Universidad Católica de Santa María

Escuela de Postgrado

Maestría en Matemática

Técnicas de análisis multivariado aplicadas al análisis del rendimiento académico

Tesis presentada por el Bachiller:

Ure Benavides, Nestor Alvaro

ORCID: 0009-0001-6528-1179

Para optar el Grado Académico de Maestro en Matemática

Asesor:

Dr. Cáceres Huambo, Alberto

ORCID: 0000-0002-9767-4946

Arequipa - Perú
2025

UCSM-ERP

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

ESCUELA DE POSTGRADO

DICTAMEN APROBACIÓN DE BORRADOR DE TESIS

Arequipa, 29 de Febrero del 2024

Dictamen: 005192-C-EPG-2024

Visto el borrador del expediente 005192, presentado por:

1993042581 - URE BENAVIDES NESTOR ALVARO

Titulado:

TÉCNICAS DE ANÁLISIS MULTIVARIADO APLICADAS AL ANÁLISIS DEL RENDIMIENTO ACADÉMICO

Nuestro dictamen es:

APROBADO

29281453 - DIAZ BASURCO LUIS FERNANDO
DICTAMINADOR

80199482 - CUEVAS ARIZACA EDY ELAR
DICTAMINADOR

45482246 - ORTIZ ROMERO DERLY DAVID
DICTAMINADOR

Técnicas de análisis multivariado aplicadas al análisis del rendimiento académico

INFORME DE ORIGINALIDAD

10%	11%	1%	4%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	idoc.pub Fuente de Internet	1%
2	hdl.handle.net Fuente de Internet	1%
3	pt.scribd.com Fuente de Internet	1%
4	1library.co Fuente de Internet	1%
5	pitagoras.sabermatematicas.com Fuente de Internet	1%
6	ddd.uab.cat Fuente de Internet	1%
7	www.coursehero.com Fuente de Internet	1%
8	repositorio.unsa.edu.pe Fuente de Internet	1%
9	repositorio.unsaac.edu.pe Fuente de Internet	1%
10	acreditacion.cs.buap.mx Fuente de Internet	1%

DEDICATORIA

A Dios por todo su amor, su protección y por todo lo que me ha dado en la vida, a la memoria de mis queridos papá y mamá que siempre están en mi corazón, a mi esposa Janet y a mis queridos hijos: Joshua y por siempre mi Francesco, y a mis hermanos y Bb que animan y motivan mi vida, todos ellos.

Néstor

AGRADECIMIENTO

Un agradecimiento especial a mis asesores y a mis amigos que en todo momento mostraron su apoyo y aportes para la concretización de este trabajo.

RESUMEN

En este trabajo se describe el modelo de regresión logística, el cual es ampliamente utilizado en diferentes áreas como la ciencia social y la medicina. El objetivo de esta investigación es aplicar el modelo de regresión logística en problemas donde se requiera un estudio observacional, de caso-control, retrospectivo y de desarrollo, como es el rendimiento académico en la materia de Algebra y geometría y su relación con las matrices de evaluación y la metodología empleada en el desarrollo de la asignatura, usando un software estadístico como el lenguaje de programación R. Metodología.- Para explicar la estimación de los parámetros de este modelo se utilizará métodos numéricos para resolver sistemas de ecuaciones no lineales tales como el método de Newton. También es posible utilizar técnicas de análisis de conglomerados, donde los algoritmos aprenden y mejoran automáticamente en base a la experiencia. Para verificar los supuestos del modelo, se realizaron pruebas de hipótesis y análisis gráfico en los datos mostrados de las notas de proceso de permanente 1 y de permanente 2 y en los exámenes Parcial y Final. Resultados. - Se obtuvo como resultados que hay una relación directa entre las evaluaciones de permanente 1 y permanente 2 con la condición de aprobar la asignatura, validándose de este modo al modelo con un nivel de significancia del 5%. También se determinó la incidencia que hay entre la metodología de enseñanza y el uso de una matriz de evaluaciones en el rendimiento académico, empleando técnicas de clasificación, logrando determinar los periodos lectivos en los que se evidencia un mejoramiento en el rendimiento de los estudiantes, como es el caso de los periodos lectivos de verano. Lo que finalmente significa que el análisis logístico, atendiendo a las singularidades propias de la asignatura, los tipos de evaluaciones, los momentos de aprendizaje y las características de las matrices de evaluación aplicadas, son predictores para analizar rendimiento del estudiante en el presente trabajo.

Palabras clave: Regresión lineal, modelo logístico, estimación, conglomerados, significancia.

ABSTRACT

This work describes the logistic regression model, which is widely used in different areas such as social science and medicine. The objective of this research is to apply the logistic regression model in problems where an observational, case-control, retrospective and developmental study is required, such as academic performance in the subject of Algebra and geometry and its relationship with the matrices of evaluation and the methodology used in the development of the subject, using statistical software such as the R programming language. Methodology. - To explain the estimation of the parameters of this model, numerical methods will be used to solve systems of nonlinear equations such as the Newton's method. It is also possible to use cluster analysis techniques, where algorithms automatically learn and improve based on experience. To verify the assumptions of the model, hypothesis tests and graphical analysis were carried out on the data shown in the permanent 1 and permanent 2 process notes and in the Midterm and Final exams. Results. - The results were obtained that there is a direct relationship between the evaluations of permanent 1 and permanent 2 with the condition of passing the subject, thus validating the model with a significance level of 5%. The incidence between the teaching methodology and the use of an evaluation matrix on academic performance was also determined, using classification techniques, managing to determine the school periods in which an improvement in student performance is evident, such as This is the case of the summer school periods. Which finally means that the logistical analysis, taking into account the singularities of the subject, the types of evaluations, the learning moments and the characteristics of the evaluation matrices applied, are predictors to analyze student performance in the present work.

Keywords: Linear regression, logistics model, estimate, conglomerates, significance.

ÍNDICE GENERAL

Índice de tablas

Índice de figuras

INTRODUCCIÓN

La regresión logística es una técnica estadística multivariante que permite relacionar una variable de respuesta categórica en función de variables predictoras numéricas o categóricas. El objetivo del análisis de regresión logístico es poder realizar predicciones, es decir, estimar la probabilidad de un suceso. Esta técnica es parte de los modelos lineales generalizados, introducidos por (McCullagh, P., & Nelder, 1989), apoyándose también en los estudios realizados de regresión multivariada para datos categóricos por (Liang, K., Zeger, S. & Qaqish, B., 1992) y se aplica en diferentes áreas de la ciencia como la epidemiologia, ecología, sociología, en los sectores bancario y asegurador, entre otros.

El modelo de regresión logística, es ampliamente utilizado en diferentes áreas de las ciencias sociales y la medicina, lo que ha permitido realizar también estudios acerca de los factores que influyen en el progreso de los estudiantes en universidades abiertas realizados por (Arifin, M.H., 2016). En el presente trabajo se pretende aplicar el modelo de regresión logística en problemas donde se requiera un estudio observacional, usando un software estadístico como el lenguaje de programación R, que permita implementar un algoritmo y compararlo con un software comercial tal como el programa SPSS. Se fundamentará la teoría necesaria para establecer un modelo predictivo, que a su vez utilice técnicas de clasificación y que permitan agrupar los objetos de estudio de acuerdo a su alta homogeneidad al interior del grupo al que pertenece y heterogeneidad al exterior, en comparación con otros de características no similares.

HIPÓTESIS

Dado que el rendimiento académico de los estudiantes en la asignatura de Algebra Lineal y Geometría requiere ser analizado mediante un modelo basado en regresión logística y mediante técnicas multivariadas de agrupación o clasificación y que, al ser debidamente fundamentados, es probable que permitan evaluar la incidencia de la metodología de enseñanza y el uso de una matriz de evaluación en la identificación de las actividades evaluativas más influyentes.

OBJETIVOS

i. Describir el modelo lineal $logit$ basado en regresión logística para analizar el rendimiento académico de los estudiantes universitarios y analizar la metodología de enseñanza mediante técnicas de agrupación multivariada, correspondientes a estudiantes de la Universidad Católica San Pablo, de la asignatura de Algebra lineal y Geometría desde el periodo lectivo 2008-II al 2019-II.

ii. Estimar los parámetros del modelo de regresión logística mediante la implementación de un programa y comparar dicha estimación de los parámetros obtenidos, con la estimación obtenida mediante un software libre tal como el que ofrece las librerías del CRAN del lenguaje de programación R, así como con la estimación obtenida mediante un programa comercial como por ejemplo el programa SPSS.

iii. Analizar los registros de notas usando el modelo logístico para determinar la incidencia que tiene la metodología de enseñanza y el uso de una matriz de evaluaciones en el rendimiento académico.

iv. Analizar las actividades de evaluación más influyentes de la metodología de enseñanza mediante técnicas de agrupación o clasificación.

CAPÍTULO I

MARCO TEÓRICO

El análisis multivariado proporciona técnicas estadísticas que son utilizadas para establecer relaciones entre variables, reconocer patrones o eliminación de variables poco representativas. Es posible establecer modelos de predicción mediante modelos de regresión múltiple donde el número de variables, representa un problema y no permite establecer fácilmente la relación entre grupos de variables.

Actualmente las técnicas estadísticas multivariadas están implementadas en diversos softwares estadísticos, sin embargo, cada caso de estudio determina un Modelo matemático, que debe satisfacer condiciones específicas, para ello la formulación del modelo y su construcción debe estar fundamentada correctamente para su uso. Se propone analizar el rendimiento académico de estudiantes de la universidad, para lo cual se establecerá un modelo de predicción, que permita evaluar estrategias de aprendizaje y predecir el rendimiento académico de estudiantes universitarios.

1.1. Antecedentes del problema

Generalmente los modelos de regresión múltiple son utilizados en investigación en el área ciencias de la salud, con el objetivo de explicar las interrelaciones que existen entre ciertas variables o para determinar los factores que afectan a la presencia o ausencia de un episodio adverso determinado. (Núñez et al., 2011)

En un trabajo de investigación (Yolvi, 2011), el objetivo fue dar a conocer cuáles son las variables académicas y la importancia de éstas en el rendimiento académico de los estudiantes universitarios. Se describe de igual manera, el proceso de cambios de la

educación superior en el último siglo, tanto a nivel mundial como nacional, se describen y analizan las variables académicas que influyen en el rendimiento de los estudiantes universitarios. A pesar del riesgo que implica usar exclusivamente las calificaciones para medir el rendimiento académico en educación superior, debido fundamentalmente a la subjetividad de los docentes, las calificaciones no dejan de ser el medio más usado para operacionalizar el rendimiento académico.

Con el objetivo de determinar los factores asociados al rendimiento académico, para el año lectivo 2010-2011, de los estudiantes de primer año de las carreras de Trabajo Social, Ingeniería, Derecho y Humanidades de la Universidad de Atacama (UDA), se tomó un total de 258 alumnos. Primeramente, se utilizó un modelo de Regresión Múltiple con datos de corte transversal para determinar las variables predictoras del rendimiento académico; posteriormente se estimó las variables que inciden en la probabilidad de mejorar el rendimiento académico del estudiante haciendo uso del Modelo de Regresión Logística. Luego del análisis, se estableció que las variables género, estudia y trabaja, conformidad con la carrera, notas de prueba verbal y matemática, resultaron ser estadísticamente significativas. Es decir, tendrían un efecto positivo sobre el rendimiento académico del estudiante. (Cabana et al., 2018)

Se reconoce en la enseñanza universitaria como muy importante la estadística para aplicar el rigor de carácter científico. Para comparar el rendimiento académico en bioestadística y la competencia disciplinar de pensamiento matemático en estudiantes universitarios del área de ciencias biológicas, se realizó una investigación transversal y exploratoria. Se consideró 187 alumnos de ambos géneros con edades entre 19 a 20 años que cursaron de manera regular la materia. Se halló una asociación significativa entre aquellos que aprobaron el examen de bioestadística y el nivel de pensamiento matemático que ostentaban los estudiantes en sus pruebas de ingreso. Con lo cual se concluyó que si se posee

un nivel alto de pensamiento matemático se cuenta con una probabilidad del 90% de aprobar el examen de bioestadística (Cantú Martínez & Santoyo Stephano, 2019).

En el proyecto denominado "Impacto de variables institucionales y pedagógicas en el rendimiento académico de los estudiantes universitarios" de la Facultad de Ciencias Económicas y Estadística de Rosario, se viene estudiando la problemática del rendimiento académico de los estudiantes universitarios desde el año 2010. Se logró identificar qué aspectos resultan relevantes, desde la opinión estudiantil, para su propio rendimiento. En el año 2013, se analizó los factores de mayor impacto en el rendimiento académico de un grupo de estudiantes, según su propia perspectiva, replicando el estudio anterior. Se escogió deliberadamente para el estudio a los alumnos cursantes de otra asignatura de cuarto año que presenta el menor porcentaje de promoción de la carrera Contador Público en el año 2011. Estos hallazgos coinciden con los datos obtenidos del relevamiento del rendimiento académico estudiantil durante el trienio 2010 a 2012 en asignaturas con cursado regular plasmado en Cavallo, M. y Vázquez, C. (2013) (Enrique & Pacheco, 2016)

1.2. Marco Teórico

Las estrategias de aprendizaje se entienden como un conjunto organizado, consciente e intencional de lo que hace el aprendiz para lograr con eficacia un objetivo en un contexto social dado (Remesal, A. F., López, B. G., & Rodríguez, 2007).

La interactuación que puede tener el docente con su alumnado es a través de estilos interpersonales muy diferentes, un estilo controlador frente al apoyo a la autonomía. El estilo controlador hace mención al uso de presiones externas, actuar de manera preconcebida, uso

de medios coercitivos, imposiciones, etc., que son entendidos por el alumnado origen de sus comportamientos, afectando de forma negativa el esfuerzo personal, la iniciativa personal y el autoconocimiento (Trigueros Ramos & Navarro Gómez, 2019).

En un trabajo de investigación realizado por (Roux & Anzures González, 2015) de 41 estrategias de aprendizaje cuyo análisis de fiabilidad resultó adecuado y fueron analizadas en relación con el rendimiento académico, únicamente 19 obtuvieron una correlación positiva significativa. La estrategia que obtuvo una mayor relación con el rendimiento académico de todas las analizadas en el cuestionario fue la de tomo apuntes en clase. Asimismo, quienes son capaces de tomar apuntes relevantes en clase fueron quienes obtuvieron mayores calificaciones.

Según (Yolvi, 2011), los investigadores suelen considerar un conjunto bastante amplio de variables académicas asociadas al rendimiento en la educación superior, entre las cuales destacan las que se analizan a continuación:

i. **Características académicas del colegio de procedencia:** Saber si el tipo de formación que se brinda en los colegios de procedencia influye sobre el rendimiento académico universitario, ha sido motivo de investigaciones en países latinoamericanos y de otras latitudes. Los tipos de colegios de donde proceden los estudiantes universitarios peruanos son: Por el financiamiento, por la admisión de estudiantes según su sexo, por la evaluación de ingreso a sus alumnos, por la cantidad de estudiantes, por el acompañamiento, por los niveles académicos, por la cantidad de horas, por si es bilingüe.

ii. **El rendimiento escolar:** Las calificaciones son muchas veces el indicador principal que permitirá conocer si el estudiante universitario podrá tener éxito en su vida universitaria o no. El rendimiento previo explica el rendimiento

presente, dado que, por un lado, sintetiza las aptitudes y el esfuerzo del estudiante y, por otro, mide el nivel de conocimientos previos; es decir, la solidez sobre los cuales se asociarán los nuevos conocimientos.

iii. **El rendimiento en las evaluaciones de aptitud y los exámenes de admisión:** En nuestro país, el examen de admisión aún es un instrumento de selección utilizado para decidir el ingreso de un estudiante a la universidad, pero no es uniforme el tipo de evaluación a aplicar. Existen universidades que toman una sola prueba para todos sus postulantes; otras toman pruebas diferenciadas según la carrera a que se postula.

iv. **El rendimiento previo en cursos prerrequisitos u otras asignaturas universitarias:** De acuerdo a (Tejedor, 2003) se determinó que la nota media del alumno en el periodo bianual anterior era el mejor explicador del rendimiento académico universitario.

v. **El esfuerzo y los efectos de las cargas laborales o académicas:** La variable esfuerzo se puede operacionalizar tomando en cuenta los siguientes criterios:

- La asistencia a clases
- Estrategias de estudio
- Trabajos presentados a tiempo y aprobados
- La participación durante la clase

a mayor asistencia a clases, dedicarle más horas a la semana al estudio, entregar puntualmente los trabajos académicos y participar activamente durante las sesiones, el rendimiento académico será definitivamente mayor.

vi. **La vocación del estudiante:** Los cursos, talleres o seminarios de orientación vocacional son hoy una necesidad que busca guiar al estudiante escolar a tomar la decisión correcta según sus capacidades e intereses. No podemos negar que la presión del mercadeo y la imagen social de muchas carreras influyen también en la decisión de los adolescentes.

vii. **Las facilidades académicas:** La infraestructura equipada con tecnología, aulas multimedia, wifi, bibliotecas amplias y actualizadas, etc. se asocia a la posibilidad de tener un mayor rendimiento académico.

1.3. Marco Conceptual

El análisis de regresión es una técnica estadística utilizada para investigar la relación entre variables dependientes y un conjunto de variables independientes. En el método de regresión, la variable dependiente es un predictor. El análisis de clúster nos permitirá la clasificación de individuos en grupos homogéneos.

1.3.1. Regresión lineal simple

El análisis de regresión, es una técnica estadística para el modelado y la investigación de la relación entre dos o más variables que permite determinar una función que se ajuste a los puntos de un diagrama de dispersión con la finalidad de poder predecir en forma aproximada una de las variables a través de la otra.

En algunas aplicaciones que provienen de un proceso experimental existe una sola variable dependiente o de respuesta ***y*** que no se controla en el experimento. La respuesta depende de una o más variables independientes (regresoras).

Si existe una sola variable regresora $\boldsymbol{x}$ entonces los pares(x_i, y_i) corresponden a una variable aleatoria Y, donde

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad \ldots (1)$$

Es la ecuación del Modelo de regresión lineal simple

Sin embargo

$$E(Y) = \beta_0 + \beta_1 x \quad \ldots (2)$$

Es la ecuación de regresión lineal simple

Es decir, en el modelo de regresión lineal simple ε de la ecuación (1) representa el error que explica la variabilidad de y, ε es una variable aleatoria y se espera que su media o valor esperado sea cero $E(\varepsilon) = 0$

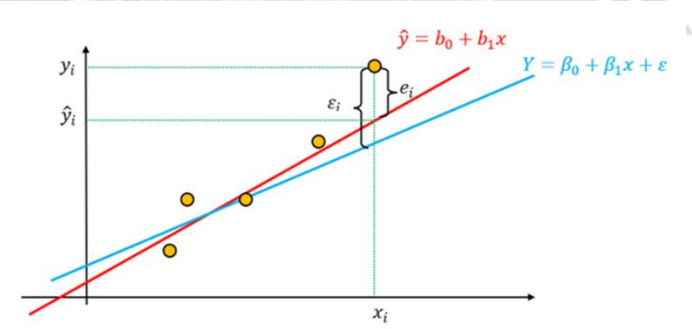

Figura 1: Recta de regresión lineal simple y residuos $e_i = y_i - \hat{y}_i$
Fuente: Elaboración propia

Estimación de la recta de regresión

Estimar la recta de regresión consiste en estimar los coeficientes de la regresión β_0 y β_1 para obtener la recta:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{ó} \quad \hat{y} = b_0 + b_1 x \qquad \ldots (3)$$

Donde $\hat{y}$ denota el valor de y predicho por la recta para el valor observado de x = x.

