con respecto a la variable dependiente Egreso con una significancia inferior al valor $2{,}2 \times 10^{-16}$.

Por lo tanto se tomaron todas estas variables académicas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018 para la determinación del modelo estadístico predictivo utilizando la regresión logística.

5.4. Determinación de las variables demográficas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

Las variables demográficas estudiadas en esta investigación se citan a continuación:

Sexo de los estudiantes.

Estado civil de los estudiantes.

Ciudad de residencia los estudiantes.

Departamento de residencia del estudiante.

En base a los test χ^2 aplicados a las variables demográficas, de las variables que se estudiaron todas, tuvieron una relación dependencia con respecto a la variable dependiente Egreso con una significancia inferior al valor $2{,}2 \times 10^{-16}$, excepto la ciudad de residencia y departamento donde reside el estudiante que no tuvieron una relación dependencia con respecto a la variable dependiente Egreso, resultaron con significancias superiores al valor 0,05.

Por lo tanto se tomaron variables demográficas, sexo y estado civil de los estudiantes con el fin de estimar la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018 para la determinación del modelo estadístico predictivo utilizando la regresión logística.

5.5. Modelo estadístico utilizando la regresión logística con mejor bondad de ajuste que estima la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.

5.5.1. Stepwise para seleccionar las variables que estiman la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.

Para seleccionar las variables que definen el mejor modelo se utiliza el procedimiento conocido como Stepwise, siguiendo las recomendaciones de Hosmer, D. y Lemeshow, S. (2000). Para facilitar la escritura del código de los diferentes modelos en el paquete estadístico R, se realiza la siguiente reparametrización de variables:

Variable dependiente.: Y: Alumno egresado o egresado de la UNVES.

Variables explicativas.

X_1: Cantidad de materias aprobadas en el primer semestre.

X_2: Calificación promedio en el primer semestre.

X_3: Cantidad de materias aprobadas en el segundo semestre.

X_4: Calificación promedio en el segundo semestre.

X_5: Tipo de Ingeniería.

X_6: Sexo.

X_7: Estado civil.

Para ajustar los distintos modelos se utiliza la función glm de R. En principio se ajusta el modelo sin variables explicativas (esto es, $Y_i = \beta_0$) y a seguir, se van incorporando o eliminado variables. El test condicional de razón de verosimilitudes contrasta el modelo seleccionado en el paso anterior con cada uno de los nuevos modelos planteados en el nuevo paso. En base a este resultado, se decide la variable que se introduce, o se saca, del modelo. El criterio de entrada y de salida se basa en el tamaño de la devianza residual y del AIC cuando se introduce una variable, esta debe reducirse y ser significativa con un p menor a

0,05. Para eliminar una variable se compara el último modelo con aquel que elimina una de las variables y el valor de la devianza residual y del AIC no deben sufrir un cambio significativo, es decir, el ajuste del modelo no debe empeorar con la eliminación de esa variable. Entonces, el criterio será en ambos casos la disminución significativa de la devianza residual y del AIC. Los pasos que conducen al modelo final son:

Paso 1: En este primer paso se ajustan varios modelos, entre ellos, el modelo nulo (sin ninguna variable), siendo éste el de referencia con el cual se comparan los otros modelos con una única variable explicativa (esto es, $Y_i = \beta_0$ contra cada uno de los siete modelos de la forma $Y_i = \beta_0 + \beta_k X_{ki}$ con $k = 1, 2, \dots, 7$).

Tabla 11: Test de Razón de Verosimilitudes para contrastar el Modelo Nulo contra los Modelos con una única variable explicativa.

Modelos	Resid. Dev.	df	Devianza	p valor	AIC
$M_0: Y_i = \beta_0$	57726.96				57728.96
$M_1: Y_i = \beta_0 + \beta_1 X_{1i}$	48635.48	1	9091.4821	0.000001e-291	48639.48
$M_2: Y_i = \beta_0 + \beta_1 X_{2i}$	52973.47	1	4753.4920	0.000000e+00	52977.47
$M_3: Y_i = \beta_0 + \beta_1 X_{3i}$	48866.52	1	8860.4430	2.360134e-291	48870.52
$M_4: Y_i = \beta_0 + \beta_1 X_{4i}$	56396.22	1	1330.7405	0.000001e-291	56400.22
$M_5: Y_i = \beta_0 + \beta_1 X_{5i}$	42839.88	4	14887.0822	0.000001e-291	42849.88
$M_6: Y_i = \beta_0 + \beta_1 X_{6i}$	54531.24	1	3195.7232	00.000001e-291	54535.24
$M_7: Y_i = \beta_0 + \beta_1 X_{7i}$	57609.99	1	116.9714	2.912422e-27	57613.99

La Tabla 11 resumió los resultados de este paso, la variable Tipo de Ingeniería es la primera a ser incluida en el modelo buscado. Esto se debe a que la reducción en devianza residual y la disminución del AIC, ocasionada por la inclusión de la variable X_{5i}, con respecto al modelo nulo es la de mayor reducción, resultando altamente significativa al ser comparada con una χ^2 con 4 grados de libertad.

Paso 2: Se partió del modelo con la primera variable X_{5i}: Tipo de Ingeniería incorporada. Contra este modelo se van a contrastar un total de seis modelos de la forma $M_5: Y_i = \beta_0 + \beta_1 X_{5i} + \beta_k X_{ki}$ con $k = 1, 2, \ldots, 6$. Identificando con φ_1 al modelo obtenido en el Paso 1, φ_1 es $Y_i = \beta_0 + \beta_1 X_{5i}$, se obtuvo la Tabla XXX, el cual muestra que la variable X_{6i}: Sexo es la que se sumó, al modelo buscado, ya que disminuyó el AIC, y es la que produjo una mayor reducción en la devianza residual siendo ésta significativa con un valor p del orden 0.000001E-291.

Tabla 12: Test de Razón de Verosimilitudes para contrastar el Modelo φ_1 contra los que agregan una de las restantes variables explicativas a la vez.

Modelos	Resid. Dev.	df	Devianza	p valor	AIC
$M_0: \varphi_1$	42839.88	NA	NA	NA	42849.88
$M_1: \varphi_1 + \beta_2 X_1$	42687.87	1	1.520.096	6.31E-29	42699.87
$M_2: \varphi_1 + \beta_2 X_2$	41764.91	1	10.749.701	9.10E-230	41776.91
$M_3: \varphi_1 + \beta_2 X_3$	41668.14	1	11.717.408	8.45E-251	41680.14
$M_4: \varphi_1 + \beta_2 X_4$	41829.31	1	10.105.667	9.07E-216	41841.31
$M_5: \varphi_1 + \beta_2 X_{6i}$	42839.88	4	148.870.822	0.0001E-198	41304.43
$M_6: \varphi_1 + \beta_2 X_{7i}$	41292.43	1	15.474.470	0.0001E-199	42800.52

Paso 3: Se partió del modelo con las variables Tipo de Ingeniería y Sexo incorporadas, contra este modelo se contrastaron un total de cinco modelos de la forma $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_k X_{ki}$ con $k = 1, 2, \ldots, 5$, se denotó por φ_2 al modelo obtenido en el Paso 2, se obtuvo que el modelo incluyó a la variable X_2: Calificación promedio en el primer semestre que presentó mejor ajuste siendo significativo con un valor p de 2.37E-204, como se obtuvo en la Tabla 13.