Si solamente se requiere determinar la recta, basta con considerar el criterio de Mínimos Cuadrados, este criterio también denominado minimización del error cuadrático medio, consiste en minimizar las distancias entre los puntos observados y los predichos por la recta de ajuste, como se observa en la figura 2.

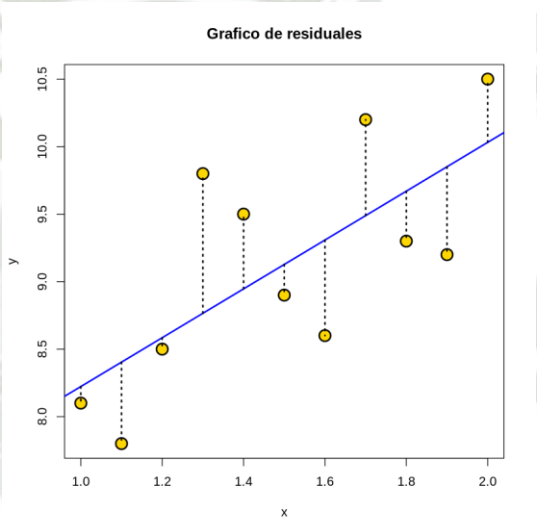

Figura 2: Grafica de residuos
Fuente: Elaboración propia

La desviación vertical del punto (x_i, y_i) desde la recta $\hat{y} = b_0 + b_1 x$ es la altura del punto a la altura de la recta $e_i = y_i - (b_0 + b_1 x_i)$. La suma de las desviaciones verticales cuadradas desde los puntos $(x_1, y_1), \ldots, (x_n, y_n)$ a la recta es entonces

$$f(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 \qquad \ldots (4)$$

Los puntos estimados de β_0 y β_1, denotados por $\hat{\beta}_0$ y $\hat{\beta}_1$ a los cuales se les llama estimaciones de mínimos cuadrados, son aquellos valores que minimizan $f(b_0, b_1)$ de la

ecuación (4). Es decir $\hat{\beta}_0$ y $\hat{\beta}_1$ son tales que $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1)$ para cualesquiera b_0 y b_1. La recta de regresión estimada o recta de mínimos cuadrados es entonces una recta cuya ecuación es $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Luego, las siguientes igualdades se obtienen de la ecuación (4)

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2\big(y_i - (b_0 + b_1 x_i)\big)(-1) = 0 \qquad \ldots (5)$$
$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2\big(y_i - (b_0 + b_1 x_i)\big)(-x_i) = 0$$

Las que dan lugar a las Ecuaciones normales,

$$n b_0 + \Big(\sum x_i\Big) b_1 = \sum y_i \qquad \ldots (6)$$
$$\Big(\sum x_i\Big) b_0 + \Big(\sum {x_i}^2\Big) b_1 = \sum x_i\, y_i$$

resolviendo el sistema de ecuaciones para *b* de la ecuación (6) se obtiene:

$$b = b_1 = \hat{\beta}_1 = \frac{\sum(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{y_i} - \overline{\boldsymbol{y}})}{\sum(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2} \quad \text{o} \quad b = \frac{n\sum(\boldsymbol{x_i y_i}) - (\sum \boldsymbol{x_i})(\sum \boldsymbol{y_i})}{n\sum \boldsymbol{x_i^2} - (\sum \boldsymbol{x_i})^2} \qquad \ldots (7)$$

y dividiendo la primera ecuación normal de (6) se obtiene

$$a = b_0 = \hat{\beta}_0 = \bar{y} - b\bar{x} \qquad \ldots (8)$$

Ejemplo: Los siguientes datos son temperatura "***y***" (°C), humedad "***x***" *(%)*.

Tabla 1
Datos de temperaturas y humedad

x	y
20	21
25	18
30	17
35	16
50	14
55	13
60	10

Fuente: Elaboración propia

Para analizar estos datos usaremos el lenguaje de programación R. Primeramente, hallamos su diagrama de dispersión, obteniendo

```
> x = c(20,25,30,35,50,55,60)
> y = c(21,18,17,16,14,13,10)
```

```
> plot(x, y)
```

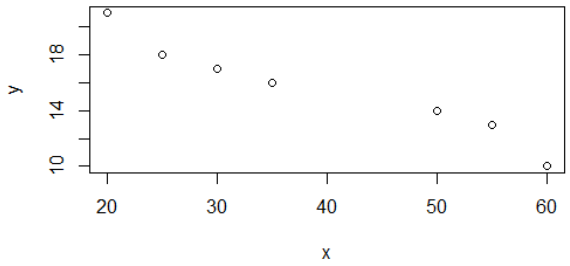

Figura 3: Diagrama de dispersión
Fuente: Elaboración propia

observamos gráficamente que aparentemente existe una relación lineal inversa, lo cual se verifica hallando su coeficiente de correlación lineal

```
>cor(x,y)
          [1] -0.9674
```

lo cual indica que existe una relación lineal dado que su correlación es fuerte y negativa la recta de mejor ajuste obtenida por mínimos cuadrados la obtenemos como sigue

```
> regLIN=lm(y~x)
> regLIN

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)          x
24.3058         -0.2223
```

de aquí la recta de mejor ajuste está dada por $y = 24.3058 - 0.2223x$ y su gráfica se obtiene como sigue

```
>abline(regLIN)
```

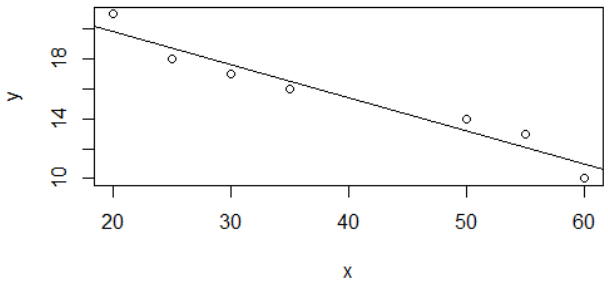

Figura 4: Diagrama de dispersión y gráfica de la recta de regresión
Fuente: Elaboración propia

En conclusión, tenemos que siendo $r = -0.9674$, su coeficiente de determinación es $r^2 = 0.9359$. Por lo cual, de acuerdo a los datos, la relación entre la humedad y la temperatura es fuerte e inversa, a mayor temperatura menor humedad, el 93.59% de los datos se describen a través de la recta de regresión, correspondiendo una descripción de 6.55 de los 7 datos a través de la recta de regresión.

Para determinar si las variables tienen una relación significativa tenemos

```
> summary(regLIN)

Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5       6       7
 1.1408 -0.7476 -0.6359 -0.5243  0.8107  0.9223 -0.9660

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.30583    1.08955  22.308 3.36e-06 ***
x           -0.22233    0.02602  -8.545 0.000361 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9981 on 5 degrees of freedom
Multiple R-squared:  0.9359, Adjusted R-squared:  0.9231
F-statistic: 73.02 on 1 and 5 DF,  p-value: 0.0004
```

Obtenemos que $Valor\ P = 0.0004 < 0.05$ lo cual nos indica que existe una relación significativa entre las variables.

1.3.2. Regresión múltiple

El análisis de regresión múltiple, consiste en determinar como una variable dependiente **y** se relaciona con dos o más variables independientes $x_1, x_2, x_3, \dots, x_k$

El modelo de regresión lineal múltiple, tiene la siguiente forma

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p + \varepsilon \quad \dots (9)$$

La ecuación de regresión múltiple está dada por

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \quad \dots (10)$$

La ecuación de regresión múltiple estimada es

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_k x_k \quad \dots (11)$$

donde

$b_1, b_2, b_3, \dots, b_k$ son los estimadores de $\beta_1, \beta_2, \beta_3, \dots, \beta_k$ de la ecuación (10)

$\hat{y}$ es el valor estimado de la variable dependiente y de la ecuación (11)

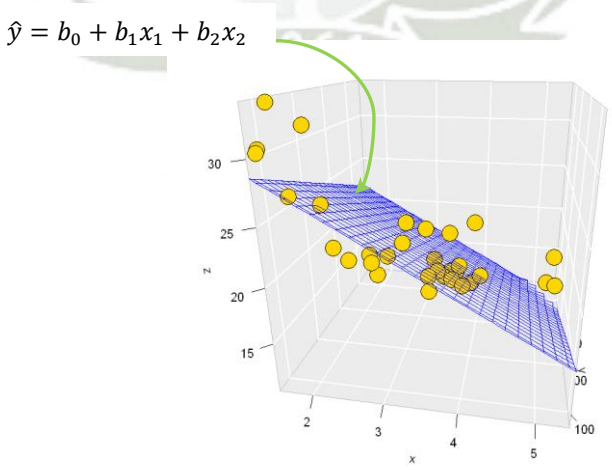

Figura 5: Regresión lineal múltiple en $\mathbb{R}^3$
Fuente: Elaboración propia

Ejemplo: Se sometió a prueba un grupo de camiones ligeros con motores que utilizan diésel como combustible para saber si la humedad, la temperatura del aire y la presión barométrica influyen en la cantidad de óxido nitroso que emiten (en ppm). Las emisiones se midieron en distintos momentos y en diversas condiciones experimentales. Los datos se presentan en la siguiente tabla

Tabla 2
Factores influyentes en la emisión de óxido nitroso

Oxido	Humedad	Temperatura	Presión
0.9	72.4	76.3	29.18
0.91	41.6	70.3	29.35
0.96	34.3	77.1	29.24
0.89	35.1	68	29.27
1	10.7	79	29.78
1.1	12.9	67.4	29.39
1.15	8.3	66.8	29.69
1.03	20.1	76.9	29.48
0.77	72.2	77.7	29.09
1.07	24	67.7	29.6
1.07	23.2	76.8	29.38
0.94	47.4	86.6	29.35
1.1	31.5	76.9	29.63
1.1	10.6	86.3	29.56
1.1	11.2	86	29.48
0.91	73.3	76.3	29.4
0.87	75.4	77.9	29.28
0.78	96.6	78.7	29.29
0.82	107.4	86.8	29.03
0.95	54.9	70.9	29.37

Fuente: Elaboración propia

Suponiendo que los datos se ajustan a un modelo de la forma

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Podemos hallar el modelo lineal de mejor ajuste, para lo cual luego de importar los datos al lenguaje de programación R

```
> library(readxl)
> reg_multiple <- read_excel("C:/Users/Desktop/reg_multiple.xlsx")
> View(reg_multiple)
```

```
> regMultiple=lm(reg_multiple$Oxido~reg_multiple$Humedad+reg_multiple$Temperatura
+reg_multiple$Presion)
> regMultiple

Call:
lm(formula = reg_multiple$Oxido ~ reg_multiple$Humedad + reg_multiple$Temperatura +
    reg_multiple$Presion)

Coefficients:
             (Intercept)      reg_multiple$Humedad  reg_multiple$Temperatura
                -3.5078                   -0.0026                    0.0008
    reg_multiple$Presion
                 0.1542
```

Por lo cual la ecuación de estimación de mejor ajuste está dada por

$$\hat{y} = -3.5078 - 0.0026x_1 + 0.0008x_2 + 0.1542x_3$$

luego para verificar si existe una relación significativa tenemos que importando los datos al lenguaje de programación R:

```
> summary(regMultiple)

Call:
lm(formula = reg_multiple$Oxido ~ reg_multiple$Humedad + reg_multiple$Temperatura +
    reg_multiple$Presion)

Residuals:
     Min       1Q   Median       3Q      Max
-0.1178  -0.0253   0.0135   0.0410   0.06523

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -3.5078    3.0049      -1.167  0.2602
reg_multiple$Humedad     -0.0026     0.0007     -4.008  0.0010 **
reg_multiple$Temperatura  0.0008     0.0020      0.391  0.7012
reg_multiple$Presion      0.1542     0.1014      1.521  0.1478
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0562  on 16 degrees of freedom
Multiple R-squared:  0.8005, Adjusted R-squared:  0.763
F-statistic:  21.4 on 3 and 16 DF,  p-value: 7.609e-06
```

obteniendo que $Valor\ P = 7.609e - 06 < 0.05$ lo cual indica que existe una relación significativa entre las variables.

Así al utilizar la ecuación de mejor ajuste como estimador, podemos hallar la cantidad de óxido nitroso que emiten los camiones en las siguientes condiciones: 50% de humedad, temperatura de 76˚F y una presión barométrica de 29.30 obteniendo

$$\hat{y} = -3.5078 - 0.0026(50) + 0.0008(76) + 0.1542(29.30) = 0.9384ppm$$

Los modelos lineales pueden ser descritos matricialmente, lo cual permite determinar sus coeficientes usando mínimos cuadrados.

1.3.3. Modelo de regresión lineal en forma matricial

Al ajustar un modelo de regresión lineal múltiple, en particular cuando contiene más de dos variables, tenemos que el experimentador tiene k variables independientes $x_{1,}x_{2,}\dots,x_k$ y n observaciones $y_{1,}y_2,\dots,y_n$, cada una de las cuales se puede expresar con la ecuación

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i \quad \dots(12)$$

$$\mu_{Y|_{x_1,x_2,\dots,x_k}} = E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \quad \dots(13)$$

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} \quad \dots(14)$$

Este modelo representa en esencia a n ecuaciones que describen como se generan los siguientes valores de respuesta durante el proceso científico. Si usamos la notación matricial. Podemos escribir la ecuación siguiente

$$\boldsymbol{y} = \boldsymbol{X\beta} + \epsilon \quad \dots(15)$$

$$\hat{\boldsymbol{y}} = \boldsymbol{Xb} \quad \dots(16)$$

donde

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

el método de mínimos cuadrados para la estimación de β, implica calcular b, para lo cual

$$SCE = (y + Xb)^t(y - Xb) \quad \ldots(17)$$

se minimiza. Este proceso de minimización implica resolver para b en la ecuación (17),

$$\frac{\partial}{\partial b}(SCE) = 0$$

El resultado se reduce a determinar b en

$$(X^tX)b = X^ty \quad \ldots(18)$$

Observamos que en la matriz X, el $i - ésimo$ renglón representa los valores de x que dan lugar a la respuesta y_i, denotando por

$$A = X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix}$$

$$g = X'y = \begin{bmatrix} g_0 = \sum_{i=1}^{n} y_i \\ g_1 = \sum_{i=1}^{n} x_{1i}y_i \\ \vdots \\ g_k = \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix}$$

Nos permite escribir las ecuaciones normales en la forma de matriz

$$Ab = g$$

Si la matriz A es no singular, la solución para los coeficientes de regresión se escribe como

$$b = A^{-1}g = \left(X'X\right)^{-1}(X'y) \quad \ldots(19)$$

De esta manera, obtenemos la ecuación de predicción o regresión resolviendo un conjunto de $k + 1$ ecuaciones con un número igual de incógnitas. Esto implica el hallar la inversa de la matriz $X^t X$ de orden $k + 1$ por $k + 1$.

1.3.4. Regresión logística

Para estimar un modelo de regresión cuando la variable dependiente es una variable continúa, el modelo de regresión más usado es la regresión lineal, pero cuando la variable de interés es dicotómica surgen ciertos inconvenientes como:

a) La distribución de los errores aleatorios no es normal

b) No acota la probabilidad, ya que los valores predichos pueden tomar valores que no estén dentro de 0 y 1

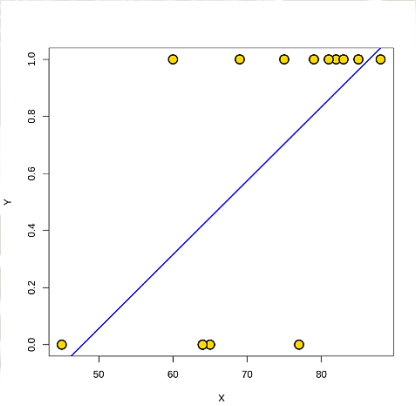

Figura 6: Modelo lineal ajustado a valores binarios
Fuente: Elaboración propia

La regresión logística resuelve este tipo de problema usando una función no lineal como es la función logística. Con esta función se pueden efectuar predicciones comprendidas entre un mínimo y un máximo

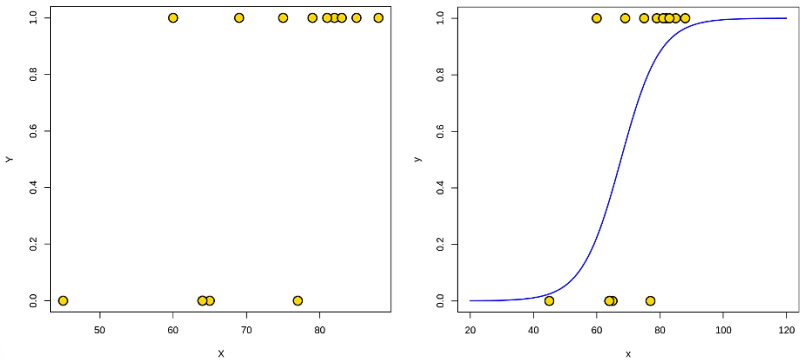

Figura 7: Modelo Logístico ajustado a valores binarios
Fuente: Elaboración propia

El modelo de regresión logística binaria considera dos sucesos de un fenómeno o variable Y, excluyentes que se codifican con valores 0 y 1. Si la probabilidad de que ocurra un evento A es p, la probabilidad de que no ocurra el evento A es a $1 - p$, es decir

$$P[Y = 1] = p$$

$$P[Y = 0] = 1 - p$$

Los valores x de una variable aleatoria X, pueden utilizarse para relacionar la probabilidad de que Y tome el valor de 1 con el valor x de la variable X, así la función logística puede definirse como

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \quad \ldots (20)$$

La variable Y es de Bernoulli, por lo cual la distribución condicional de Y sobre $X = x$ es una distribución de Bernoulli

$$P[Y = 1|X = x] = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

$$P[Y = 0|X = x] = 1 - \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

de forma equivalente, como

$$\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

entonces

$$P[Y=1|X=x] = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad \ldots(21)$$

Los valores de la función logística oscilan entre 0 y 1, lo cual puede ser verificado utilizando límites a la ecuación (21)

$$\lim_{x\to-\infty} \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = 0 \quad y \quad \lim_{x\to+\infty} \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = 1$$

Lo cual indica que la función es monótona creciente dado que

$$e^{\beta_0+\beta_1 x} < 1+e^{\beta_0+\beta_1 x}$$

En general los modelos logísticos pueden tener diferentes formas de representación, por ejemplo, siendo

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad \ldots(22)$$

Se obtiene la razón

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x} \quad \ldots(23)$$

A esta razón se le denomina $odss$, que establece la razón de probabilidad entre dos sucesos

$$odss = \frac{p}{1-p} = \frac{Probabilidad\ de\ que\ ocurra\ un\ suceso}{Probabilidad\ de\ que\ NO\ ocurra\ un\ suceso}$$

que cuantifica que tan probable es que ocurra un suceso, es decir su riesgo. También a la razón $odds$ se le denomina razón de ventaja a favor de éxito.