Tabla 13: Test de Razón de Verosimilitudes para contrastar el Modelo φ_2 contra los que agregan una de las restantes variables explicativas a la vez.

Modelos	Resid. Dev.	df	Devianza	p valor	AIC
M_0: φ_2	41292.43	NA	NA	NA	41304.43
M_1: $\varphi_2 + \beta_3 X_1$	41194.85	1	97.574.670	5.19E-17	41208.85
M_2: $\varphi_2 + \beta_3 X_2$	40334.38	1	958.045.382	2.37E-204	40348.38
M_3: $\varphi_2 + \beta_3 X_3$	40376.78	1	915.644.341	3.90E-195	40390.78
M_4: $\varphi_2 + \beta_3 X_4$	40523.26	1	769.167.833	2.73E-163	40537.26
M_5: $\varphi_2 + \beta_3 X_{7i}$	41284.74	1	7.688.801	5.56E+03	41298.74

Paso 4: Antes de realizar el análisis a efectos de incorporar una nueva variable, se indagó la posibilidad de eliminar alguna de las variables introducidas hasta el paso 2. Para ello, se contrastó el modelo $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2$ denotado por φ_3 con los dos modelos resultantes de excluir del anterior una variable a la vez. En ambos casos empeoraron los ajustes y no se eliminaron ninguna de las dos primeras variables introducidas como se obtuvo en la Tabla 14.

Tabla 14: Test de Razón de Verosimilitudes para contrastar el Modelo φ_3 contra los que resultan de eliminar una a la vez las dos primeras variables introducidas.

Modelos	Resid. Dev.	Df	Devianza	p valor	AIC
M_0: φ_3	40334.38	NA	NA	NA	40348.38
M_1: $\varphi_3 - \beta_1 X_{5i}$	50307.11	-4	-99.727.291	0.00E+00	50313.11
M_2: $\varphi_3 - \beta_2 X_{6i}$	41764.91	-1	-14.305.223	4.90E-307	41776.91

Paso 5. Se partió del modelo con las variables Tipo de Ingeniería, Sexo y Calificación promedio en el primer semestre incorporadas, contra este modelo se contrastaron un total de modelos de la forma $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_k X_{ki}$ con $k = 1, 2, \ldots, 4$, se denotó por φ_3 al modelo obtenido en el Paso 3, se obtuvo que el modelo

incluyó a la variable X_3: Cantidad de materias aprobadas en el segundo semestre que presentó mejor ajuste siendo significativo con un valor p de 1.93E-66, como se obtuvo en la Tabla 15.

El modelo hasta aquí es $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_4 X_3$

Tabla 15: Test de Razón de Verosimilitudes para contrastar el Modelo φ_3 contra los que agregan una de las restantes variables explicativas a la vez.

Modelos	Resid. Dev.	df	Devianza	p valor	AIC
M_0: φ_3	40334.38	NA	NA	NA	40348.38
M_1: $\varphi_3 + \beta_4 X_1$	40296.82	1	37.561.780	8.86E-04	40312.82
M_2: $\varphi_3 + \beta_4 X_3$	40010.36	1	324.019.112	1.93E-66	40026.36
M_3: $\varphi_3 + \beta_4 X_4$	40224.68	1	109.700.293	1.14E-19	40240.68
M_4: $\varphi_3 + \beta_4 X_{7i}$	40327.36	1	7.026.421	8.03E+03	40343.36

Paso 6. Antes de realizar el análisis a efectos de incorporar una nueva variable, se indagó la posibilidad de eliminar alguna de las variables introducidas hasta el paso 3. Para ello, se contrastó el modelo $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_4 X_3$ denotado por φ_4 con los tres modelos resultantes de excluir del anterior una variable a la vez. En ambos casos empeoraron los ajustes y no se eliminaron ninguna de las tres primeras variables introducidas, como se obtuvo en la Tabla 16.

Tabla 16: Test de Razón de Verosimilitudes para contrastar el Modelo φ_4 contra los que resultan de eliminar una a la vez las tres primeras variables introducidas.

Modelos	Resid. Dev.	df	Devianza	p valor	AIC
M_0: φ_4	40010.36	NA	NA	NA	40026.36
M_1: $\varphi_4 - \beta_1 X_{5i}$	46374.59	-4	-63.642.227	0.00E+00	46382.59
M_2: $\varphi_4 - \beta_2 X_{6i}$	41293.73	-1	-12.833.681	4.65E-275	41307.73
M_3: $\varphi_4 - \beta_3 X_2$	40376.78	-1	-3.664.202	1.13E-75	40390.78

Paso 7. Se partió del modelo con las variables incorporadas: Tipo de Ingeniería, Sexo, Calificación promedio en el primer semestre y Cantidad de materias aprobadas en el segundo semestre, contra este modelo se contrastaron un total de tres modelos de la forma $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_4 X_3 + \beta_k X_{ki}$ con $k = 1, 2, \ldots, 3$, se denotó por φ_4 al modelo obtenido en el Paso 5, se obtuvo que el modelo incluyó a la variable X_1: Cantidad de materias aprobadas en el primer semestre que presentó mejor ajuste siendo significativo con un valor p de 4.64E-230, como se obtuvo en la Tabla 17.

El modelo hasta aquí es $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_1$

Tabla 17: Test de Razón de Verosimilitudes para contrastar el Modelo φ_4 contra los que agregan una a la vez las restantes variables explicativas.

Modelos	Resid. Dev.	df	Devianza	p valor	AIC
M_0: φ_4	40010.36	NA	NA	NA	40026.36
M_1: $\varphi_4 + \beta_5 X_1$	38934.05	1	1076.3145	4.64E-230	38952.05
M_2: $\varphi_4 + \beta_5 X_4$	39986.84	1	23.527176	1.23E+00	40004.84
M_3: $\varphi_4 + \beta_5 X_{7i}$	40006.07	1	4.290311	3.83E+04	40024.07

Paso 8. Antes de realizar el análisis a efectos de incorporar una nueva variable, se indagó la posibilidad de eliminar alguna de las variables introducidas hasta el paso 5. Para ello, se contrastó el modelo $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_4 X_2 + \beta_5 X_1$ denotado por φ_5 con los cuatro modelos resultantes de excluir del anterior una variable a la vez. En ambos casos empeoraron los ajustes y no se eliminaron ninguna de las cuatro primeras variables introducidas.

Como se obtuvo en la Tabla 18 que se expone a continuación.

Tabla 18: Test de Razón de Verosimilitudes para contrastar el Modelo φ_5 contra los que resultan de eliminar una a la vez las cuatro primeras variables introducidas.

Modelos	Resid. Dev.	df	Devianza	p valor	AIC
M_0: φ_5	38934.05	NA	NA	NA	38952.05
M_1: $\varphi_5 - \beta_1 X_{5i}$	46093.53	-4	-7159.4837	0.00E+00	46103.53
M_2: $\varphi_5 - \beta_2 X_{6i}$	40134.21	-1	-1200.1638	5.62E-257	40150.21
M_3: $\varphi_5 - \beta_3 X_2$	39575	-1	-640.9478	2.08E-135	39591
M_4: $\varphi_5 - \beta_4 X_3$	40296.82	-1	-136.27719	2.58E-292	40312.82

Paso 9. Se partió del modelo con las variables incorporadas, Tipo de Ingeniería, Sexo, Calificación promedio en el primer semestre, Cantidad de materias aprobadas en el segundo semestre y Cantidad de materias aprobadas en el primer semestre, contra este modelo se contrastaron un total de dos modelos de la forma $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_1 + \beta_k X_{ki}$ con $k = 1, 2$, se denotó por φ_5 al modelo obtenido en el Paso 7, se obtuvo que el modelo incluyó a la variable X_{7i}: Estado civil que presentó mejor ajuste siendo significativo con un valor p de 1.38E-02, como se obtuvo en la Tabla 19.

El modelo hasta aquí es $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_1 + \beta_6 X_{7i}$

Tabla 19: Test de Razón de Verosimilitudes para contrastar el Modelo φ_5 contra los que agregan una a la vez las restantes variables explicativas.

Modelos	Resid. Dev.	df	Devianza	p valor	AIC
M_0: φ_5	38934.05	NA	NA	NA	38952.05
M_1: $\varphi_5 + \beta_6 X_4$	38933.93	1	0	7.31E-01	38953.93
M_2: $\varphi_5 + \beta_6 X_{7i}$	38927.99	1	6.057.991	1.38E-02	38947.99

Paso 10. Antes de realizar el análisis a efectos de incorporar una nueva variable, se indagó la posibilidad de eliminar alguna de las variables introducidas hasta el paso 7. Para ello, se contrastó el modelo $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_4 X_2 + \beta_5 X_1 + \beta_6 X_{7i}$ denotado por φ_6 con los cinco modelos resultantes de excluir del anterior una variable a la

vez. En ambos casos empeoraron los ajustes y no se eliminaron ninguna de las cinco primeras variables introducidas, como se obtuvo en la Tabla 20.

Tabla 20: Test de Razón de Verosimilitudes para contrastar el Modelo φ_6 contra los que resultan de eliminar una a la vez las cuatro primeras variables introducidas.

Modelos	Resid. Dev.	df	Devianza	p valor	AIC
M_0: φ_6	38927.99	NA	NA	NA	38947.99
M_1: $\varphi_6 - \beta_1 X_{5i}$	46090.36	-4	-7162.3640	0.00E+00	46102.36
M_2: $\varphi_6 - \beta_2 X_{6i}$	40092.67	-1	-1164.6818	2.89E-249	40110.67
M_3: $\varphi_6 - \beta_3 X_2$	39570.67	-1	-642.6759	8.75E-136	39588.67
M_4: $\varphi_6 - \beta_4 X_3$	40288.56	-1	-1360.5657	7.79E-292	40306.56
M_5: $\varphi_6 - \beta_4 X_3$	40006.07	-1	-1078.0822	1.92E-230	40024.07

Paso 11. Se partió del modelo con las seis variables incorporadas, Tipo de Ingeniería, Sexo, Calificación promedio en el primer semestre, Cantidad de materias aprobadas en el segundo semestre, Cantidad de materias aprobadas en el primer semestre y Estado civil, contra este modelo se contrastó un modelo de la siguiente forma $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_1 + \beta_6 X_{7i} + \beta_7 X_{ki}$ con $k = 1$, se denotó por φ_6 al modelo obtenido en el Paso 9, se obtuvo que el modelo no incluyó a la variable X_4: Calificación promedio en el segundo semestre ya que presentó un leve peor ajuste siendo no significativo con un valor p de 0.6526774, , como se obtuvo en la Tabla 21.

Tabla 21: Test de Razón de Verosimilitudes para contrastar el Modelo φ_6 contra los que agregan la restante variable explicativa.

Modelos	Resid. Dev.	df	Devianza	p valor	AIC
M_0: φ_6	38927.99	NA	NA	NA	38947.99
M_1: $\varphi_6 + \beta_6 X_4$	38927.79	1	0.2025412	0.6526774	38949.79

En resumen la regresión logística incorporó seis variables explicativas que se citan a continuación:

X_1: Cantidad de materias aprobadas en el primer semestre.

X_2: Calificación promedio en el primer semestre.

X_3: Cantidad de materias aprobadas en el segundo semestre.

X_{5i}: Tipo de Ingeniería.

X_{6i}: Sexo.

X_{7i}: Estado civil.

Por lo mencionado y por el principio de Parsimonia el modelo final es:

$$Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_1 + \beta_6 X_{7i}$$

5.5.2. Estimación y contrastes sobre los parámetros del modelo que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.

La siguiente Tabla contiene los coeficientes del último modelo ajustado obtenido por selección Stepwise. Las filas identifican las variables incluidas (con sus categorías de agrupamiento), además de la constante (Intercepto) del modelo. En las columnas se encuentran los valores estimados de los distintos parámetros (Estimación), el error estándar de cada estimación (Error Est.), el valor del estadístico del Test de Wald (Valor z) y su significación $\left(Pr(>|z|)\right)$.

Con respecto a los valores estimados de los β_k (y sus respectivas categorías), puede apreciarse que todas las variables incluidas en el modelo resultan altamente significativas con valores p inferiores a $0{,}05$. En las variables cualitativas, cada categoría de referencia se denota por (*).

Tabla 22: Estimación de los parámetros del modelo final ajustado por Regresión Logística, mediante selección Stepwise.

Variable	Categoría o valor de la variable	Estimación	Error Est.	Valor z	$Pr(>\|z\|)$
	(Intercepto)	-341.266	0.07261	-46.999	<2e-16 ***
X_{5i}	Zootecnia (*)				
	Química	182.501	0.0668	27.322	<2e-16 ***
	Ambiental	492.348	0.12694	38.785	<2e-16 ***
	Informática	-0.76055	0.07579	-10.034	<2e-16 ***
	Azúcares	441.763	0.10908	40.500	<2e-16 ***
X_{6i}	Masculino (*)				
	Femenino	0.89389	0.02653	33.698	<2e-16 ***
X_{7i}	Soltero (*)				
	Casado	0.19645	0.08003	2.455	0.0141 *
X_2	Promedio Primer Semestre (0 a 5)	0.32072	0.01271	25.225	<2e-16 ***
X_1	Cantidad de materias aprobadas en el primer semestre (0 a 6)	0.49748	0.01602	31.064	<2e-16 ***
X_3	Cantidad de materias aprobadas en el segundo semestre (0 a 6)	0.52018	0.01527	34.066	<2e-16 ***

5.5.3. Cálculo de las OR para la interpretación de los parámetros del modelo que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.