Al aplicar logaritmo natural a la ventaja o razón de $odds$ obtenemos

$$ln\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x \quad \ldots(24)$$

ecuación que expresa el modelo de forma lineal. A esta transformación del modelo logístico se le denomina $logit$, así estableceremos dos modelos logísticos

- El modelo $logit$:

$$\boldsymbol{logit}\big(p(x)\big) = ln\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x \quad \ldots(25)$$

- El modelo $probit$

  Para el modelo $probit$ se considera la función de distribución de una función normal estándar $N\ (0,1)$

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \ldots(26)$$

  Así, si tomamos $z = \beta_0 + \beta_1 x$ la función de distribución de probabilidad acumulada normal estándar estaría dada por

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}t^2}\, dt \quad \ldots(27)$$

$$P[Y = 1|X = x] = F(\beta_0 + \beta_1 x)$$

  la función $probit$ se acerca más rápidamente a probabilidades de 0 y 1 que la función $logit$ como se observa en la figura 8.

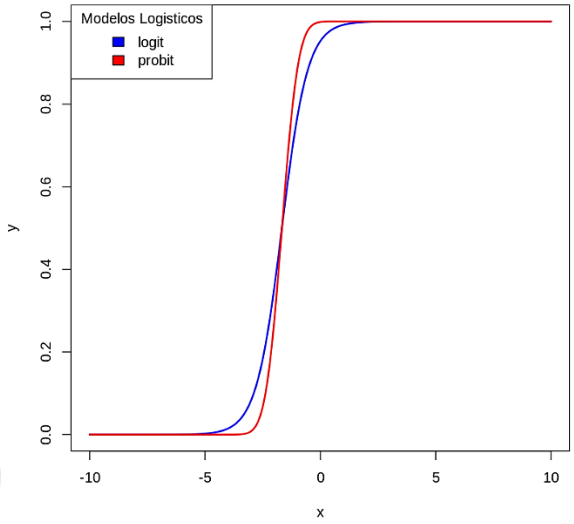

Figura 8: Modelos Logit y Probit para $\beta_0 = 3 \quad y \quad \beta_1 = 1.8$
Fuente: Elaboración propia

En general los modelos logísticos se pueden describir como

$$P[Y = 1|X = x] = F(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_R x_R) \qquad \ldots (28)$$

La regresión logística se puede clasificar en dos tipos:

- **La regresión logística binaria:** Cuando se pretende explicar una característica dicotómica, una variable dicotómica es aquella que solo puede tomar dos valores, por ejemplo, estar desempleado o no, abstenerse en las elecciones o no, tener una enfermedad o no tenerla, un paciente muere o no antes del alta, una persona deja o no de fumar después de un tratamiento, al ensayar la efectividad de un fármaco se estudia su dosis si es efectiva o no, etc. Estos valores, habitualmente se denominan como cero, en ausencia, o uno, en presencia de una característica o fenómeno.

- **La regresión logística multinomial:** Explica una variable cualitativa politómica que es un tipo de variable que puede tomar tres o más valores. Es decir, una variable politómica es toda aquella variable que tiene más de dos

opciones posibles, por ejemplo, el deporte favorito de una persona, elección de una marca de un producto, filiación política, nivel salarial, etc.

1.3.4.1. Modelo de regresión logística binaria simple

En algunos procesos naturales se realiza una progresión temporal desde un nivel mínimo hasta acercarse a un máximo durante un tiempo determinado, la función sigmoide permite describir esta evolución. Su gráfica tiene una forma de "S". Un caso de particular es la función logística que está definida por

$$f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad \ldots (29)$$

O equivalentemente

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad \ldots (30)$$

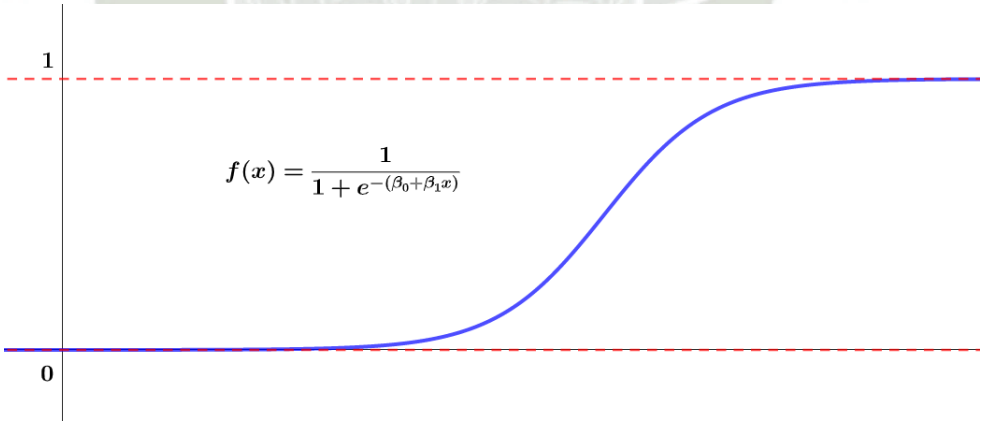

Figura 9: Función logística univariable
Fuente: Elaboración propia

La regresión logística binaria nos permite cuantificar la relación entre las variables independientes y la variable dependiente, clasificando los datos en dos categorías de la variable dependiente según su probabilidad de ocurrencia, tiene una variable de respuesta binaria y una variable explicativa. La variable de respuesta como Y la cual tomará el valor

de 1 si ocurre el suceso y 0 si no ocurre, por ello este modelo es denominado modelo de respuesta binaria o dicotómica.

Por ejemplo, la probabilidad de comprar una vivienda para un individuo de acuerdo a sus ingresos económicos o salario, no es la misma para los diferentes incrementos de salarios

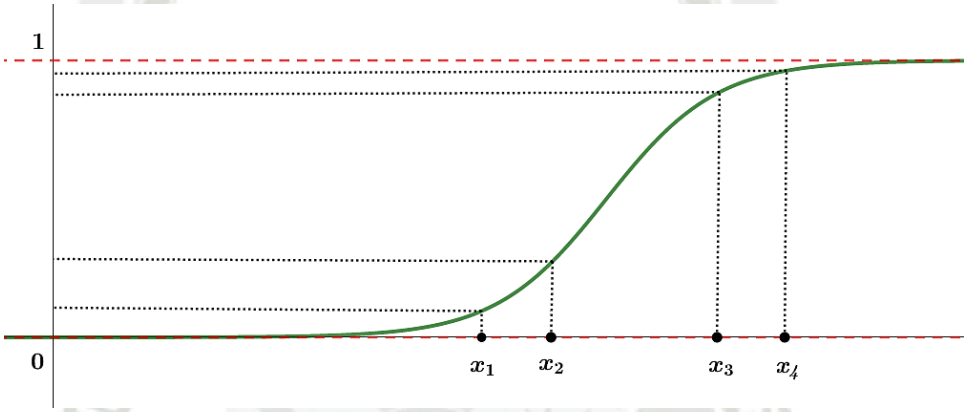

Figura 10: Función logística para variable Salarios
Fuente: Elaboración propia

Según la Figura 10, existe una mayor probabilidad de que Y valga 1 cuando los incrementos de los salarios de la variable X varían de x_1 a x_2, en comparación de variar de x_3 a x_4, por lo cual, algunos incrementos en los valores de X generarán una mayor probabilidad de que Y valga 1.

1.3.4.2. Modelo de regresión logística binaria múltiple

El modelo de regresión logística para una variable independiente puede generalizarse a utilizar dos o más variables independientes. Sean R valores $x_1, \dots, x_R$ de las variables $X_1, \dots, X_R$ definimos $P[Y = 1 | X_1 = x_1, \dots, X_R = x_R] = p(x_1, \dots, x_R)$ como la probabilidad de que Y tome el valor de 1 cuando las variables $X_1, \dots, X_R$ tomen los valores respectivos $x_1, \dots, x_R$, así

$$p(x_1, \dots, x_R) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_R x_R}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_R x_R}} \quad \dots (31)$$

O equivalentemente

$$p(x_1, \dots, x_R) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_R x_R)}} \quad \dots (32)$$

El modelo de regresión logística múltiple para Y en términos de los valores de las variables X, se describe en forma matricial por

$$p(\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}^t \boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}^t \boldsymbol{x}}} \quad \dots (33)$$

con $\boldsymbol{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_R \end{bmatrix}$ y $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_R \end{bmatrix}$

análogamente al modelo de regresión logística con una sola variable independiente, tenemos que la ecuación

$$\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} = e^{\boldsymbol{\beta}^t \boldsymbol{x}} \quad \dots (34)$$

denominaremos $odds$ a la razón

$$\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} = \frac{P[Y = 1 | X_1 = x_1, \dots, X_R = x_R]}{1 - P[Y = 1 | X_1 = x_1, \dots, X_R = x_R]} \quad \dots (35)$$

y al logaritmo natural de la relación

$$\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} = e^{\boldsymbol{\beta}^t \boldsymbol{x}}$$

se denominará modelo $logit$ de $p(\boldsymbol{x})$

$$\boldsymbol{logit}\big(p(\boldsymbol{x})\big) = ln\left[\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_R x_R \quad \dots (36)$$

$$\boldsymbol{logit}\big(p(\boldsymbol{x})\big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_R x_R$$

o equivalentemente

$$\boldsymbol{logit}\big(p(\boldsymbol{x})\big) = \sum_{r=0}^{R} \beta_r x_r \qquad \ldots (37)$$

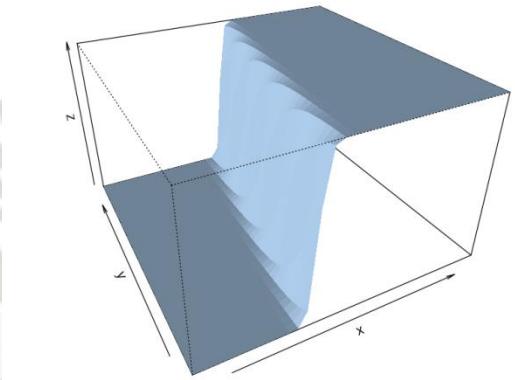

Figura 11: Grafica de la función $f(x_1, x_2) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$
Fuente: Elaboración propia

La regresión logística es una técnica estadística multivariable destinada al análisis de una relación de dependencia entre una variable dependiente y un conjunto de variables independientes, análoga a la regresión lineal simple. El objetivo de la regresión logística es poder efectuar predicciones del comportamiento. Así, a diferencia de la regresión lineal, la regresión logística explica o pronostica la pertenencia a un grupo, a partir de una variable dependiente categórica o cualitativa, en función de una o más variables independientes que pueden ser tanto cuantitativas como cualitativas.

1.3.4.3. Ajuste del modelo

1.3.4.3.1. Estimación de los coeficientes

Para realizar el ajuste del modelo se utiliza el método de máxima verosimilitud, el cual se encuentra implementado en algunos programas estadísticos.

Asumimos que se dispone de una muestra de n observaciones independientes (x_i, y_i), para $i = 1,2,\dots,n$ donde y_i toma valores 0 ó 1, para estimar los parámetros β_i desconocidos.

En el caso del modelo de regresión lineal múltiple se utiliza el método de mínimos cuadrados para estimar los parámetros β_i, donde se minimiza la suma de cuadrados del error, pero cuando las variables son binarias este método no cumple con todas las propiedades por ese motivo se tiene que usar el método de máxima verosimilitud, ya que se obtiene parámetros estimados que maximizan la probabilidad de obtener un conjunto de datos observados estimados de forma significativa.

La función de verosimilitud expresa la probabilidad de los datos observados como una función de parámetros desconocidos. Los estimadores de máxima verosimilitud de esos parámetros son aquellos que están en concordancia con los datos observados.

En muchos procedimientos estadísticos se desconocen algunos de sus parámetros, los cuales deben ser estimados a partir de cierta información obtenida en algún estudio. Existen diferentes formas y métodos de poder estimar dichos parámetros, pero el más usado es el método de máxima verosimilitud.

Hoy en día mediante los programas estadísticos, es posible calcular los parámetros para un modelo de regresión logístico. Sin embargo, puede suceder que en alguna situación se utilice los softwares estadísticos sin saber el proceso del cálculo de los coeficientes.

El método de máxima verosimilitud consiste en maximizar la función de verosimilitud. Consideremos que $X_1, X_{2,}, \cdots, X_n$ son variables aleatorias independientes tomadas de una distribución de probabilidad discreta o de una función de densidad de

probabilidad, la cual se representa por $f(x;\theta)$, donde θ es el parámetro de la distribución, así la función

$$L(x_1, x_2, \cdots, x_n; \theta) = f(x_1, x_2, \cdots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta) \quad \ldots (38)$$

es la distribución conjunta de las variables aleatorias, a la cual frecuentemente se le denomina función de probabilidad, en la cual θ es la variable de la función de probabilidad.

Al representar por $x_1, x_2, \cdots, x_n$ a los valores observados en una muestra, la cantidad $L(x_1, x_2, \cdots, x_n; \theta)$ es la verosimilitud de la muestra, suponiendo que los valores observados $x_1, x_2, \cdots, x_n$ son fijos mientras que θ puede variar libremente, para la función de verosimilitud

$$L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \quad \ldots (39)$$

y se suele utilizar el logaritmo de esta función:

$$\hat{l}(x_1, x_2, \cdots, x_n; \theta) = lnL = \sum_{i=1}^{n} ln\, f(x_i; \theta) \quad \ldots (40)$$

Para la cual el método de la máxima verosimilitud busca estimar un valor θ_0 que maximice $\hat{l}(x;\theta)$.

Asumiendo que para la muestra de n observaciones independientes existe una respuesta

$$(x_i, y_i),\ i = 1,2, \ldots, n\ ;\ \text{donde cada } y_i \text{ toma valores } y_i = 0 \text{ ó } y_i = 1$$

Sea $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$ donde $y_i \sim B(1, p_i)$

$$P[y_i = 1|X = x] = p(x) = \frac{e^{\beta^t x}}{1 + e^{\beta^t x}} \quad \ldots (41)$$

$$P[y_i = 0|X = \boldsymbol{x}] = 1 - p(\boldsymbol{x}) = 1 - \frac{e^{\boldsymbol{\beta}^t \boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}^t \boldsymbol{x}}} = \frac{1}{1 + e^{\boldsymbol{\beta}^t \boldsymbol{x}}}$$

Es decir

$$1 - p(\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{\beta}^t \boldsymbol{x}}} \quad \ldots (42)$$

El vector a estimar es $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_R \end{bmatrix}$, siendo $\boldsymbol{x_i} = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{iR} \end{bmatrix}$ tenemos

$$\lambda_i = ln\left[\frac{p(\boldsymbol{x_i})}{1 - p(\boldsymbol{x_i})}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_R x_{iR} = \sum_{r=0}^{R} \beta_r x_{ir}$$

$$\lambda_i = \sum_{r=0}^{R} \beta_r x_{ir} \quad \ldots (43)$$

$$f(y_i; p(\boldsymbol{x_i})) = \prod_{i=1}^{n} p(\boldsymbol{x_i})^{y_i} (1 - p(\boldsymbol{x_i}))^{1-y_i} \quad \ldots (44)$$

$$= \prod_{i=1}^{n} \left(\frac{p(\boldsymbol{x_i})}{1 - p(\boldsymbol{x_i})}\right)^{y_i} (1 - p(\boldsymbol{x_i}))$$

$$= \prod_{i=1}^{n} \left(\frac{p(\boldsymbol{x_i})}{1 - p(\boldsymbol{x_i})}\right)^{y_i} \prod_{i=1}^{n} (1 - p(\boldsymbol{x_i}))$$

$$= \prod_{i=1}^{n}(1 - p(\boldsymbol{x_i})) \prod_{i=1}^{n} exp\left(ln\left(\frac{p(\boldsymbol{x_i})}{1 - p(\boldsymbol{x_i})}\right)^{y_i}\right)$$

$$= \prod_{i=1}^{n}(1 - p(\boldsymbol{x_i})) \prod_{i=1}^{n} exp\left(y_i ln\left(\frac{p(\boldsymbol{x_i})}{1 - p(\boldsymbol{x_i})}\right)\right)$$

$$= \prod_{i=1}^{n}(1 - p(\boldsymbol{x_i})) \prod_{i=1}^{n} exp\left(y_i \sum_{r=0}^{R} \beta_r x_{ir}\right)$$

$$= \prod_{i=1}^{n}(1 - p(\boldsymbol{x_i})) exp\left(\sum_{r=0}^{R}\left(\sum_{i=1}^{n} y_i x_{ir}\right)\beta_r\right)$$

$$\rightarrow L = f(y_i; p(\boldsymbol{x_i})) = \prod_{i=1}^{n}(1 - p(\boldsymbol{x_i}))\, exp\left(\sum_{r=0}^{R}\left(\sum_{i=1}^{n} y_i x_{ir}\right)\beta_r\right) \quad \ldots (45)$$

Tomando el logaritmo de la función:

$$\hat{l}(x_1, x_2, \cdots, x_n; \boldsymbol{\beta}) = lnL = \sum_{i=1}^{n} ln\,(1 - p(\boldsymbol{x_i})) + \sum_{r=0}^{R}\left(\sum_{i=1}^{n} y_i x_{ir}\right)\beta_r \quad \ldots (46)$$

Sabemos que

$$1 - p(\boldsymbol{x}) = \frac{1}{1 + e^{\boldsymbol{\beta^t x}}}$$

$$l\,n\left(1 - p(\boldsymbol{x_i})\right) = -l\,n\left(1 + e^{\boldsymbol{\beta^t x_i}}\right) = -l\,n\left(1 + exp\left(\sum_{r=0}^{R} \beta_r x_{ir}\right)\right)$$

Luego

$$\hat{l}(x_1, x_2, \cdots, x_n; \boldsymbol{\beta}) = lnL = -\sum_{i=1}^{n} ln\left(1 + exp\left(\sum_{r=0}^{R} \beta_r x_{ir}\right)\right) + \sum_{r=0}^{R}\left(\sum_{i=1}^{n} y_i x_{ir}\right)\beta_r \quad \ldots (47)$$

Para hallar los estimadores hallaremos, las ecuaciones que verifiquen $\frac{\partial \hat{l}}{\partial \beta_j} = 0$

$$\frac{\partial \hat{l}}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ir} - \sum_{i=1}^{n} x_{ir} \left(\frac{exp(\sum_{r=0}^{R} \beta_r x_{ir})}{1 + exp(\sum_{r=0}^{R} \beta_r x_{ir})} \right) \quad \dots (48)$$