Cuando la variable explicativa es cualitativa nominal, como es el caso de las variables X_{5i}, X_{6i} y X_{7i}; lo que se analiza con los cocientes de ventaja en términos probabilísticos, es el cambio marginal de pasar de la categoría de referencia a otra categoría de la misma variable. Cuando la variable explicativa es cuantitativa como es el caso de las variables X_1, X_2 y X_3; lo que se analiza con los cocientes de ventaja en términos probabilísticos, es el incremento de r unidades en X_k. En las variables cualitativas de la Tabla 23, cada categoría de referencia se denota por (*).

Tabla 23: Estimación de los OR del modelo final ajustado por Regresión Logística, mediante selección Stepwise. Con sus intervalos de confianza ($\alpha = 5\%$)

Variable	Categoría o valor de la variable	OR	2,5%	97,5%
(Intercept)		1.62E+148	1.62E+147	1.62E+150
X_{5i}	Zootecnia (*)			
	Química	6.20	5.42	6.98
	Ambiental	137.48	125.78	149.18
	Informática	0.47	0.43	0.51
	Azúcares	82.90	77.85	87.95
X_{6i}	Masculino (*)			
	Femenino	2.45	2.33	2.56
X_{7i}	Soltero (*)			
	Casado	1.22	1.12	1.33
X_2	Promedio Primer Semestre (0 a 5)	1.38	1.28	1.47
X_1	Cantidad de materias aprobadas en el primer semestre (0 a 6)	1.65	1.55	1.76
X_3	Cantidad de materias aprobadas en el segundo semestre (0 a 6)	1.68	1.57	1.79

Si la OR resulta inferior a la unidad, esto sucede cuando β_k tiene valor negativo, como sucede con la OR correspondiente a la variable X_{5i} en la categoría Informática, en este caso se calcula la inversa OR^{-1} para una mejor interpretación, siendo ahora la ventaja a favor de la categoría de referencia que es Zootecnia.

5.5.4. Bondad de ajuste del modelo que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.

Para determinar qué tan bueno resultó el modelo final ajustado por Regresión Logística, se realizaron los siguientes procedimientos: el Test de Hosmer-Lemeshow, el Área bajo la curva ROC y La Tasa de Clasificaciones correctas

5.5.4.1. Test de Hosmer-Lemeshow

La hipótesis nula de este test es que el modelo propuesto es el apropiado para explicar la probabilidad que un estudiante de ingeniería de la UNVES sea egresado en su carrera, con lo cual lo conveniente es no rechazarla.

Siendo el modelo final $Y_i = \beta_0 + \beta_1 X_{5i} + \beta_2 X_{6i} + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_1 + \beta_6 X_{7i}$, el resumen del test se muestra a continuación:

Tabla 24: Test de Hosmer-Lemeshow para la bondad de ajuste del modelo final que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.

χ^2	df	p valor
7,862	8	0,34

En base a los resultados de este test se concluyó que no existen evidencias estadísticamente significativas para rechazar la hipótesis nula, no podemos rechazar la hipótesis que el modelo final propuesto ajusta bien, es decir, el modelo final es adecuado.

5.5.4.2. Área bajo la curva ROC

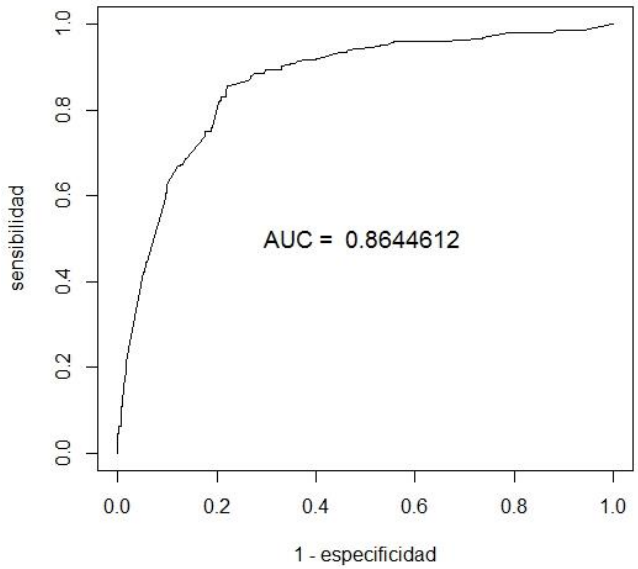
El área bajo la respectiva curva (estadístico AUC) representa la probabilidad de que un estudiante de ingeniería con valor $Y = 1$ (probabilidad de ser egresado) tenga un valor en la escala de medida considerada mayor que un estudiante de ingeniería con valor $Y = 0$ (probabilidad de no ser egresado). Se recuerda que el AUC es mejor cuanto más cercano esté a la unidad. Hosmer y Lemeshow (2000) sugieren que:

Si $0{,}7 < AUC \leq 0{,}8$: se considera aceptable la discriminación.

Si $0{,}8 < AUC \leq 0{,}9$: se considera excelente la discriminación.

Si $AUC > 0{,}9$: se considera la discriminación excepcional.

Gráfico 12: Área bajo la curva ROC del modelo final ajustado por Regresión Logística que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.



Como se obtuvo el $AUC = 0{,}8644612$, la capacidad de discriminación del modelo final es excelente.

5.5.4.3. Punto de Corte y Tasa de Clasificaciones correctas.

De acuerdo con el mecanismo sugerido por Hosmer y Lemeshow (2000) para seleccionar el mejor punto de corte, en el Grafico 13 se obtuvo un punto de corte igual a 0,38 que maximiza tanto la sensibilidad como la especificidad del modelo final.

Gráfico 13: Sensibilidad y Especificidad versus Punto de Corte del modelo final ajustado por Regresión Logística que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.

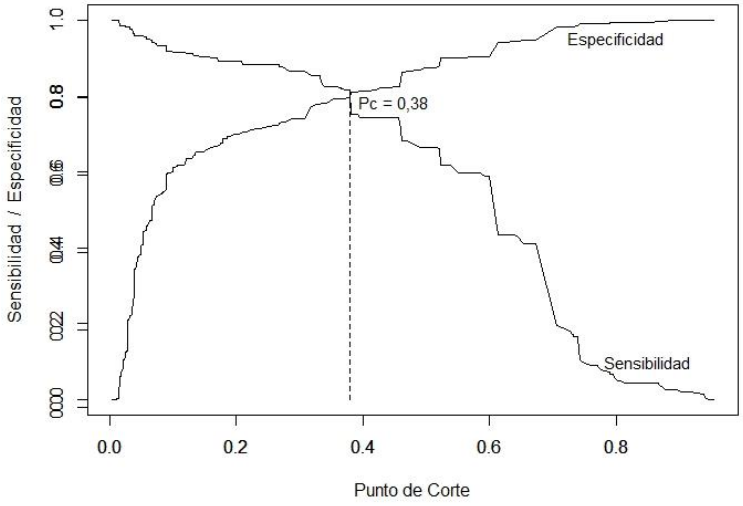


Tabla 25: Índices con punto el corte seleccionado para medir la bondad de ajuste del modelo final que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.

Precisión	Sensibilidad	Especificidad
0,8027466	0,8195337	0,8059482
Punto de corte igual a 0,38		

Con el $p_c = 0{,}38$ se obtuvieron excelente índice de bondad de ajuste del modelo final, todas tienen un porcentaje similar y superior al 80%, esto significa que el modelo final que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES identifica a 8 de cada 10 estudiantes que van a ser egresados en el primer año de la carrera.

5.5.5. Diagnosis y validación del modelo final que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.

El análisis de los residuos y la Distancia de Cook (medidas de influencia) son los indicadores que se aplican para realizar la diagnosis y la validación del modelo final que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.

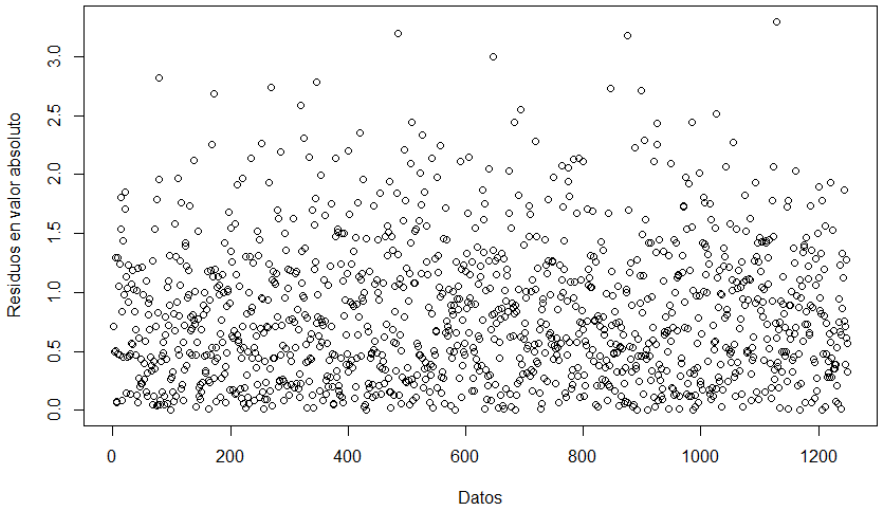
5.5.5.1. Análisis de los residuos

Al observar la Tabla 26 y el Gráfico 14 se constató que existen residuos en valor absoluto superiores al valor máximo 2, éstos a su vez, podrían señalar que existen valores observados que afectan el ajuste global del modelo final, más puntualmente, de 1250 residuos, 52 son mayores a 2 en valor absoluto, es decir, alrededor de un 4,16 %.

Tabla 26: Cuartiles de los residuos estimados del modelo final ajustado por regresión logística que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.

Residuos				
Min	Q_1	Q_2	Q_3	Máximo
-2,64741	-1,39587	0,14373	1,79805	3,30104

Gráfico 14: Residuos estimados en valor absoluto del modelo final ajustado por regresión logística que estima la probabilidad de egreso de los estudiantes de ingeniería de la UNVES.



5.5.5.2. Distancia de Cook.

La Distancia de Cook del i-ésimo elemento consiste en buscar una medida que indique la separación entre los parámetros estimados, caso incluyan la i-ésima observación y caso no la incluyan. Cada observación tiene asociada una distancia y se considera significativa si es mayor que 1. Se calculó las 1250 distancias con el software R y el valor máximo que se halló es aproximadamente igual a 0,0007693125, menor al valor límite 1. Por tanto con este resultado, ninguna observación es potencialmente influyente en el buen ajuste del modelo final estimado por regresión logística, y se dio por validado el modelo.

Con este último paso se logró el último objetivo específico cual es la determinación del modelo estadístico utilizando la regresión logística con mejor bondad de ajuste que estime la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay

En base a todos los resultados obtenidos se utilizando el modelo de regresión logística se alcanzó estimar la probabilidad de egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.

6. DISCUSIÓN FINAL

A partir de los resultados obtenidos se ha podido realizar la estimación de la probabilidad de egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay utilizando la ficha académica de los estudiantes de ingeniería obrantes en la base de datos del Centro Tecnológico de Informática y Comunicaciones CETIC, dependiente de la Dirección General Académica de la Universidad Nacional de Villarrica del Espíritu Santo UNVES. Por tanto, se ha podido cumplir con el objetivo general y los objetivos específicos de la investigación.

6.1. Del análisis descriptivo del egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

Realizando un exhaustivo análisis de la información contenida en las fichas académicas de los estudiantes de ingeniería de la UNVES (1250 fichas), se observó que aproximadamente tan sólo 3 de cada 10 estudiantes de ingeniería culminaron su carrera tal como se observó en el Gráfico 5.

Referente al sexo de los estudiantes existió una distribución general equitativa, siendo 50,32% del sexo masculino y el 49,68% al sexo femenino como se observó en el Gráfico 6. En la tabla 4 se observa que existe un mayor porcentaje de sexo femenino que egresaron comparado con el de sexo masculino.

Con relación al estado civil de los estudiantes, se observó una notable mayor proporción de estudiantes con estado civil soltero con respecto al estado civil casado, 8 de cada 10 estudiantes de ingeniería de la UNVES son solteros, tal como se indicó en el Gráfico

11. Sin embargo en la Tabla 5 se observa que en la categoría casado existe una mayor proporción de egresado que dentro de la categoría soltero.

Con respecto a la variable calificación promedio por semestre se observó que a medida que aumenta el promedio del estudiante por semestre aumenta el porcentaje de egresados, las variables son directamente proporcionales, como se indicó en la Tablas 6 y Tabla 7.

En base a la variable cantidad de materias aprobadas por semestre observó que a medida que aumenta el valor de esta variable aumenta el porcentaje de egresados, las variables son directamente proporcionales, como se indicó en la Tablas 8 y Tabla 9.

Sobre la carrera de ingeniería elegida por los estudiantes de la UNVES se observó que aproximadamente 45,36% de los estudiantes optaron por estudiar Ingeniería Química, 44,48% de los estudiantes optaron por estudiar Ingeniería Informática, y en lo que sobra se distribuyen las demás ingenierías.

6.2. De las variables académicas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

En este punto se realizó el análisis de dependencia entre la variable respuesta (egreso) y las variables académicas cada una de ellas en una tabla de contingencia aplicando un test χ^2 para observar si las variables eran independientes o dependientes entre sí.

Con respecto a la variable cuantitativa, calificación promedio al final del primer semestre; en relación con el egreso se obtuvo una de dependencia con alta significancia aplicando el test χ^2 como resultó en la Tabla 6, por lo que fue tenida en cuenta en el Stepwise para su inclusión en el modelo para estimar la probabilidad de egreso de los estudiantes de ingeniería de la UNVES. Con lo que se concluye que existen evidencias

estadísticamente significativas para concluir que existe una dependencia entre el egreso y la calificación promedio al final del primer semestre.