Siendo

$$\hat{p}(\boldsymbol{x_i}) = \frac{exp(\sum_{r=0}^{R} \beta_r x_{ir})}{1+exp(\sum_{r=0}^{R} \beta_r x_{ir})} \ , i = 1,2, \dots, n \quad \dots (49)$$

$$\sum_{i=1}^{n} y_i x_{ir} - \sum_{i=1}^{n} x_{ir}\, \hat{p}(\boldsymbol{x_i}) = 0 \ , \qquad r = 0,1,2, \dots, R, \text{con } x_{i0} = 1 \quad \dots (50)$$

De aquí

$$\sum_{i=1}^{n} x_{ir}(y_i - \hat{p}(\boldsymbol{x_i})) = 0 \ , \quad r = 0,1,2, \dots, R, \text{con } x_{i0} = 1 \quad \dots (51)$$

Equivalentemente en forma matricial

$$\boldsymbol{X}^t\big(\boldsymbol{Y} - \hat{p}(\boldsymbol{x})\big) = \boldsymbol{0} \quad \dots (52)$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1R} \\ 1 & x_{21} & x_{22} & \cdots & x_{1R} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nR} \end{bmatrix}$$

$$\frac{\partial^2 \hat{l}}{\partial \beta_r^2} = -\sum_{i=1}^{n} {x_{ir}}^2 \hat{p}(\boldsymbol{x_i})\big(1 - \hat{p}(\boldsymbol{x_i})\big) \ , \quad r = 0,1,2, \dots, R \quad \dots (53)$$

$$\frac{\partial^2 \hat{l}}{\partial \beta_r \partial \beta_k} = -\sum_{i=1}^{n} x_{ir} x_{ik} \hat{p}(\boldsymbol{x_i})\big(1 - \hat{p}(\boldsymbol{x_i})\big) \ , \quad r = 0,1,2, \dots, R \quad \dots (54)$$

1.3.4.3.2. Solución Numérica

Dada la función g y un escalar β, para encontrar un valor que maximice $g(\beta)$ para un valor b_k usando la serie de Taylor tenemos que

$$g(\beta) = g(b_k) + (\beta - b_k)\frac{dg(\beta)}{d\beta} + \frac{1}{2}(\beta - b_k)^2\frac{d^2g(\beta)}{d\beta^2} + R \quad \ldots(55)$$

Para hallar el estimador $g'(\beta) = 0$, así

$$\beta - b_k = \left[-\frac{d^2g(\beta)}{d\beta^2}\right]^{-1}\left[\frac{dg(\beta)}{d\beta}\right] \quad \ldots(56)$$

Por lo cual tomando $b_{k+1} \approx \boldsymbol{\beta}$ obtenemos

$$b_{k+1} = b_k + \left[-\frac{d^2g(\boldsymbol{\beta})}{d\boldsymbol{\beta}^2}\right]^{-1}\left[\frac{dg(\boldsymbol{\beta})}{d\boldsymbol{\beta}}\right] \quad \ldots(57)$$

Análogamente en forma vectorial

$$g(\boldsymbol{\beta}) = g(\boldsymbol{b_k}) + (\beta - \boldsymbol{b_k})^t\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \frac{1}{2}(\beta - \boldsymbol{b_k})^t\frac{\partial^2 g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(\partial \boldsymbol{\beta})^t}(\boldsymbol{\beta} - \boldsymbol{b_k}) + R \quad \ldots(58)$$

podemos obtener

$$\boldsymbol{b_{k+1}} = \boldsymbol{b_k} + \left[-\frac{\partial^2 g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}(\partial \boldsymbol{\beta})^t}\right]^{-1}\left[\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] \quad \ldots(59)$$

Usando la notación matricial tenemos que

$$\boldsymbol{b_{k+1}} = \boldsymbol{b_k} + (\boldsymbol{X}^t V_k \boldsymbol{X})^{-1}\boldsymbol{X}^t(\boldsymbol{Y} - \hat{p}^k(\boldsymbol{x_i})) \quad \ldots(60)$$

$$V_k = Diag\left(\hat{p}^k(\boldsymbol{x_i})\left(1 - \hat{p}^k(\boldsymbol{x_i})\right)\right)$$

Algoritmo

```
lreg1 = function(X, y, max.iter = 10, tol = 1e-06, verbose = FALSE){
X = cbind(1, X)
b.last = rep(0, ncol(X))
b = b.last
it = 1
while (it <= max.iter) {
if (verbose)
cat("\niteration = ", it, ",", b)
p= as.vector(1/(1 + exp(-X %*% b)))
V= diag(p * (1 - p))
var.b = solve(t(X) %*% V %*% X)
b = b + var.b %*% t(X) %*% (y - p)
if (max(abs(b - b.last)/(abs(b.last) + 0.001 * tol)) < tol)
break
b.last = b
it = it + 1
}
if (verbose)
cat("\n")
if (it > max.iter)
warning("número máximo de iteraciones excedido")
list(coefficients = as.vector(b), var = var.b, iterations = it)
}
```

1.3.4.3.3. Contraste de hipótesis para el coeficiente de la regresión logística simple

La validez del modelo ajustado depende del coeficiente β_1, este coeficiente debe ser estadísticamente distinto de cero con un nivel de significancia especificado que por lo general es del 5% ($p < 0.05$), por lo cual la hipótesis a contrastar es

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Considerando $\alpha = 0.05$, podemos utilizar la prueba $t - Student$ o la prueba de $Wald$

El coeficiente β_1 se distribuye normalmente, su estadístico de prueba está dado por

$$t_{n-2} = \frac{b_1 - \beta_{10}}{S_{b_1}}$$

Como el valor propuesto de beta es $\beta_{10} = 0$

$$t_{n-2} = \frac{b_1}{S_{b_1}}$$

Siendo S_{b_1} su error estándar muestral, se utiliza la distribución $t - Student$ con $n - 2$ grados de libertad, dado que el tamaño de la muestra es menor que 200, caso contrario se utilizará la distribución normal (Cáceres, 2007)

Para utilizar la prueba de $Wald$, consideramos el estadístico

$$Wald = \frac{{b_1}^2}{{S_{b_1}}^2}$$

Luego aplicamos el criterio del valor crítico, obtenido con la distribución $chi - cuadrado$ con un grado de libertad χ_1^2

1.3.4.3.4. Contraste de hipótesis para el coeficiente de la regresión logística múltiple

Análogamente a la regresión logística simple, luego de haber especificado y estimado el modelo de regresión se debe aplicar pruebas para validar el modelo, por lo cual la hipótesis

a contrastar consiste en determinar si las variables explicativas tienen coeficientes iguales a cero serán:

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

Considerando $\alpha = 0.05$, la prueba de $Wald$ considera el estadístico

$$W^2 = \frac{\hat{\beta}_j^{\,2}}{\sigma^2(\hat{\beta}_j)}$$

$$Wald = W^2 \sim \chi_1^2$$

1.3.4.3.5. Tabla de clasificación

La tabla de clasificación o tabla de confusión nos permite evaluar el ajuste del modelo de regresión logística, la tabla muestra la distribución de los objetos que pertenecen a las categorías, cuando $Y = 0$; $Y = 1$. En esta tabla se consignan las frecuencias de los diferentes tipos de aciertos y errores que se cometen cuando se aplica el modelo

Tabla 3: Tabla de Clasificación o de confusión

Valores observados de Y	**Valores pronosticados de Y**		Total (Observado)
	Positiva (+1)	Negativa (-1)	
Positiva (+1)	VP	FN	VP+FN
Negativa (-1)	FP	VN	FP+VN
Total (Pronosticado)	VP+FP	FN+VN	N

Fuente: Elaboración propia

Donde:

- VP: Verdadero positivo, numero de predicciones correctas
- VN: Verdadero Negativo, numero de predicciones correctas para valores negativos
- FP: Falso positivo, numero de predicciones incorrectas
- FN: Falso Negativo, numero de predicciones incorrectas para valores negativos
- $N = VP + VN + FP + FN$, numero de observaciones

La interpretación se realiza mediante el porcentaje de objetos bien clasificados

$$\frac{VP + VN}{N} \times 100\% \quad \ldots (61)$$

También es posible describir la tabla de clasificación como en la Tabla 3, donde en cada una de las casillas de la tabla aparece el número de casos que cumplen las condiciones indicadas por la fila y la columna correspondiente

Tabla 4: Tabla de clasificación alternativa a la tabla de confusión

Observaciones	**Predicciones del modelo**	
	P[Y=1]>0.5	P[Y=1]<0.5
Y=1		
Y=0		

Fuente: Retirada de (Levy Manquin, J. P., Varela Mallou, J., 2008)

1.3.4.3.6. Curva ROC

La curva ROC ("*Receiver Operating Characteristic*", Características operativas para el receptor) es una curva que representa la tasa verdadera positiva (sensitividad) en función de la tasa falsa positiva (1-especificidad). La curva ROC permite determinar un punto de corte de tal modo que los valores de la probabilidad mayores que él indican que $Y = 1$.

El modelo tiene mejor desempeño si la curva ROC correspondiente se aleja más de la diagonal principal. La Tabla 4 proporciona una interpretación del área bajo la curva ROC

Tabla 5: Clasificación del área ROC

	Poder discriminante
$Área\ ROC = 0.5$	Nulo
$0.7 \leq Área\ ROC < 0.8$	Aceptable
$0.8 \leq Área\ ROC < 0.9$	Excelente
$Área\ ROC \geq 0.9$	Excepcionalmente buena

Fuente: Elaboración propia

Ejemplo: Consideremos un conjunto de personas para el cual se determinará la probabilidad de comprar un producto de acuerdo a su salario dado en unidades monetarias, la tabla dada describe la información

Tabla 6: Tabla de salarios y condición de compra

X	Y
82	1
65	0
79	1
85	1
60	1
45	0
69	1
77	0
88	1
64	0
75	1
83	1

Fuente: Elaboración propia

Donde

- La variable independiente $X: Salario$
- La variable dependiente $Y = \begin{cases} 1\,, si\ el\ cliente\ compra\ el\ producto \\ 0\,,\ si\ el\ cliente\ no\ compra\ el\ producto \end{cases}$

El modelo logístico está dado por

$$f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Para estimar sus coeficientes $b_0 \approx \beta_0; b_1 \approx \beta_1$, podemos utilizar el programa estadístico SPSS, con lo cual obtenemos

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	X	,162	,086	3,527	1	,060	1,176
	Constante	-10,973	6,123	3,211	1	,073	,000

a. Variables especificadas en el paso 1: X.

Figura 12: Estimación de parámetros, para la compra de un producto usando SPSS
Fuente: Elaboración propia

Obteniendo $b_0 = -10.9730; b_1 \approx 0.1620$, luego el modelo logístico estimado está dado por

$$f(x) = \frac{1}{1 + e^{-(-10.9730+0.1620x)}}$$

Usando el programa R- Studio obtenemos resultados análogos, en efecto:

Si cargamos la base de datos a la variable M, al ejecutar lo comandos

```
M=EJEMPLO_LOG_02
X1=M$X
Y1=M$Y
glm.fit=glm(Y1~X1,data=M,family=binomial)
summary(glm.fit)
```

obtenemos

```
Call:
glm(formula = Y1 ~ X1, family = binomial, data = M)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8520  -0.7598   0.3702   0.5235   1.7290

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.9726     6.1234  -1.792   0.0731 .
X1            0.1622     0.0864   1.8780   0.0604 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 18.2490  on 13  degrees of freedom
Residual deviance: 11.9890  on 12  degrees of freedom
AIC: 15.9890
```

```
Number of Fisher Scoring iterations: 5
```

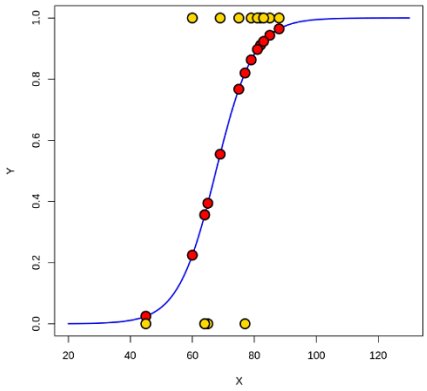

Figura 13: Modelo logístico, para la compra de un producto
Fuente: Elaboración propia

Para mostrar el ajuste, utilizamos el histograma de datos agrupados

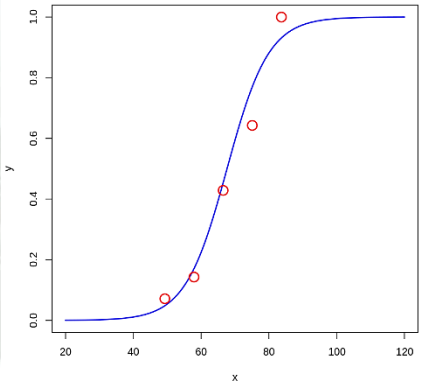

Figura 14: Modelo logístico para datos agrupados, para la compra de un producto
Fuente: Elaboración propia

Para analizar, el modelo logístico estimado tenemos:

- Si una persona tiene un ingreso de 67 unidades monetarias, la probabilidad de que compre el producto está dada por

$$P[Y = 1|x = 67\ ] = p(67) = \frac{1}{1 + e^{-(-10.9730+0.1620(67))}} = 0.4703$$

Es decir, existe una probabilidad de 47.03% de que una persona que tiene un ingreso de 67 unidades monetarias compre el producto.

La razón denominada $odss$, establece la razón de probabilidad entre dos sucesos

$$odss = \frac{p(x)}{1 - p(x)} = \frac{Probabilidad\ de\ que\ ocurra\ un\ suceso}{Probabilidad\ de\ que\ NO\ ocurra\ un\ suceso}$$

cuantifica que tan probable es que se compre el producto, es decir su riesgo

$$odss = \frac{p(67)}{1 - p(67)} = \frac{0.4703}{1 - 0.4703} = 0.8879 \approx \frac{9}{10}$$

- Cuando el salario es de $x = 67$ unidades monetarias, un éxito es 9 veces tan probable como diez fallas es decir las chances $(odds)$ de comprar el producto son de 9 a 10

- Si una persona de tiene un ingreso de 86 unidades monetarias, la probabilidad de que compre el producto está dada por

$$P[Y = 1|x = 86\ ] = p(86) = \frac{1}{1 + e^{-(-10.9730 + 0.1620(67))}} = 0.9507$$

Es decir, existe una probabilidad de 95.07% de que una persona que tiene un ingreso de 86 unidades monetarias compre el producto.

La razón denominada $odss$, establece la razón de probabilidad entre dos sucesos

$$odss = \frac{p(x)}{1 - p(x)} = \frac{Probabilidad\ de\ que\ ocurra\ un\ suceso}{Probabilidad\ de\ que\ NO\ ocurra\ un\ suceso}$$

Esta razón cuantifica que tan probable es que se compre el producto, es decir su riesgo

$$odss = \frac{p(86)}{1 - p(86)} = \frac{0.9507}{1 - 0.9507} = 19.28 \approx 19$$

- Cuando el salario es de $x = 86$ unidades monetarias, un éxito es 19 veces tan probable como una falla es decir las chances ($odds$) de comprar el producto son de 19 a 1

Contraste de hipótesis para el coeficiente de la regresión logística simple

La validez del modelo ajustado depende del coeficiente β_1, este coeficiente debe ser estadísticamente distinto de cero con un nivel de significancia especificado que por lo general es del 5%, por lo cual la hipótesis a contrastar es

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Considerando $\alpha = 0.05$, podemos utilizar la prueba $t - Student$ o la prueba de $Wald$

El coeficiente β_1 se distribuye normalmente, su estadístico de prueba está dado por

$$t_{n-2} = \frac{b_1 - \beta_{10}}{S_{b_1}}$$

Como el valor propuesto de beta es $\beta_{10} = 0$

$$t_{n-2} = \frac{b_1}{S_{b_1}}$$

Siendo S_{b_1} su error estándar muestral, se utiliza la distribución $t - Student$ con $n - 2$ grados de libertad, dado que el tamaño de la muestra es menor que 200, caso contrario se utilizará la distribución normal (Cáceres, 2007)

Mediante el uso del programa SPSS, obtenemos que:

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	X	,162	,086	3,527	1	,060	1,176
	Constante	-10,973	6,123	3,211	1	,073	,000

a. Variables especificadas en el paso 1: X.

Figura 15: Determinación de parámetros para la prueba de Wald y la Prueba t-Student, para la compra de un producto
Fuente: Elaboración propia

Dado que $n = 14, T \sim t(12)$

$$t_{12} = \frac{b_1}{S_{b_1}} = \frac{0.1622}{0.0860} = 1.8777$$

Su $valor\ P = 0.0849 \not< 0.05 = \alpha$ por lo cual el coeficiente de regresión logística no es significativo, es decir $\beta_1 = 0$

Utilizando la distribución normal estándar para el estadístico de prueba

$$z = \frac{b_1}{S_{b_1}} = \frac{0.1622}{0.0860} = 1.8777$$

Su $valor\ P = 0.0604 \not< 0.05 = \alpha$ por lo cual el coeficiente de regresión logística no es significativo, es decir $\beta_1 = 0$

Obtenemos un resultado análogo usando la prueba de $Wald$

$$Wald = \frac{{b_1}^2}{{S_{b_1}}^2} = \frac{0.1622^2}{0.0864^2} = 3.5268$$

Hallamos el valor crítico de la distribución $chi - cuadrado$ con un grado de libertad χ_1^2

$$x_{1-\alpha}^2 = x_{0.95}^2 = 3.8415$$

Siendo el cociente de $Wald$ menor que el valor critico no rechazamos la hipótesis nula es decir aceptamos que $\beta_1 = 0$.

En conclusión, el coeficiente del modelo no es significativo, el modelo de regresión logístico no se adecua a toda la población con la variable independiente X (Salario).

Tabla de clasificación

Usando el programa SPSS obtenemos

Tabla de clasificación[a]

Observado		Pronosticado: Y = 0	Pronosticado: Y = 1	Porcentaje correcto
Paso 1 Y	0	4	1	80,0
	1	1	8	88,9
Porcentaje global				85,7

a. El valor de corte es ,500

Tabla de clasificación[a]

Observado		Pronosticado: No comprar	Pronosticado: Comprar	Porcentaje correcto
Paso 1 Y	No comprar	4	1	80,0
	Comprar	1	8	88,9
Porcentaje global				85,7

a. El valor de corte es ,500

Figura 16: Tabla de clasificación, para la compra de un producto usando el programa SPSS
Fuente: Elaboración propia

Tabla 7: Tabla de clasificación, para la compra de un producto

Valores observados de Y	Valores pronosticados de Y		Total(Observado)
	Positiva (+1)	Negativa (-1)	
Positiva (+1)	8	1	9
Negativa (-1)	1	4	5
Total (Pronosticado)	9	5	14

Fuente: Elaboración propia

Donde:

- $VP = 8$
- $VN = 4$
- $N = 14$

El porcentaje de objetos bien clasificados

$$\frac{VP + VN}{N} \times 100\% = \frac{8 + 4}{14} \times 100\% = 85.71\%$$

El 85.71% de los casos están correctamente clasificados.