En relación a la variable cuantitativa, calificación promedio al final del segundo semestre; con respecto al egreso se obtuvo una relación de dependencia con alta significancia aplicando el test χ^2 como resultó en la Tabla 7, por lo que fue tenida en cuenta en el Stepwise para su inclusión en el modelo para estimar la probabilidad de egreso de los estudiantes de ingeniería de la UNVES. Con lo que se concluye que existen evidencias estadísticamente significativas para concluir que existe una dependencia entre el egreso y la calificación promedio al final del segundo semestre.

En la variable cuantitativa, cantidad de materias aprobadas en el primer semestre; con respecto al egreso se obtuvo una relación de dependencia con alta significancia aplicando el test χ^2 como resultó en la Tabla 8, por lo que fue tenida en cuenta en el Stepwise para su inclusión en el modelo para estimar la probabilidad de egreso de los estudiantes de ingeniería de la UNVES. Con lo que se concluye que existen evidencias estadísticamente significativas para concluir que existe una dependencia entre el egreso y la cantidad de materias aprobadas en el primer semestre.

En la variable cuantitativa, cantidad de materias aprobadas en el segundo semestre; con respecto al egreso se obtuvo una relación de dependencia con alta significancia aplicando el test χ^2 como resultó en la Tabla 9, por lo que fue tenida en cuenta en el Stepwise para su inclusión en el modelo para estimar la probabilidad de egreso de los estudiantes de ingeniería de la UNVES. Con lo que se concluye que existen evidencias estadísticamente significativas para concluir que existe una dependencia entre el egreso y la cantidad de materias aprobadas en el segundo semestre.

En la variable cualitativa, Ingeniería que estudia; con respecto al egreso se obtuvo una relación de dependencia con alta significancia aplicando el test χ^2 como resultó en la

Tabla 10, por lo que fue tenida en cuenta en el Stepwise para su inclusión en el modelo para estimar la probabilidad de egreso de los estudiantes de ingeniería de la UNVES. Con lo que se concluye que existen evidencias estadísticamente significativas para concluir que existe una dependencia entre el egreso y el tipo de ingeniería que estudia.

6.3. De las variables demográficas que estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay cohorte 2009-2018.

En la variable cualitativa, sexo de los estudiantes; con respecto al egreso se obtuvo una relación de dependencia con alta significancia aplicando el test χ^2 como resultó en la Tabla 4, por lo que fue tenida en cuenta en el Stepwise para su inclusión en el modelo para estimar la probabilidad de egreso de los estudiantes de ingeniería de la UNVES. Con lo que se concluye que existen evidencias estadísticamente significativas para concluir que existe una dependencia entre el egreso y el sexo de los estudiantes.

En la variable cualitativa, estado civil de los estudiantes; con respecto al egreso se obtuvo una relación de dependencia con alta significancia aplicando el test χ^2 como resultó en la Tabla 5, por lo que fue tenida en cuenta en el Stepwise para su inclusión en el modelo para estimar la probabilidad de egreso de los estudiantes de ingeniería de la UNVES. Con lo que se concluye que existen evidencias estadísticamente significativas para concluir que existe una dependencia entre el egreso y estado civil de los estudiantes.

En relación a las variables cualitativas, ciudad y departamento de residencia los estudiantes; con respecto al egreso se obtuvo una relación de dependencia no significativa aplicando el test χ^2, por lo que no fueron tenidas en cuenta en el Stepwise para su inclusión en el modelo para estimar la probabilidad de egreso de los estudiantes de ingeniería de la UNVES. Con lo que se concluye que existen evidencias estadísticamente significativas para concluir que existe una dependencia entre el egreso y éstas últimas dos variables.

6.4. Del modelo estadístico utilizando la regresión logística con mejor bondad de ajuste que estima la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay.

Utilizando la técnica Stepwise para la selección de variables explicativas a introducir en el modelo de regresión logística para estimar la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay, fueron seleccionadas las siguientes variables, confirmando lo que han puesto en evidencia los estudios de Goberna (1987), House, Hurst, Keely (1996), Jiménez (1987), Wilson y Hardgrave (1995):

X_1: Cantidad de materias aprobadas en el primer semestre.

X_2: Calificación promedio en el primer semestre.

X_3: Cantidad de materias aprobadas en el segundo semestre.

X_{5i}: Tipo de Ingeniería.

X_{6i}: Sexo.

X_{7i}: Estado civil.

Con respecto a la interpretación de los parámetros del modelo, de acuerdo a los valores de los parámetros obtenidos en la Tabla 22 y sus respectivos OR de la Tabla 23 se concluyó lo siguiente:

Para variables cuantitativas.

Cuando aumenta la cantidad de materias aprobadas en el primer semestre aumenta la probabilidad de egreso del estudiante de ingeniería de la UNVES, más específicamente por cada unidad que aumente en la cantidad de materias aprobadas en el primer semestre, manteniendo fijas las demás variables, aumenta 1,65 veces las chances de egreso del estudiante de ingeniería de la UNVES.

Cuando aumenta la calificación promedio en el primer semestre aumenta la probabilidad de egreso del estudiante de ingeniería de la UNVES, más específicamente por cada unidad que aumente en la calificación promedio en el primer semestre, manteniendo fijas las demás variables, las chances de egreso del estudiante de ingeniería de la UNVES aumenta 1,38 veces.

Cuando aumenta la cantidad de materias aprobadas en el segundo semestre aumenta la probabilidad de egreso del estudiante de ingeniería de la UNVES, más específicamente por cada unidad que aumente en la cantidad de materias aprobadas en el segundo semestre, manteniendo fijas las demás variables, aumenta 1,68 veces las chances de egreso del estudiante de ingeniería de la UNVES.

Para variables cualitativas.

Tipo de Ingeniería: la probabilidad de egreso del estudiante de ingeniería de la UNVES aumenta cuando pasa de la categoría de referencia (Zootecnia) a las siguientes categorías de esta variable (Química, Azúcares, Ambiental), manteniendo fijas las demás variables; más específicamente, el estudiante de Ingeniería Química tiene 6,20 veces más chances de egreso que el estudiante de Ingeniería en Zootecnia; el estudiante de Ingeniería en Azúcares tiene 82,90 veces más chances de egreso que el estudiante de Ingeniería en Zootecnia; el estudiante de Ingeniería en Azúcares tiene 137.48 veces más chances de egreso que el estudiante de Ingeniería en Zootecnia. Sin embargo la probabilidad de egreso del estudiante de ingeniería de la UNVES disminuye cuando pasa de la categoría de referencia (Zootecnia) a la categoría (Informática), manteniendo fijas las demás variables; más específicamente, el estudiante de Ingeniería Zootecnia tiene 2,13 veces más chances de egreso que el estudiante de Ingeniería Informática.

Sexo: la probabilidad de egreso del estudiante de ingeniería de la UNVES aumenta cuando pasa de la categoría de referencia (Masculino) a la otra categoría de esta variable (Femenino), manteniendo fijas las demás variables; más específicamente, el estudiante de ingeniería de sexo femenino tiene 2,45 veces más chances de egreso que el estudiante de sexo masculino.

Estado civil: la probabilidad de egreso del estudiante de ingeniería de la UNVES aumenta cuando pasa de la categoría de referencia (Soltero) a la otra categoría de esta variable (Casado), manteniendo fijas las demás variables; más específicamente, el estudiante de ingeniería con estado civil casado tiene 1,22 veces más chances de egreso que el estudiante con estado civil soltero.

Con respecto a la bondad de ajuste del modelo, aplicando el test de Hosmer-Lemeshow como se observó en la Tabla 24 se concluyó que el modelo final es el apropiado para explicar la probabilidad de egreso del estudiante de ingeniería de la UNVES. Clasificando a los estudiantes con probabilidad de egreso mayor a 0,38 como estudiantes egresados se obtiene los índices para medir la bondad de ajuste del modelo final ajustado por regresión logística, siendo la Precisión igual a 80,27 %, la Especificidad igual a 81,95% y la Sensibilidad igual a 80,59%, como resultó en el Gráfico 12 y la Tabla 25; esto es, el modelo identifica 8 de cada 10 egresados aproximadamente, por lo que la capacidad predictiva del modelo matemático propuesto es excelente. Similarmente, como resultó en el Gráfico 11 el valor del área bajo la curva ROC, es igual a 0,8644612, esto significa que el modelo tiene capacidad de discriminación también excelente.

Finalmente, el modelo propuesto es validado mediante el análisis de los residuos, como se observó en el Gráfico 13 y la Tabla 26, sólo el 4,16% de estos residuos ajustados están fuera del intervalo ± 2, pero ninguno de estos errores es potencialmente influyente en

el modelo, ya que en el cálculo de las distancias de Cook arrojó un valor máximo mucho menor al valor límite 1, valor igual a 0,0007693125.

Por todo lo expuesto es posible cumplir con el objetivo general de la investigación cual es estimar la probabilidad de egreso en base a variables académicas y demográficas de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo de la república del Paraguay, y así comprobar la hipótesis de este trabajo de investigación, que las variables demográficas y las variables académicas estiman la probabilidad de egreso de los estudiantes de ingeniería de la Universidad Nacional de Villarrica del Espíritu Santo, cohorte 2009-2018.

Con esto es posible identificar al término del primer año de la carrera de ingeniería a los posibles estudiantes egresados, como así también lo más importante, identificar a los posibles estudiantes con probabilidades bajas de culminar la carrera, esto permitirá determinar de manera temprana a estos estudiantes y así poder implementar políticas educativas que mejoren la calidad académica de la Universidad para poder aumentar el número de egresados y así poder disminuir la brecha existente entre la matrícula y el egreso.

7. RECOMENDACIONES

Resulta razonable sugerir las siguientes recomendaciones:

1. Realizar en un futuro investigaciones similares en carreras de ingeniería a los efectos de verificar la permanencia o no de las variables explicativas que componen el modelo.
2. Efectuar investigaciones en otras carreras universitarias diferentes a las de ingeniería a fin de poseer un panorama más general de los estudiantes de toda la universidad.
3. Realizar una investigación con los datos en base a la teoría de Análisis de Supervivencia.
4. Realizar una investigación con los datos en base a la teoría de Árboles de Decisión.
5. Ampliar el estudio introduciendo otras variables explicativas, tales como variables académicas del curso probatorio de admisión, calificaciones obtenidas en la educación media y en la educación escolar básica, a fin de comprobar si resultan significativas.
6. Instaurar en las Universidades tanto públicas como privadas políticas relacionadas a la utilización de modelos predictivos para la probabilidad de egreso de sus estudiantes

REFERENCIAS

Acevedo, C., & Rocha, F. (2011). Estilos de aprendizaje, género y rendimiento académico. Journal of Learning Styles, 71-84.

Arkin, H. & Colton, R. R. (1995). Tables for statisticians. New York: Barnes & Noble.

Baird, K.E., & Elías, R. (2014). Factores Asociados al Logro Académico en Paraguay: Un Análisis Multinivel.

Bisquerra, R. (2009). Metodología de la Investigación educativa. Madrid: La Muralla.

Bisquerra, R. (1989). Métodos de Investigación Educativa Barcelona: CEAC.

Bobadilla de Almada, G. & la Red Martínez, D. (2017). Detección de Perfiles de Rendimiento Académico en la Universidad Nacional del Este de Paraguay. XIX. Workshop de Investigadores en Ciencias de la Computación. Volumen 1. 1149-1153. Disponible en: http://sedici.unlp.edu.ar/handle/10915/61343

Briones, G. (2003). Métodos y Técnicas de Investigación para las Ciencias Sociales. México: Trillas.

Bunge, M. (2000). La investigación Científica. Barcelona: Siglo XXI.

Campoy, T. (2016). Metodología de la investigación científica: Manual para la elaboración de tesis y trabajos de investigación. Asunción: Librería Cervantes.

Cardona, C. (2002). Introducción a los métodos de Investigación en Educación. Madrid: EOS.

Cook, R. (1977). Detection of Infuential Observations in Linear Regression. Technometrics. Vol. 19, pp. 15-18.

Cook, R. y Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, Vol. VIII. New York–London.

Di Gresia, L. (2007). Rendimiento académico universitario. Trabajo de tesis doctoral, Doctorado en Economía, Universidad Nacional de La Plata. Argentina.

Di Gresia, L., & Porto, A. (2004). Dinámica del desempeño académico. Departamento de Economía, Universidad Nacional de La Plata.

Di Gresia, L., & Porto, A., & Ripani, L. (2002). Rendimiento de los estudiantes de las Públicas Argentinas. Departamento de Economía, Universidad Nacional de La Plata. Argentina.

Fernández, P., y Pértegas, S. (2002). Investigación cuantitativa y cualitativa: Investigación Cuantitativa y Cualitativa. Pp. 76-78.

Fonseca Grandón, G. R. (2018). Trayectorias de permanencia y abandono de estudios universitarios: una aproximación desde el currículum y otras variables predictoras. (Spanish). Latin American Journal of Content & Language Integrated Learning, 21(2), 239–256. Disponible en: https://doi.org/10.5294/edu.2018.21.2.4

García J, M., Alvarado I, J.; & Jiménez, A. (2000). La predicción del rendimiento académico: Regresión Lineal versus Regresión Logística. Psicothema, 12(2); 248 - 252.

García Tinisaray, D. K (2015). Construcción de un modelo para determinar el rendimiento académico de los estudiantes basado en learning analytics (análisis del aprendizaje), mediante el uso de técnicas multivariantes (Tesis Doctoral). Universidad de Sevilla. Sevilla, España.