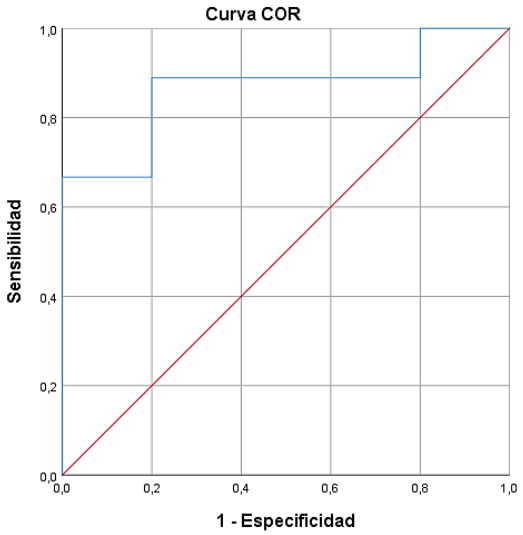

Figura 17: Gráfica de la curva ROC, para la compra de un producto
Fuente: Elaboración propia

Área bajo la curva

Variables de resultado de prueba: X

Área	Desv. Error[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,867	,101	,028	,669	1,000

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

Figura 18: Área de la curva ROC, usando SPSS
Fuente: Elaboración propia

Ejemplo: Tomando una muestra de 40 hombres respecto a una enfermedad coronaria (EC), se observa su evolución durante cinco años y se observa quienes desarrollaron la enfermedad coronaria codificando con un quienes desarrollaron la enfermedad coronaria y con un 0 quienes no desarrollaron la enfermedad coronaria, la variable FUMA se codifica con un cero para los no fumadores y con un uno para los fumadores, la variable EDAD en años es la edad que tenían los integrantes de la muestra al ser incluidos en el estudio (Cáceres, 2007), la Tabla 6 dada, describe la información

Donde

- La variable independiente X_1: $FUMA$
- La variable independiente X_2: $EDAD$
- La variable dependiente $Y = \begin{cases} 1\,, Desarrollo\ la\ enfermedad\ coronaria \\ 0\,, No\ desarrollo\ la\ enfermedad\ coronaria \end{cases}$

El modelo logístico está dado por

$$p(x_1, x_2) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

Para estimar sus coeficientes podemos utilizar el programa estadístico SPSS, con lo cual obtenemos

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	X1	2,890	1,103	6,869	1	,009	17,997
	X2	,147	,061	5,723	1	,017	1,158
	Constante	-11,359	3,673	9,566	1	,002	,000

a. Variables especificadas en el paso 1: X1, X2.

Figura 19: Parámetros del modelo logístico para la enfermedad coronaria, usando SPSS
Fuente: Elaboración propia

La función logística estaría dada por

$$p(x_1, x_2) = \frac{1}{1 + e^{-(-11.3590 + 2.890 x_1 + 0.1470 x_2)}}$$

Tabla 8: Enfermedad Coronaria

Y	X1	X2
1	1	42
0	1	43
0	1	43
0	1	45
0	1	47
0	1	47
0	1	49
0	1	49
0	1	51
1	1	53
0	1	53
0	1	54
1	1	54
0	1	54
0	1	55
0	1	56
1	1	56
1	1	57
1	1	57
0	1	57
0	1	58
0	1	58
0	1	58
1	1	58
1	1	59
1	1	60
1	1	62
1	1	62
1	1	63
1	1	64
1	1	67
0	1	67
1	1	67
0	1	68
0	0	45
0	0	46
0	0	46
0	0	47
0	0	48
0	0	48
0	0	48
0	0	49
0	0	49
0	0	49
0	0	50
0	0	52
0	0	52
0	0	53
0	0	53
0	0	53
0	0	57
0	0	58
0	0	58
0	0	59
0	0	59
0	0	60
0	0	61
0	0	62
0	0	62
1	0	63

Fuente: (Cáceres, 2007)

Usando el programa R- Studio obtenemos resultados análogos, en efecto:

Si cargamos la base de datos a la variable M, al ejecutar los comandos

```
M=EJEMPLO_LOG_01
X1=M$X1
X2=M$X2
Y1=M$Y
glm.fit=glm(Y1~X1+X2,data=M,family=binomial)
summary(glm.fit)
```

obtenemos

```
Call: glm(formula = Y1 ~ X1 + X2, family = binomial, data = M)
Deviance Residuals:

          Min      1Q      Median      3Q      Max
       -1.8462  -0.6130  -0.2337  0.6769  2.1928

Coefficients:
                  Estimate    Std. Error   z value    Pr(>|z|)
(Intercept)      -11.3594     3.6728       -3.093     0.0019 **
X1                 2.8902     1.1028        2.621     0.0087 **
X2                 0.1467     0.0613        2.392     0.0167 *

--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 69.590 on 59 degrees of freedom
Residual deviance: 47.711 on 57 degrees of freedom
AIC: 53.711 Number of Fisher Scoring iterations: 6
```

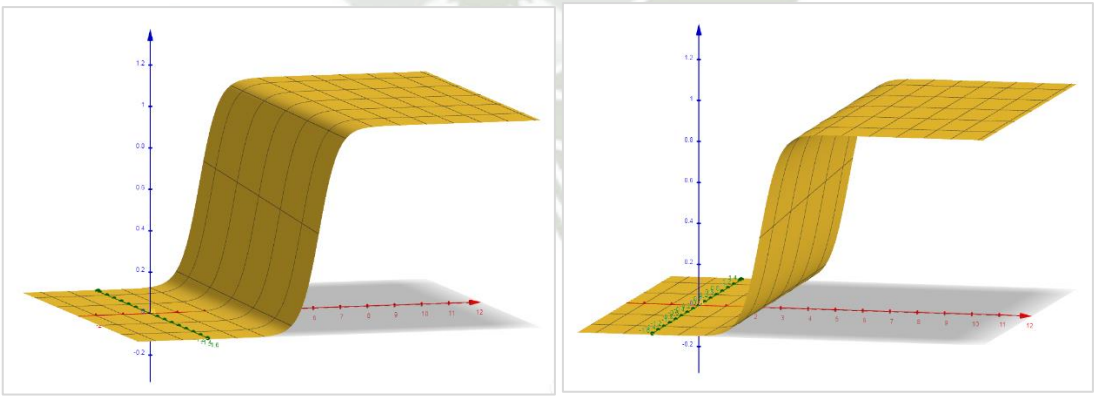

Figura 20: Grafica del Modelo logístico, enfermedad coronaria, usando Geogebra
Fuente: Elaboración propia

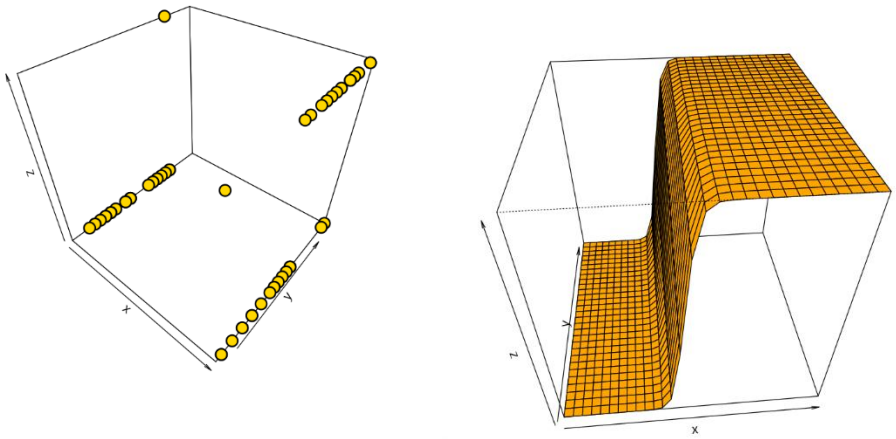

Figura 21: Modelo logístico, para el ejemplo de enfermedad coronaria usando el lenguaje R
Fuente: Elaboración propia

Para un hombre de 60 años que es fumador, la probabilidad que desarrolle enfermedad coronaria antes de cinco años está dada por

$$P[Y = 1|x_1 = 1, x_2 = 60\ ] = p(1{,}60) = \frac{1}{1 + e^{-(-11.3590+2.8900(1)+0.1470(60))}} = 0.5869$$

Es decir, existe una probabilidad de 58.69% de que un varón de 60 años que fuma desarrolle enfermedad coronaria antes de cinco años.

Si un hombre de 60 años que no es fumador, la probabilidad que desarrolle enfermedad coronaria antes de cinco años está dada por

$$P[Y = 1|x_1 = 0, x_2 = 60\ ] = p(0{,}60) = \frac{1}{1 + e^{-(-11.3590+2.8900(0)+0.1470(60))}} = 0.0732$$

Es decir, existe una probabilidad de 7.32% de que un varón de 60 años que no es fumador desarrolle enfermedad coronaria antes de cinco años.

1.3.5. Análisis de conglomerados

El análisis de conglomerados tiene por objetivo la clasificación o agrupamiento de individuos u objetos en clase o conglomerados a partir de mediciones realizadas en ellos de tal manera que dentro de los grupos se reúnan los elementos más homogéneos y que entre los grupos exista la mayor heterogeneidad (Véliz Capuñay, 2017). Las técnicas de formar conglomerados son diversas, se pueden dividir como en: Técnicas jerárquicas aglomerativas y Técnicas no jerárquicas, estas técnicas utilizan el concepto de distancia para la matriz de datos.

La matriz de datos

Está conformada por los diferentes valores de las variables de manera que cada individuo se identifica por cada fila y los valores de cada variable para cada individuo esta dado en cada columna

El concepto de distancia

Existen distintas maneras de considerar la distancia que separa a dos objetos, estas en general se denominan métricas se denotan con la letra d y se define como una aplicación en base a un conjunto E, denominado espacio métrico de la siguiente manera:

$$d: E \times E \to [0, \infty >$$

$$(a, b) \mapsto r$$

Donde $r = d(a, b)$ verifica las condiciones

a) $d(a, b) \geq 0 \quad y\, d(a, a) = 0$

b) $d(a, b) = d(b, a)$

c) $d(a,b) \leq d(a,c) + d(c,b)$

Algunos ejemplos de distancias son:

Distancia Euclídea:

$$d(i,j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$

Distancia Euclídea al cuadrado:

$$d^2(i,j) = \sum_k (x_{ik} - x_{jk})^2$$

Distancia de Minkowsky:

$$d_q(i,j) = \left(\sum_k (x_{ik} - x_{jk})^q \right)^{1/q}$$

Distancia de City-Block o de Manhattan:

$$d_1(i,j) = \sum_k |x_{ik} - x_{jk}|$$

Distancia de Tchebysheff:

$$d_\infty(i,j) = \max_k \{|x_{ik} - x_{jk}|\}$$

Distancia de Camberra:

$$d_{canb}(i,j) = \sum_k \frac{|x_{ik} - x_{jk}|}{(x_{ik} + x_{jk})}$$

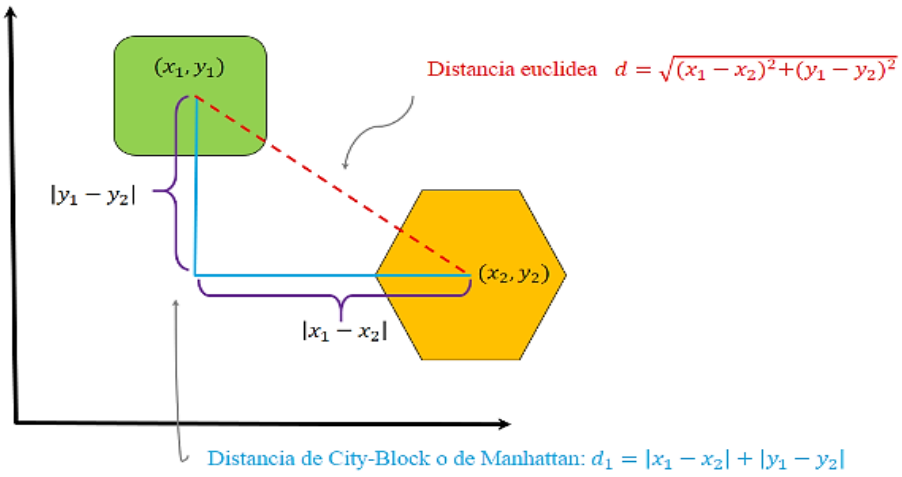

Figura 22: Interpretación geométrica de las distancias
Fuente: Elaboración propia

1.3.5.1. Técnicas jerárquicas aglomerativas:

Se forman grupos sucesivos partiendo de tantos grupos como elementos se tengan hasta formar un único grupo con todos los elementos

Pasos

i. Se parte de tantos conglomerados como elementos existan
ii. Se calculan las distancias entre los conglomerados iniciales
iii. Con los dos conglomerados más próximos se forma un nuevo grupo
iv. Con los nuevos elementos se procede como en los pasos 2 y 3 hasta obtener un solo grupo formado con todos los elementos

Los pasos que se siguen para llevar a cabo la partición de los elementos se representan mediante un diagrama denominado dendograma.

Dendograma

Es una representación gráfica de los datos en forma de árbol que organiza los datos en subcategorías dividiéndose en niveles. Para formar este diagrama se forman conglomerados de observaciones en cada paso y sus niveles de similitud. El nivel de

similitud se mide en el eje vertical (o el nivel de distancia), y las diferentes observaciones se especifican en el eje horizontal.

Ejemplo: *Técnica Jerárquica* (Véliz Capuñay, 2017).

Se consideran 6 clientes de una entidad financiera para los cuales se han calculado los valores de las variables:

X_1: "Edad"
X_2: "Sueldo mensual"

Los valores de las variables aparecen en la siguiente tabla:

Tabla 9: Tabla de edades y sueldos

	Edad	Sueldo
1	28	2800
2	35	3500
3	33	4700
4	50	5500
5	48	4500
6	25	7000

Fuente: Veliz Capuñay, 2017

La matriz de distancias euclidianas entre los clientes es la siguiente:

Tabla 10: Tabla de distancias euclidianas entre clientes

	1	2	**3**	4	5	6
1	0					
2	700.0350	0				
3	1900.0066	1200.0017	0			
4	2700.0896	2000.0562	800.1806	0		
5	1700.1176	1000.0845	**200.5617**	1000.0020	0	
6	4200.0011	3500.0143	2300.0139	1500.2083	2500.1058	0

Fuente: Elaboración propia

En la matriz observamos que los elementos más cercanos son el elemento 3 y el 5

$$\|(33{,}4700) - (48{,}4500)\| = 200.5617$$

Tabla 11: Tabla de edades y sueldos con señalización de distancias

	Edad	Sueldo
1	28	2800
2	35	3500
3	**33**	**4700**
4	50	5500
5	**48**	**4500**
6	25	7000

Fuente: Elaboración propia

$$d(1,[3,5]) = min(d(1,3), d(1,5)) = 1700.1180$$
$$d(2,[3,5]) = min(d(2,3), d(2,5)) = 1000.0840$$
$$d(4,[3,5]) = min(d(4,3), d(4,5)) = 800.1810$$
$$d(6,[3,5]) = min(d(6,3), d(6,5)) = 2300.0140$$

Usando el lenguaje de programación R, asignamos la matriz de datos “SUELDO” a la variable `M`

Tabla 12: Tabla de clientes, edades y sueldos

CLIENTES	Edad	Sueldo
C1	28	2800
C2	35	3500
C3	33	4700
C4	50	5500
C5	48	4500
C6	25	7000

Fuente: Elaboración propia

```
M=read xlsx("SUELDO.xlsx")
M1=M[,2:3]
distancias=dist(as.matrix(M1))
cluster=hclust(distancias)
cluster
```

```
Call: hclust(d = distancias)
Cluster method : complete
Distance : euclidean Number of objects: 6
```

```
plot(cluster,main = "dendrograma",labels = M$CLIENTES,col="red")
```

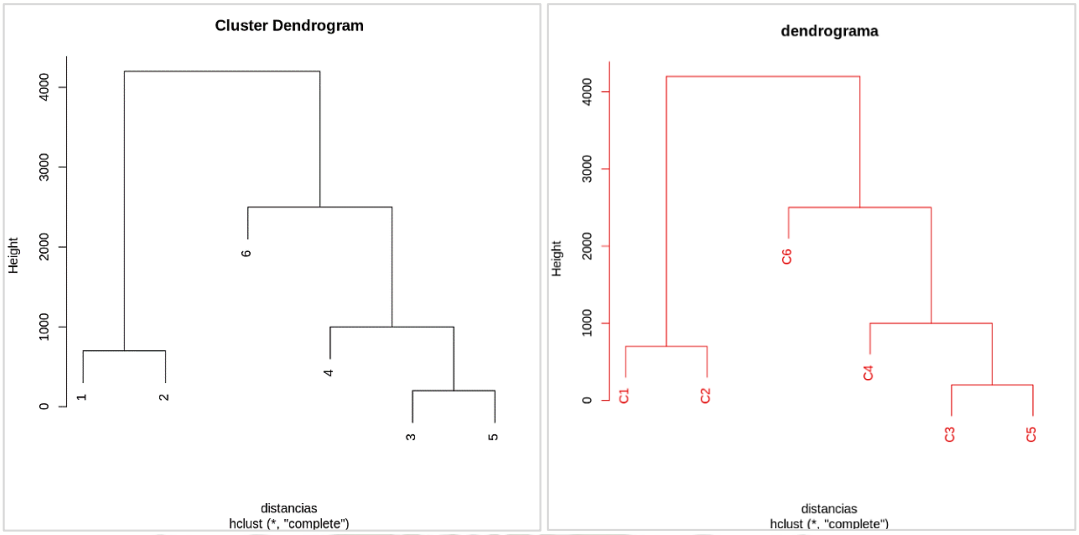

Figura 23: Dendograma para clientes
Fuente: Elaboración propia

Las líneas verticales indican la distancia en que los elementos se unen, si el dendograma se corta a cierto nivel de distancia se obtienen diferentes grupos

Grupo 1: Cliente 1 y Cliente 2

Grupo 2: Cliente 6 y en otras subcategorías Cliente 4, Cliente 3 y Cliente 5

Tabla 13: Medias de edades y sueldo en cada conglomerado de clientes

	Grupo1	Grupo2
Edad	31.5	39
Sueldo	3150	5425

Fuente: Elaboración propia

Si dividimos dentro del dendograma mediante una línea horizontal determinaremos diferentes subcategorías de acuerdo a una distancia especifica.

La decisión de elegir el número óptimo de clústeres es subjetiva, dado que, si se seleccionan pocos clústeres, puede generarse conglomerados heterogéneos.