Garzón, R., Rojas, M., del Riesgo, L., & Pinzón, M. (2010). Factores que pueden influir en el rendimiento académico de estudiantes de Bioquímica que ingresan en el programa de Medicina de la Universidad del Rosario-Colombia. Educación médica, 85-96.

Glavinich, N. (2007). Guía breve para la preparación de trabajos de investigación según el Manual de Estilo de Publicaciones de la American Psychological Association (A.P.A.). Asunción. Universidad Autónoma de Asunción.

Goberna, M.A., López M.A. y Pastor J.T. (1987). La predicción del rendimiento como criterio para el ingreso en la universidad. Revista de Educación, 283, 235-248.

Goodman, L. y Kruskal, W. (1954). Measures of Association for Cross Classifications, Journal of the American Statistical Association, Vol. 49, Nro. 268, pp. 732-764.

Hernández, S., Fernández, C., y Baptista, P. (2013). Metodología de la Investigación, Sexta Edición. México: McGraw-Hill.

Hosmer, D. y Lemeshow, S. (2000). Applied Logistic Regression, Second Edition, Addison Wesley, E.E.U.U.

House, J. D., Hurst, R.S., Keely, E.J. (1996). Relacionship between learner attitudes, prior achievement, and performance in a General Education Course: A multi-Institutional Study. International Journal of instructional media, 23, 257-271.

Hueso, A., y Cascant, Mª J. (2012). Metodología y técnicas cuantitativas de investigación. Valencia: Universidad Politécnica de Valencia.

Jiménez Fernández, C. (1987). Rendimiento académico en la universidad a distancia. Un estudio empírico sobre su evolución y predicción (II). Revista de Educación, 284, 317-347.

Kerlinger, F., y Lee, H. (2001). Investigación del comportamiento. Métodos de investigación en ciencias sociales. México: McGraw-Hill.

Krejcie, R.V. & Morgan, D. W. (1970). Determining Sample Size for Research Activities. Determining Sample Size for Research Activities. Educational and Psychological Measurement. Volume: 30 issue: 3, pp: 607-610. Disponible en: https://doi.org/10.1177/001316447003000308

Latorre, A., Rincón, D. y Arnal, J. (1996). Bases metodológicas de la investigación educativa. Barcelona: Experiencia.

Mertens, D. M. (2010). Transformative Mixed Methods Research. Qualitative Inquiry, 16(6), 469–474. Disponible en: https://doi.org/10.1177/1077800410364612

Moral, J. (2006). Predicción del rendimiento académico universitario. Recuperado en junio de 2014, de Perfiles educativos: Recuperado de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-26982006000300003&lng=es&tlng=es

Pantoja, A., & Alcaide, M. (2013). La variable Género y su relación con el autoconcepto y el rendimiento académico de alumnado universitario. Revista científica electrónica de Educación y Comunicación en la Sociedad del Conocimiento, 124-140.

Porto, A. & Di Gresia, L. (2000). Características y Rendimiento de estudiantes universitarios. El caso de la Facultad de Ciencias Económicas de la Universidad Nacional de La Plata. Departamento de Economía, Universidad Nacional de La Plata.

Porto, A.; Di Gresia, L., & López A, M. (2004). Mecanismos de admisión a la Universidad y rendimiento de los estudiantes. Departamento de Economía, Universidad Nacional de La Plata.

Reyes Rocabado, J. & Escobar Flores, C. (2007). Una aplicación del modelo de regresión logística en la predicción del rendimiento estudiantil. Revista Estudios Pedagógicos XXXIII, Nro 2, 101-120.

Rodríguez Fontes, R, Díaz Rodríguez, P., Moreno Lazo, M., & Bacallao Gallestey, J. (2000). Capacidad predictiva de varios indicadores de selección para el ingreso a la carrera de medicina. Educación Médica Superior, 14(2), 128-135.

Salazar, A. (2011). Modelos de respuesta discreta en R y aplicación con datos reales. Universidad de Granada, España.

Valdés, K. & González Campos, J.A. (2019). Perfil de egreso doctoral: una propuesta desde el análisis documental y las expectativas de los doctorandos. IE Revista de Investigación Educativa de La REDIECH, (18), 161.

Vélez van Meerbeke, A. & Roa G, C. (2005). Factores asociados al rendimiento académico en estudiantes de medicina. Educación Médica, 8(2).

Wilson, R.L., Hardgrave, B.C. (1995). Predicting graduate student success in an MBA program: Regression versus classification. Educational and Psychological Measurement, 55, 186-195.

ANEXO

Ficha académica del estudiante de la UNVES.

FICHA ACADEMICA	N°:	N° Recibo:

Alumno/a :
Nro. Documento :
Año Lectivo de Ingreso :
Facultad :
Carrera :
Plan: :

Sexo :
Estado Civil :
Ciudad :
Departamento :

Curso :
Semestre:

ASIGNATURA	HS.	OBS.	ACTA N°	FECHA	PERIODO	CALIFICACION	MODALIDAD

Promedio:

Semestre:

ASIGNATURA	HS.	OBS.	ACTA N°	FECHA	PERIODO	CALIFICACION	MODALIDAD

Promedio:

Curso :
Semestre:

ASIGNATURA	HS.	OBS.	ACTA N°	FECHA	PERIODO	CALIFICACION	MODALIDAD

Promedio:

Semestre:

ASIGNATURA	HS.	OBS.	ACTA N°	FECHA	PERIODO	CALIFICACION	MODALIDAD

Promedio:

Curso :
Semestre:

ASIGNATURA	HS.	OBS.	ACTA N°	FECHA	PERIODO	CALIFICACION	MODALIDAD

		Promedio:					
Semestre:							
ASIGNATURA	HS.	OBS.	ACTA N°	FECHA	PERIODO	CALIFICACION	MODALIDAD

		Promedio:					
Curso :							
Semestre:							
ASIGNATURA	HS.	OBS.	ACTA N°	FECHA	PERIODO	CALIFICACION	MODALIDAD

		Promedio:					
Semestre:							
ASIGNATURA	HS.	OBS.	ACTA N°	FECHA	PERIODO	CALIFICACION	MODALIDAD

		Promedio:					
Curso :							
Semestre:							
ASIGNATURA	HS.	OBS.	ACTA N°	FECHA	PERIODO	CALIFICACION	MODALIDAD

Promedio:	

La escala de calificaciones es la siguiente

1- Insuficiente 2- Aprobado 3- Bueno 4- Distinguido 5- Sobresaliente

Observación

1. Esta ficha contiene datos cargados en el Sistema Académico Tekombo'e.

2. Los datos contenidos en la presente ficha están sujetos al control de la Facultad y de la Dirección General Académica del Rectorado.

3. El presente documento es de uso interno de la UNVES y del alumno solicitante y la misma por lo tanto no tiene equivalencia a un certificado de estudios.
4. En caso de dudas de alguno de los datos contenidos en la presente ficha académica el alumno podrá recurrir por escrito a su facultad, adjuntando la ficha.

Funcionario Autorizado por Resolución del Rectorado