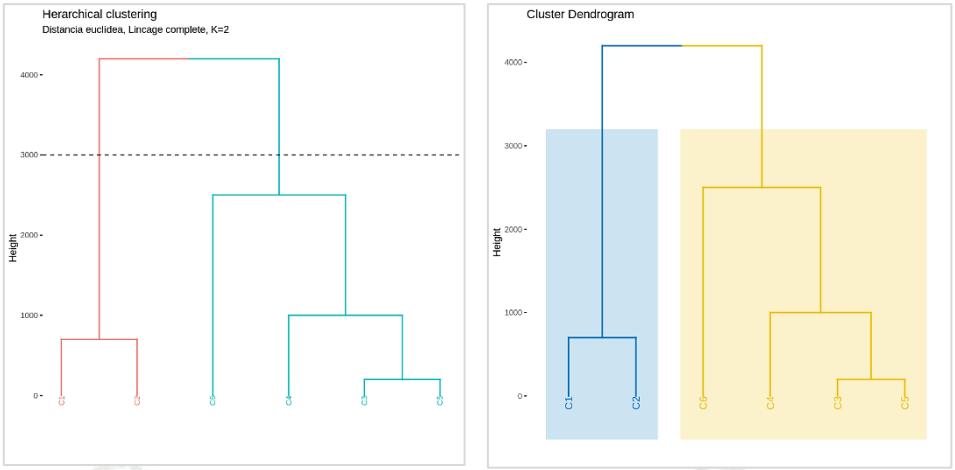

Figura 24: Dendograma de dos grupos para clientes
Fuente: Elaboración propia

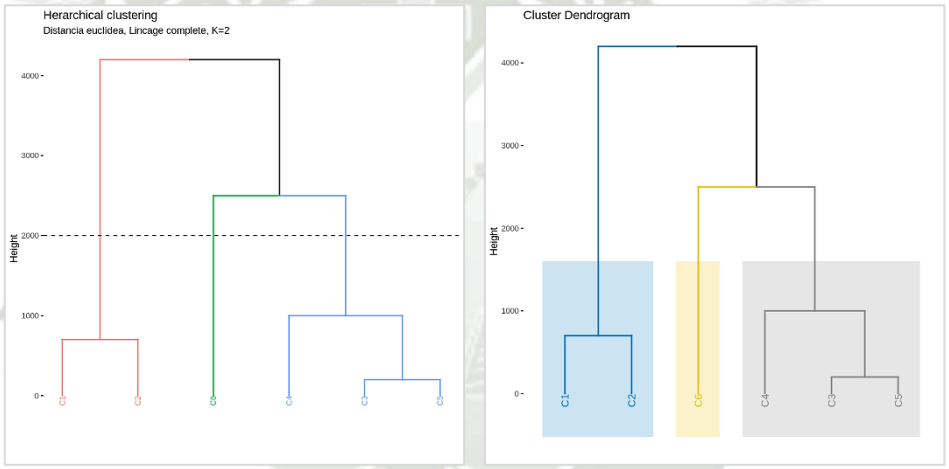

Figura 25: Dendograma de tres grupos para clientes
Fuente: Elaboración propia

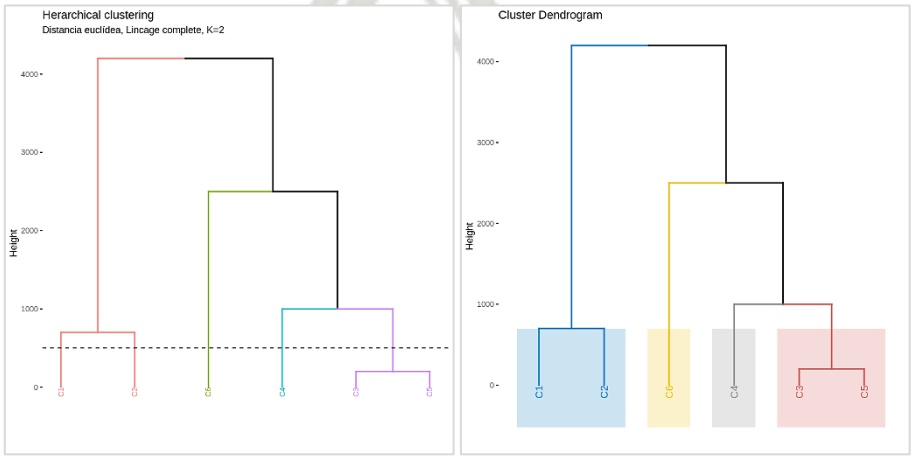

Figura 26: Dendograma de cuatro grupos para clientes
Fuente: Elaboración propia

Por otro lado, si se seleccionan demasiados clústeres, la interpretación de los mismos seria complicada por la carente cantidad de elementos.

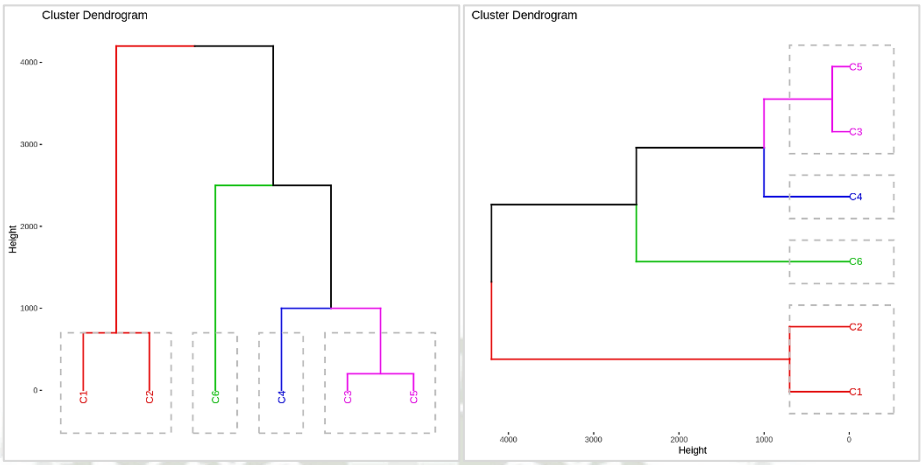

Figura 27: Clúster de cuatro grupos
Fuente: Elaboración propia

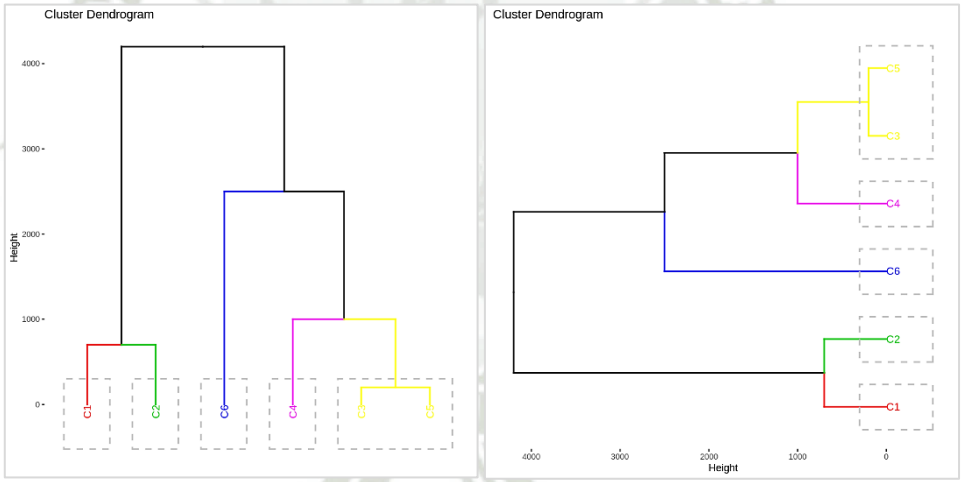

Figura 28: Clúster de cinco grupos
Fuente: Elaboración propia

Ejemplo: Se considera la cantidad diaria de Calorías(kcal), grasa(g) y proteínas(g) que consumen en 19 países de América Latina. Los datos corresponden al reporte de la organización de las naciones unidas para la alimentación entre los años 1990 y 1992. Usando la distancia euclidiana y el método del vecino más alejado se puede determinar el agrupamiento jerárquico de los países

Tabla 14: Consumo de calorías, grasas y proteínas

PAIS	CALORIAS	GRASAS	PROTEINAS
Argentina	2948	103	97
Bolivia	2031	51	52
Brasil	2791	82	64
Chile	2535	65	70
Colombia	2632	62	60
Costa Rica	2870	78	69
Cuba	3003	77	66
Ecuador	2539	90	52
El Salvador	2526	58	68
Guatemala	2282	42	58
Honduras	2307	61	56
Mexico	3190	84	80
Nicaragua	2290	52	55
Panama	2238	65	59
Paraguay	2618	68	91
Peru	1881	34	50
R. Dominicana	2273	65	50
Uruguay	2684	96	83
Venezuela	2586	95	65

Fuente: Elaboración propia

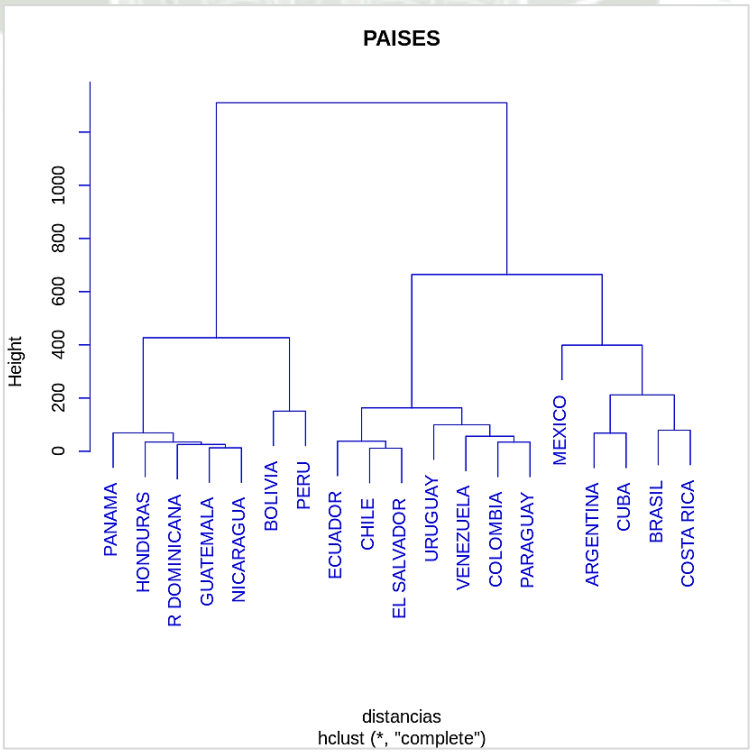

Figura 29: Dendograma para la alimentación por países
Fuente: Elaboración propia

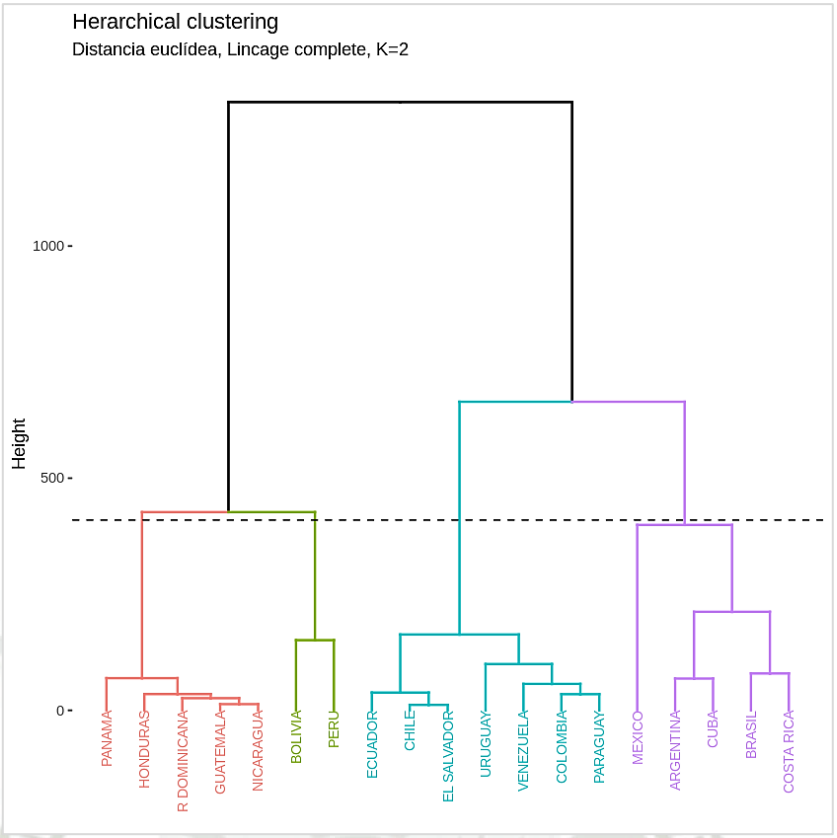

Figura 30: Distancia para Dendograma de cuatro grupos para la alimentación por países
Fuente: Elaboración propia

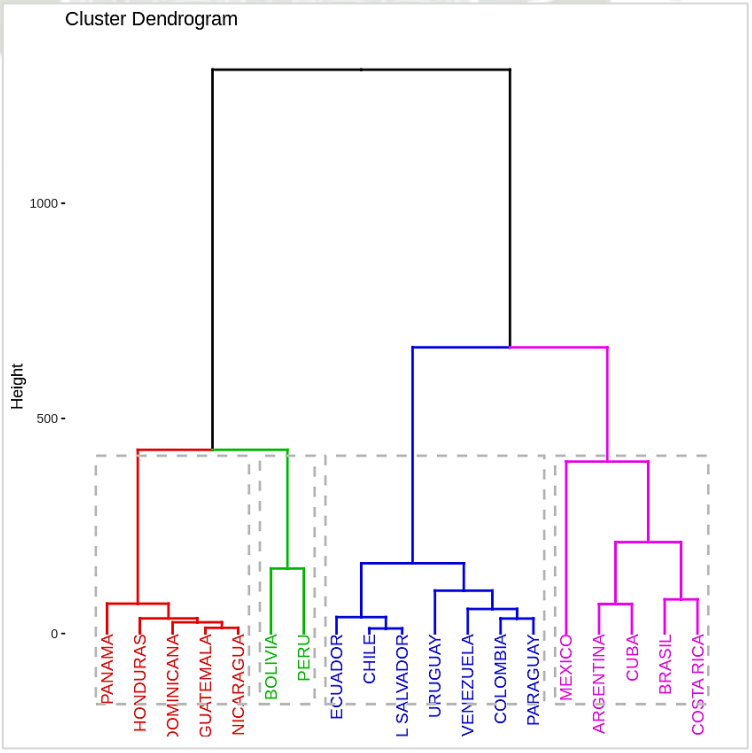

Figura 31: Dendograma de cuatro grupos para la alimentación por países
Fuente: Elaboración propia

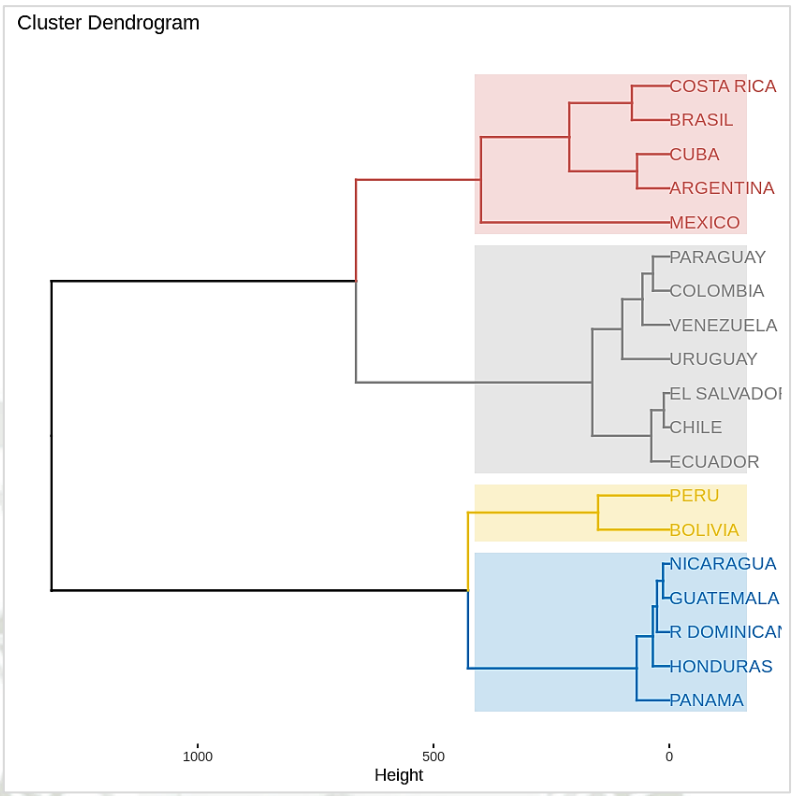

Figura 32: Dendograma horizontal de cuatro grupos para la alimentación por países
Fuente: Elaboración propia

Los grupos formados son:

- **Grupo 1:** Panamá, Honduras, República Dominicana, Guatemala y Nicaragua
- **Grupo 2:** Bolivia y Perú
- **Grupo 3:** Ecuador, Chile y El Salvador, Uruguay, Venezuela, Colombia y Paraguay
- **Grupo 4:** México, Argentina, Cuba, Brasil y Costa rica

Tabla 15: Consumo de calorías, grasas y proteínas por grupos

	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Calorías	2278	1956	2588.6	2960.4
Grasas	57	42.5	76.3	84.8
Proteínas	55.6	51	69.9	75.2

1.3.5.2. Técnicas no jerárquicas

Las técnicas no jerárquicas permiten la formación de conglomerados, se forma un número preestablecido de grupos. Entre los métodos no jerárquicos destaca el método de las $k-medias$ que permite la formación de un número k de grupos previamente determinado.

Si se tiene una muestra de N elementos para los cuales están definidas p variables numéricas, los pasos de ejecución del algoritmo son los siguientes:

i. **Elección del número de Clúster:** Se elige $k-grupos$ para formar los grupos

ii. **Inicializar las coordenadas de los centroides:** Para cada grupo se inicializan de forma aleatoria las coordenadas de los centroides.

iii. **Asignar cada punto a un clúster:** Usando la distancia euclidiana para cada elemento se calcula su distancia a cada uno de los centroides, reasignándolos al grupo cuyo centroide está más cercano.

iv. **Se recalculan los centroides de los clústeres:** Los nuevos centroides de los grupos formados se recalculan cuyas coordenadas son las medias aritméticas de las variables.

v. **Se repiten los pasos hasta llegar al criterio de parada:** Si la distancia entre los centroides iniciales y los nuevos centroides es pequeña o si se ha completado un número fijo de iteraciones o cuando los centroides dejan de cambiar así el proceso termina. De otro modo se repite el paso (iii) y (iv)

Ejemplo: Se tiene el registro de notas siete estudiantes, se desea agrupar a los estudiantes en dos grupos mediante el algoritmo de $k-means$ y la distancia media euclidiana

Tabla 16: Notas de estudiantes en dos asignaturas

	Asignatura 1	Asignatura 2
E1	11	11
E2	11.5	12
E3	13	14
E4	15.5	17
E5	13.5	15
E6	14.5	15
E7	13.5	14.5

Fuente: Elaboración propia

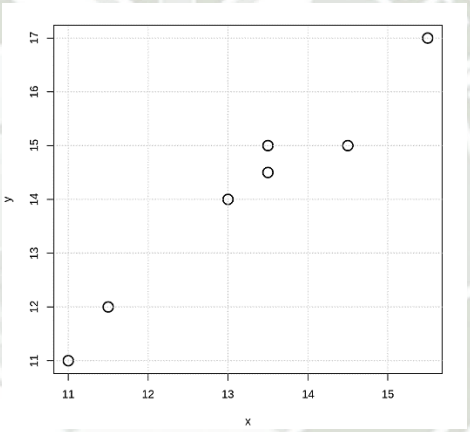

Figura 33: Grafica de dispersión del registro de notas
Fuente: Elaboración propia

Centroides iniciales

$$C_1 = (11.5,12.5)\ \ C_2 = (14,16)$$

Distancias al centroide C_1

$$d(E_1, C_1) = \sqrt{(11-11.5)^2 + (11-12.5)^2} = 1.58114$$

$$d(E_2, C_1) = \sqrt{(11.5 - 11.5)^2 + (12 - 12.5)^2} = 0.5000$$

$$\vdots$$

$$d(E_7, C_1) = \sqrt{(13.5 - 11.5)^2 + (14.5 - 12.5)^2} = 2.8284$$

Distancias al centroide C_2

$$d(E_1, C_2) = \sqrt{(11 - 14)^2 + (11 - 16)^2} = 5.8310$$

$$d(E_2, C_2) = \sqrt{(11.5 - 14)^2 + (12 - 16)^2} = 4.7170$$

$$\vdots$$

$$d(E_7, C_2) = \sqrt{(13.5 - 14)^2 + (14.5 - 16)^2} = 1.5811$$

Tabla 17: Tabla de distancias de datos a centroides y asignación de Clúster

		Distancia al centroide C1	Distancia al centroide C2	Asignación de clúster
E1	(11,11)	1.58114	5.8309	1
E2	(11.5,12)	0.5000	4.7169	1
E3	(13,14)	2.1213	2.2360	1
E4	(15.5,17)	6.0208	1.8028	2
E5	(13.5,15)	3.2016	1.1180	2
E6	(14.5,15)	3.9051	1.1180	2
E7	(13.5,14.5)	2.8284	1.5811	2

Fuente: Elaboración propia

Cálculo de nuevos Centroides

$$C_1 = \frac{E_1 + E_2 + E_3}{3};\ C_2 = \frac{E_4 + E_5 + E_6 + E_7}{4}$$

$$C_1 = (11.8333, 12.3333)\ \ C_2 = (14.2500, 15.3750)$$

Distancias al centroide C_1

$$d(E_1, C_1) = \sqrt{(11 - 11.8333)^2 + (11 - 12.3333)^2} = 1.5723$$

$$d(E_2, C_1) = \sqrt{(11.5 - 11.8333)^2 + (12 - 12.3333)^2} = 0.4714$$

$$\vdots$$

$$d(E_7, C_1) = \sqrt{(13.5 - 11.8333)^2 + (14.5 - 12.3333)^2} = 2.7335$$

Distancias al centroide C_2

$$d(E_1, C_2) = \sqrt{(11 - 14.2500)^2 + (11 - 15.3750)^2} = 5.4501$$

$$d(E_2, C_2) = \sqrt{(11.5 - 14.2500)^2 + (12 - 15.3750)^2} = 4.3535$$

$$\vdots$$

$$d(E_7, C_2) = \sqrt{(13.5 - 14.2500)^2 + (14.5 - 15.3750)^2} = 1.1524$$

Tabla 18: Tabla de distancias a nuevos centroides y asignación de Clúster

		Distancia al centroide C1	Distancia al centroide C2	Asignacion de cluster
E1	(11,11)	1.572330	5.450057	1
E2	(11.5,12)	0.471405	4.353519	1
E3	(13,14)	2.034426	1.858259	2
E4	(15.5,17)	5.934831	2.050152	2
E5	(13.5,15)	3.144660	0.838526	2
E6	(14.5,15)	3.771236	0.450694	2
E7	(13.5,14.5)	2.733537	1.152443	2

Fuente: Elaboración propia

Cálculo de nuevos Centroides

$$C_1 = \frac{E_1 + E_2}{3};\ C_2 = \frac{E_3 + E_4 + E_5 + E_6 + E_7}{4}$$

$$C_1 = (11.25,11.5)\ ;\ C_2 = (14,15.1)$$

El algoritmo termina puesto que, para el nuevo centroide, la distancia con los demás puntos no cambia es decir los conglomerados se mantienen.

Usando el lenguaje de programación R obtenemos

```
K-means clustering with 2 clusters of sizes 5, 2

Cluster means:
   x1    x2
1 14.00 15.1
2 11.25 11.5

Clustering vector:

[1] 2 2 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 9.200 0.625
between_SS / total_SS = 74.9 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss"
[6] "betweenss" "size" "iter" "ifault"
```

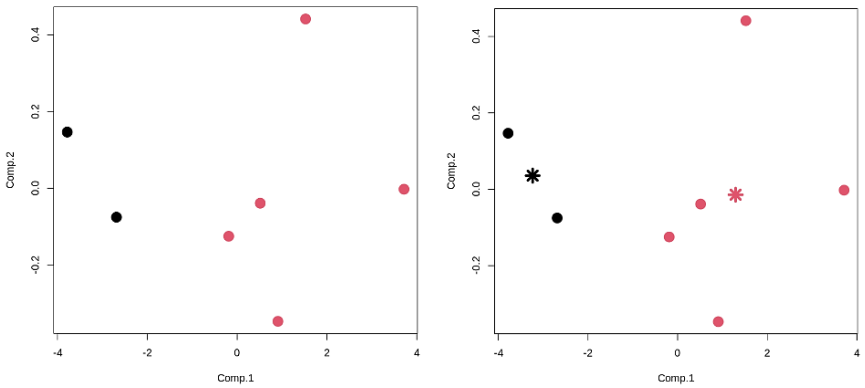

Figura 34: Grafico de k-medias para el registro de notas
Fuente: Elaboración propia

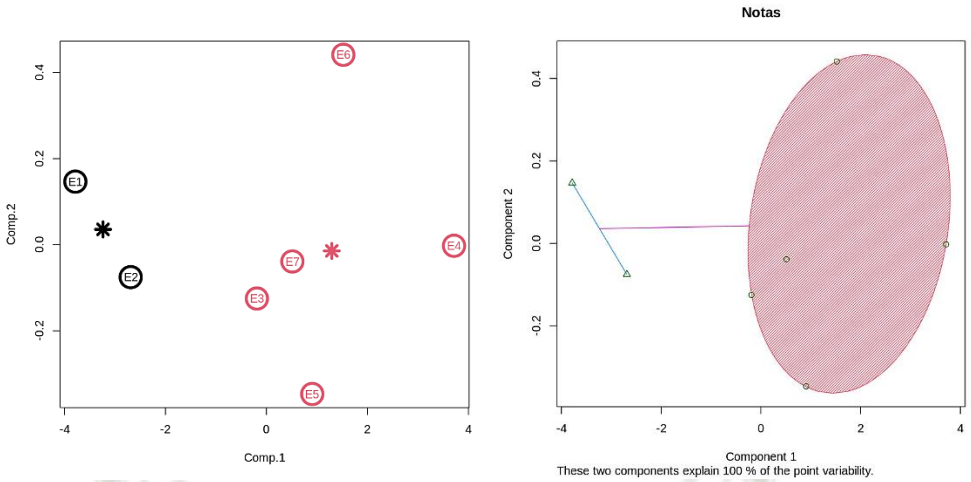

Figura 35: Grafico de k-medias para el registro de notas, dos grupos
Fuente: Elaboración propia

Clúster 1: Notas de los estudiantes E1 y E2

Tabla 19: Tabla de clúster 1

	Asignatura 1	Asignatura 2
E1	11	11
E2	11.5	12
Promedio	11.25	11.5

Fuente: Elaboración propia

Clúster 2: Notas de los estudiantes E3, E4, E5, E6 y E7

Tabla 20: Tabla de clúster 2

	Asignatura 1	Asignatura 2
E3	13	14
E4	15.5	17
E5	13.5	15
E6	14.5	15
E7	13.5	14.5
Promedio	14	15.1

Fuente: Elaboración propia

Tabla 21: Tabla de asignatura y clústeres

	Clúster 1	Clúster 2
Asignatura 1	11.25	11.5
Asignatura 2	14	15.1

Fuente: Elaboración propia

1.3.6. Hipótesis

Primera hipótesis

H_0: Es posible describir un modelo basado en regresión logística para analizar el rendimiento académico.

H_1: No es posible describir un modelo basado en regresión logística para analizar el rendimiento académico.

Segunda hipótesis

H_0: El modelo logístico permite determinar la incidencia que tiene la metodología de enseñanza y el uso de una matriz de evaluaciones en el rendimiento académico.

H_1: El modelo logístico No permite determinar la incidencia que tiene la metodología de enseñanza y el uso una matriz de evaluaciones en el rendimiento académico.

Tercera hipótesis

H_0: Las técnicas de agrupación o clasificación multivariada permiten determinar las actividades de evaluación más influyentes de la metodología de enseñanza.

H_1: Las técnicas de agrupación o clasificación multivariada No permiten determinar las actividades de evaluación más influyentes de la metodología de enseñanza.

1.3.7. Variables

Las variables tienen propiedades que se pueden definir por medio de la observación y que pueden presentar diversos valores de una unidad de observación (Tascón, 1980). Las variables en este estudio son el registro de notas de evaluaciones y la condición académica del estudiante.

Interrogante básica

Al fundamentar el uso de técnicas del análisis multivariado tales como la regresión logística como un modelo de clasificación supervisada, y el análisis de conglomerados que permitan el agrupamiento de individuos u objetos, de acuerdo a su registro de notas, se determinará su rendimiento académico, la Metodología de enseñanza y el sistema de evaluaciones está basado en una matriz de evaluaciones, con lo cual explicará si el uso de una matriz de evaluaciones influye en el rendimiento académico de los estudiantes universitarios que tienen dentro de su malla curricular cursos del área de matemáticas, como el caso de estudio, que son los estudiantes que llevan las asignaturas de Algebra Lineal y Geometría, cuyo dictado corresponde al Departamento Académico de Matemática y Estadística.

Tabla 22: Matriz de evaluación final semestre 2017-1

PROPUESTA MATRIZ DE EVALUACIÓN - EXAMEN FINAL
2017-1

DOCENTE: NESTOR URÉ BENAVIDES **ESCUELA PROF.:** INGENIERÍA INDUSTRIAL
ASIGNATURA: MATEMÁTICA II **SEMESTRE:** II

TABLA DE ESPECIFICACIONES

OBJETIVOS DE LA UNIDAD	CONTENIDOS	INDICADORES	PESO%	N° DE REACTIVOS	PUNTAJE
• Utiliza los conceptos aprendidos de la ecuación de la recta en R^2 para resolver un problema.	• Ecuación de la recta en R^2.	• Encuentra la ecuación de la recta a partir de los datos proporcionados (usando vectores y sus características).	20%	1	4
• Resolver problemas en R^3 que involucre, paralelismo ortogonalidad entre ellos. • Aplicar las fórmulas de distancias	• Rectas paralelas, perpendiculares • Planos paralelos, perpendiculares. • Fórmulas de distancias.	• Hallar la ecuación de un plano ó de una recta teniendo en cuenta su posición respecto a una recta ó un plano.	20%	1	4
• Identifica los elementos de las cónicas	• Circunferencia • Parábola • Elipse • Hipérbola	• Reconoce y diferencia los elementos y ecuaciones de las cónicas, para determinar la ecuación de otra la cónica pedida.	20%	1	4
• Determinar e identificar las superficies cuadráticas en el espacio a partir de su ecuación.	• Superficies cuadráticas	• Identifica la superficie. • Realiza la gráfica de la superficie, identificando las trazas. (independientes una de la otra)	20%	a) b)	4
• Reconocer y establecer una relación entre las coordenadas polares y las coordenadas rectangulares. • Identificar la cónica y trazar su gráfica.	• Coordenadas polares • Coordenadas rectangulares	• Convierte una ecuación de coordenadas polares a coordenadas rectangulares. • Convierte una ecuación de coordenadas rectangulares a coordenadas polares	20%	a) b)	4
			100%	4	20

MATRIZ DE EVALUACIÓN EXAMEN PARCIAL

MATEMÁTICA II - INGENIERÍAS 2017-1

OBJETIVOS DE LA UNIDAD	CONTENIDOS	INDICADORES	PESO%	N° DE REACTIVOS	PUNTAJE
					6
• Operar sistema de ecuaciones lineales, utilizando métodos de eliminación. • Identificar y operar con los tipos y propiedades de matrices y determinantes.	• Sistemas de ecuaciones lineales, matrices y determinantes.	• Determina la verdad o falsedad de proposiciones sobre sistemas de ecuaciones, matrices y determinantes	30%	1 a) b) c)	
• Aplicar operaciones con matrices en problemas cotidianos, analizando e interpretando la solución.	• Operaciones con matrices, aplicaciones.	• Utiliza matrices en la formulación y solución de un problema dado empleando diversas operaciones con las mismas	25%	1	5
• Aplicación de la inversa de una matriz en sistemas de ecuaciones lineales. • Identificar y calcular la inversa de una matriz.	• Matriz inversa y aplicaciones.	• Plantea y resuelve un problema que se formula como un sistema de ecuaciones lineales utilizando el método de la matriz inversa.	25%	1	5
• Aplicar adecuadamente definiciones y propiedades de vectores en R2 • Establecer y manejar los conceptos de norma de un vector, vector unitario. vectores paralelos, producto escalar.	• Vectores en R2.	• Resuelve un ejercicio de vectores utilizando correctamente las operaciones entre vectores y/o las normas, dirección y producto interno de vectores paralelos, perpendiculares, vectores unitarios.	20%	1	4
			100%	4	20

Tabla 23: Matriz de evaluación parcial semestre 2017-1

Fuente: Departamento de Matemática y Estadística de la Universidad Católica San Pablo

CAPITULO II

METODOLOGIA

El estudio propuesto en este proyecto tiene por finalidad establecer un modelo matemático-estadístico que permita analizar el rendimiento académico de estudiantes universitarios, así como establecer las características más influyentes, para lo cual se realizará una Investigación aplicada, con un nivel descriptivo, explicativo y correlacional.

2.1 Nivel de investigación

El nivel de la investigación es descriptivo, explicativo y correlacional, dado que se busca establecer relaciones causa-efecto, midiendo el grado de relación entre las variables.

2.2 Diseño de la Investigación

El estudio propuesto consiste en un trabajo de investigación aplicada, debido a que se utilizarán conocimientos de ciencias con la finalidad de determinar la probabilidad de ocurrencia de una categoría en función de diversas variables numéricas o categóricas, así como el análisis de clasificación de grupos, en el contexto del análisis del rendimiento académico.

a) **Análisis y Operacionalización de variables**

Variables

Las variables tienen propiedades que se pueden definir por medio de la observación y que pueden presentar diversos valores de una unidad de observación (Tascón, 1980). Las variables en este estudio son el registro de notas de evaluaciones y la condición académica del estudiante.

Análisis de Variables

Para el análisis de las variables es importante indicar la forma de estructurar y mostrar los datos obtenidos luego de la observación realizada a la población indicada (Sanchez Fernández, 2004).

Variable Independiente

Registro de notas de las evaluaciones, tipo de evaluación, aplicadas a los estudiantes universitarios desde el año 2008 al 2019.

Variable Dependiente

Condición académica del estudiante

Operacionalización de Variables

Tabla 24: Cuadro de Operacionalización de las Variables.

TIPO	VARIABLE	DIMENSIÓN	INDICADOR
Independiente	Evaluaciones	Controles	Cuatro controles semestrales
		Practicas calificadas	Cuatro practicas calificadas semestrales
		Trabajos	Dos trabajos
		Participación en aula	Intervenciones orales
Dependiente	Condición académica del Estudiante	Aprobado	• Evaluaciones aprobadas • Nota mínima aprobatoria • Evaluaciones efectuadas por los estudiantes.
		Desaprobado	• Evaluaciones desaprobadas • Notas mínima y máxima desaprobatoria

- Evaluaciones no efectuadas por los estudiantes

Fuente: Elaboración propia

Interrogante básica

Al fundamentar el uso de técnicas del análisis multivariado tales como la regresión logística como un modelo de clasificación supervisada, y el análisis de conglomerados que permitan el agrupamiento de individuos u objetos, de acuerdo a su registro de notas, se determinará su rendimiento académico, la Metodología de enseñanza y el sistema de evaluaciones está basado en una matriz de evaluaciones, con lo cual explicará si el uso de una matriz de evaluaciones influye en el rendimiento académico de los estudiantes universitarios que tienen dentro de su malla curricular cursos del área de matemáticas, como el caso de estudio, que son los estudiantes que llevan la asignatura de Algebra Lineal y Geometría, cuyo dictado corresponde al Departamento Académico de Matemática y Estadística.

2.3 Población y Muestra

Se utilizó el registro de notas correspondientes a las asignaturas de: Algebra lineal y Geometría Analítica, Algebra y Geometría y Matemática II del Programa profesional de Ingeniería Industrial, de la Universidad Católica San Pablo. Se consideraron en el análisis los registros de notas de la Evaluación permanente 1, Evaluación Permanente 2, Examen Parcial, Examen Final y la nota promedio, correspondiente a las notas de 938 estudiantes.

Los grupos estudiantes corresponden a los periodos lectivos:

2008-1, 2009-1, 2009-2, 2010-1, 2010-2, 2011-1, 2011-2, 2013-1, 2013-2, 2014-1, 2014-2, 2014-3, 2015-1, 2015-2, 2015-R, 2016-1, 2016-2, 2017-1, 2017-2, 2017-R, 2018-1, 2019-1, 2019-2

Se utilizó una muestra censal de todos los registros de notas

2.4 Técnicas e instrumentos de recolección de datos

Las técnicas e instrumentos utilizados en la presente investigación se toman de acuerdo a la siguiente clasificación (Hernández Sampieri & Otros, 2005)

2.4.1 Técnicas

- Observación: Técnica que utilizamos para la obtención de datos respecto a las evaluaciones continuas y parciales.
- Encuesta: Técnica de la cual se hace uso para obtener información sobre la metodología de enseñanza, realizada por la Universidad Católica San Pablo.

2.4.2 Instrumentos

Ficha de recolección de datos.

2.5 Campo de Verificación

Ubicación espacial

La ubicación de la investigación se determina en:

Tabla 25: Ámbito de la investigación.

UBICACIÓN ESPACIAL	
Región	Arequipa
Provincia	Arequipa
Distrito	Arequipa

Fuente: Elaboración propia.

Ubicación temporal

La investigación se desarrolla desde el mes de enero hasta abril del 2021

Unidades de estudio

Se considera como unidades de estudio los diferentes modelos de regresión logística y modelos de análisis de conglomerados, se utilizó como datos las notas de las evaluaciones de alumnos de la asignatura de Matemática II, así como los de Algebra Lineal y Geometría.

2.6 Técnicas de procesamiento y análisis de datos

- Organización del trabajo de campo: Se utilizó la base de datos proporcionada por la Universidad Católica San Pablo.
- Tratamiento estadístico: Se realizó la codificación de variables del modelo para finalmente elaborar cuadros y gráficos que muestren con claridad los resultados, utilizando el programa estadístico SPSS y el lenguaje de programación R.

CAPÍTULO III

RESULTADOS Y DISCUSIÓN

3.1 Análisis exploratorio

En el análisis de datos la falta de normalidad y homogeneidad de la varianza pueden influenciar en las pruebas de hipótesis y cometer errores del tipo I y tipo II. Generalmente los datos a analizar son asimétricos y multimodales, cuando se comparan grupos las varianzas pueden no tener varianza homogénea. Los modelos a utilizarse deben verificar sus supuestos o condiciones para lo cual se debe realizar pruebas de hipótesis y análisis gráfico de los datos.

Prueba de normalidad

Esta prueba se usa para determinar si un conjunto de observaciones proviene de una población normal, los contrastes univariados más utilizados son:

- **Shapiro y Wilks:** Este test se emplea para contrastar normalidad cuando el tamaño de la muestra es menor o igual a 50

- **Kolmogorov y Smirnov:** Este test se utiliza si el tamaño de la muestra es mayor a 50, la debilidad de este test es que asume conocida la media y varianza poblacional, lo que en general no se conoce. Para salvar esta consideración, se desarrolló una modificación del test de Kolmogorov - Smirnov conocido como test de Lilliefors. El test Lilliefors asume que la media y varianza son desconocidas.

Hipótesis para la prueba de normalidad

H_0: La distribución de los datos se aproximan a una distribución normal

H_1: La distribución de los datos no se aproximan a una distribución normal

3.1.1 Análisis de la nota permanente 1

Para determinar si los datos provienen de una población Normal, realizamos la prueba de $Lilliefors$ usando el programa R, obtenemos:

```
Lilliefors (Kolmogorov-Smirnov) normality test data: x D = 0.0665,
p-value = 1.569e-10
```

Siendo el $valor\ P = 1.569e - 10 < 0.05$ rechazamos la hipótesis nula, es decir concluimos que los datos no provienen de una población normal

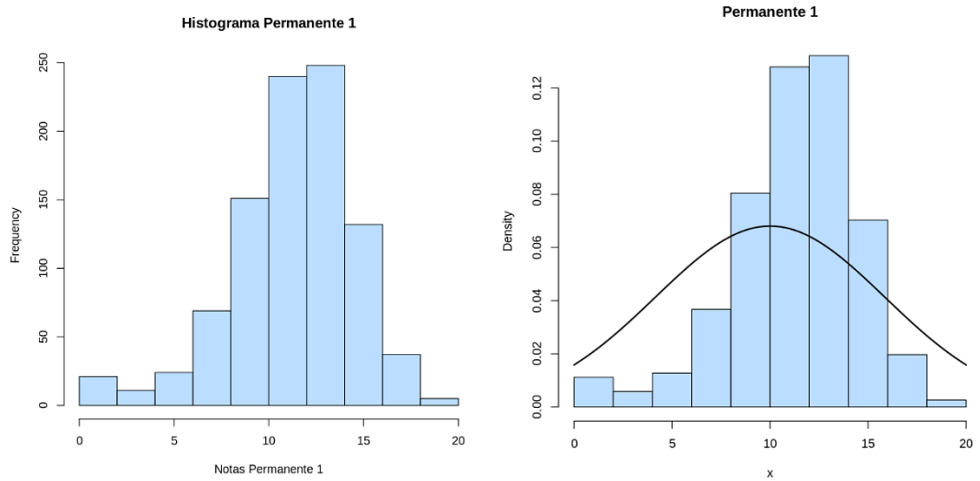

Figura 36: Grafica de la prueba de normalidad para la nota permanente 1
Fuente: Elaboración propia.

Siendo su media aritmética $\bar{x} = 11.32$ y su desviación estándar muestral $s = 3.3$ su coeficiente de variación $CV = 29.12\%$ nos indica que las notas permanentes son moderadamente homogéneas. Usando su mediana $M_e = 11.65$ establecemos que el 50% de las notas permanentes centrales alrededor de la nota mediana de 11.65 oscilan entre $Q_1 = 9.7$ y $Q_3 = 13.5$, notamos de acuerdo al diagrama de caja de la figura 37 que existen valores atípicos. Los 32 valores atípicos son:

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.05	1.3	1.3	1.42	2.17	2.33	2.39	2.83	3.08	3.15	3.15	3.38	3.45	3.67	19.4

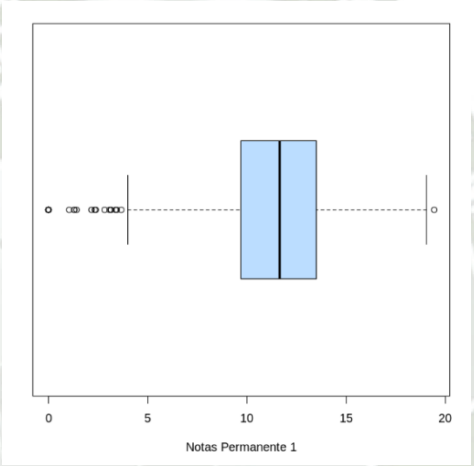

Figura 37: diagrama de caja para la nota permanente 1
Fuente: Elaboración propia.

Al realizar la gráfica de la serie de notas desde el periodo 2008-02 hasta el periodo 2019-02 podemos observar gráficamente que la serie de notas correspondientes a la *nota permanente* 1, oscilan aproximadamente entre 6 y 16 alrededor de la nota de 11, lo cual puede ser apreciado mejor usando un promedio móvil de periodo 4. Adicionalmente podemos apreciar que en los semestres 2014-03, 2015-R y 2017-R las notas de las evaluaciones permanentes tienen un rango de variación diferente posiblemente debido a la presencia de estudiantes de segunda matricula.

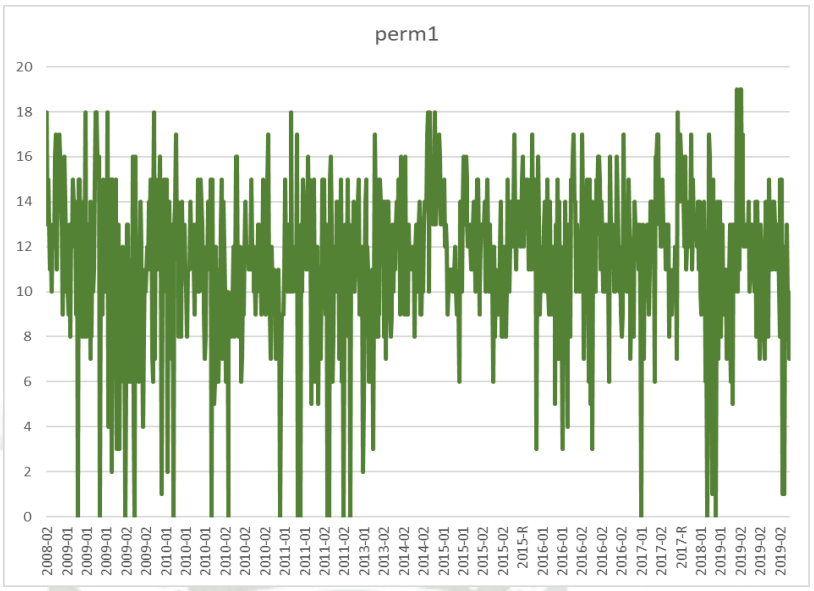

Figura 38: Serie de notas correspondiente a la nota permanente 1
Fuente: Elaboración propia.

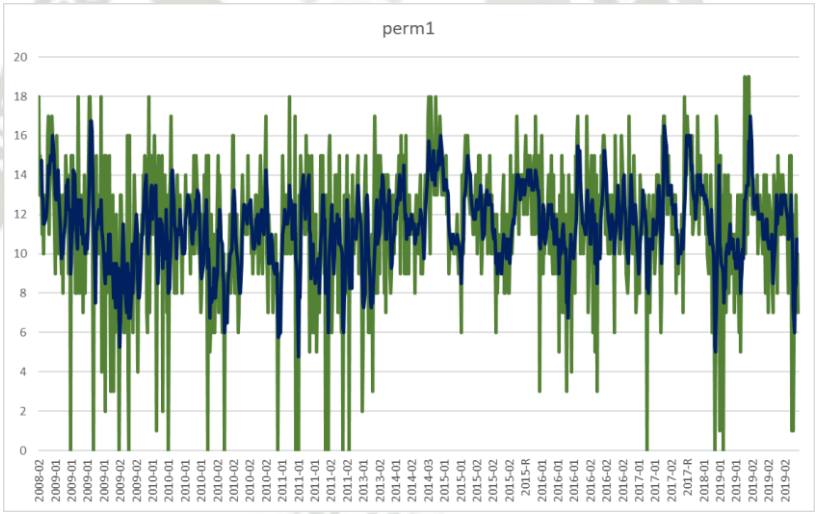

Figura 39: Serie de notas con media móvil de periodo 3, correspondiente a la nota permanente 1
Fuente: Elaboración propia.

3.1.2 Análisis de la nota permanente 2

Para determinar si los datos provienen de una población Normal, realizamos la prueba de $Lilliefors$ usando el lenguaje de programación R, obtenemos:

```
        Lilliefors (Kolmogorov-Smirnov) normality test data: x D = 0.0750,
p-value = 1.449e-13
```

Siendo el $valor\ P = 1.449e - 13 < 0.05$ rechazamos la hipótesis nula, es decir concluimos que los datos no provienen de una población normal

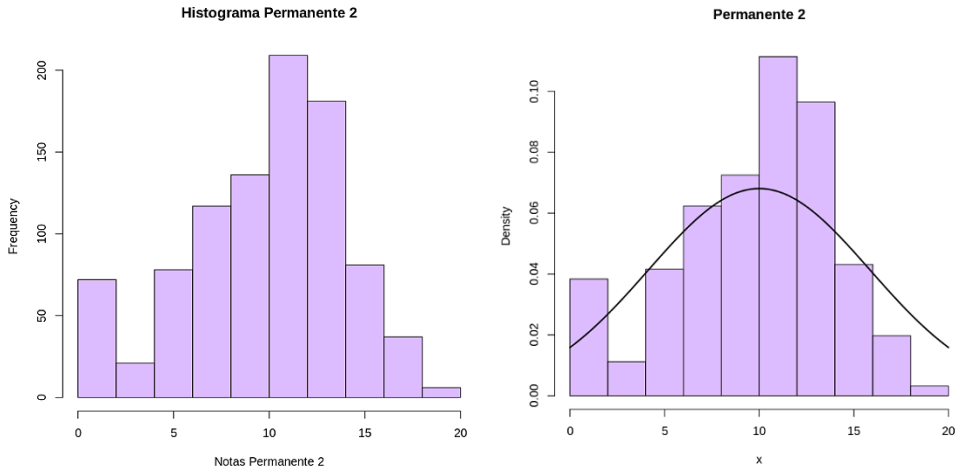

Figura 40: Gráfica de la prueba de normalidad para la nota permanente 2
Fuente: Elaboración propia.

Siendo su media aritmética $\bar{x} = 9.73$ y su desviación estándar muestral $s = 4.2500$ y siendo su coeficiente de variación $CV = 43.69\%$ lo que nos indica que las notas permanentes son moderadamente homogéneas. Usando su mediana $M_e = 10.4650$ establecemos que el 50% de las notas permanentes centrales alrededor de la nota mediana de 10.4650 oscilan entre $Q_1 = 7.1400$ y $Q_3 = 12.7500$, no existen valores atípicos.

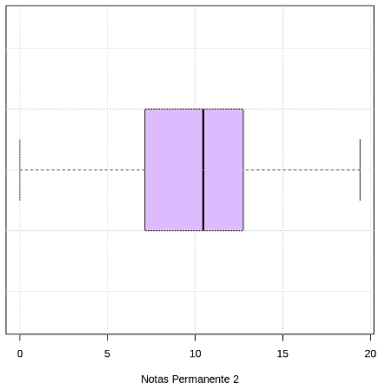

Figura 41: Diagrama de caja para la nota permanente 2
Fuente: Elaboración propia.

Al realizar la gráfica de la serie de notas desde el periodo 2008-02 hasta el periodo 2019-02 podemos observar gráficamente que la serie de notas correspondientes a la *nota permanente* 2, oscilan aproximadamente entre 5 y 15 alrededor de la nota de 10, lo cual puede ser apreciado mejor usando un promedio móvil de periodo 4. Adicionalmente podemos apreciar que en los semestres 2014-03 y 2015-R las notas de las evaluaciones permanentes tienen un rango de variación diferente posiblemente debido a la presencia de estudiantes de segunda matricula.

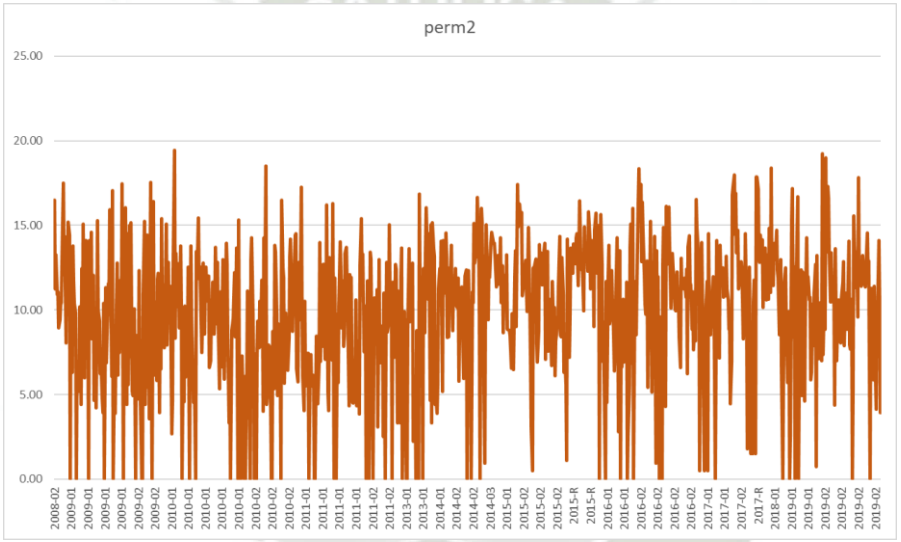

Figura 42: Serie de notas correspondiente a la nota permanente 2
Fuente: Elaboración propia.

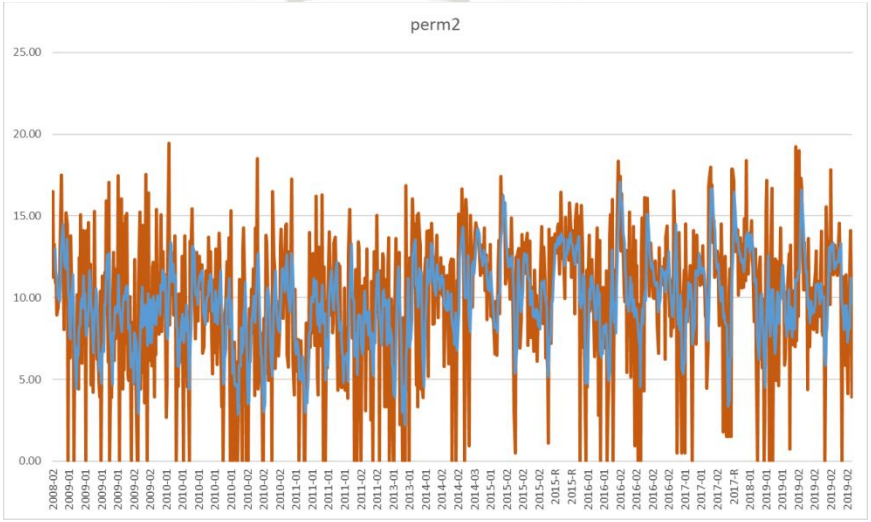

Figura 43: Serie de notas con media móvil de periodo 3, correspondiente a la nota permanente 2
Fuente: Elaboración propia.

3.1.3 Contraste de las notas permanentes

3.1.3.1 Análisis de regresión simple

Dado que el coeficiente de determinación es $r^2 = 0.5650$, su coeficiente de correlación es $r = 0.7517$, lo cual indica que existe una relación directa entre las notas permanentes es decir si aumenta la nota permanente 1 aumenta la nota permanente 2, la correlación es fuerte, siendo que el 56.5% de las notas se describen a través de la recta de regresión $\hat{y} = 0.9695x - 1.2429$.

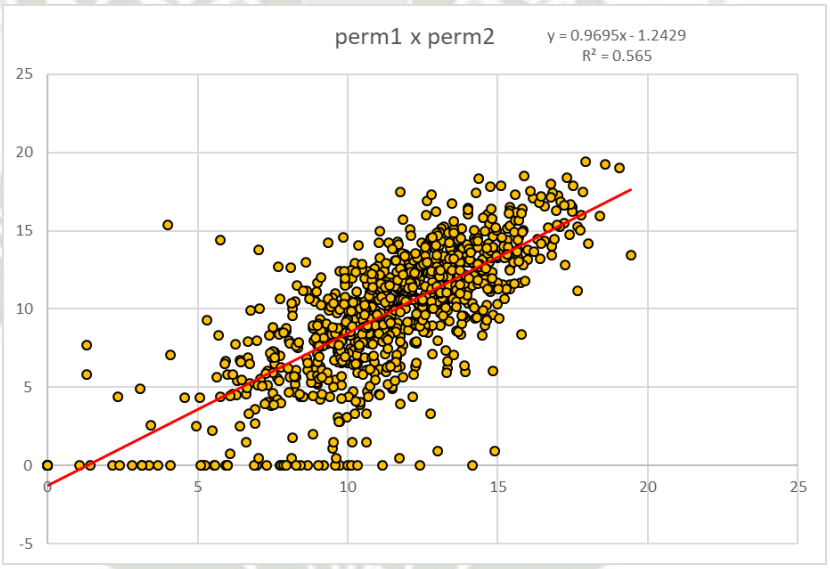

Figura 44: Correlación entre las notas permanentes
Fuente: Elaboración propia.

Según el informe ANOVA, para probar la hipótesis que existe una relación lineal entre la nota permanente 1 y el examen parcial, para un nivel de significancia del 5% tenemos

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Dado que el $valor\ P = 2.288E - 171 < 0.05 = \alpha$ por lo cual aceptamos que existe una relación lineal entre las notas permanentes

Estadísticas de la regresión	
Coeficiente (	0.751655239
Coeficiente (	0.564985599
R^2 ajustad(	0.56452084
Error típico	2.805332884
Observacion	938

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	9567.057779	9567.057779	1215.6529	2.2883E-171
Residuos	936	7366.219465	7.869892591		
Total	937	16933.27724			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	-1.242885434	0.327779498	-3.791833964	0.00015912	-1.886153252	-0.599617616	-1.886153252	-0.599617616
perm1	0.969543189	0.027807527	34.86621436	2.288E-171	0.914970871	1.024115507	0.914970871	1.024115507

Figura 45: Informe ANOVA, correspondiente a las notas permanentes
Fuente: Elaboración propia.

Dado que el coeficiente de determinación es $r^2 = 0.4567$, su coeficiente de correlación es $r = 0.6758$, lo cual indica que existe una relación directa entre la nota permanente 1 y examen parcial es decir si aumenta la nota permanente 1 aumenta la nota del examen parcial, la correlación es fuerte, siendo que el 45.67% de las notas se describen a través de la recta de regresión $\hat{y} = 0.8707x - 0.9747$.

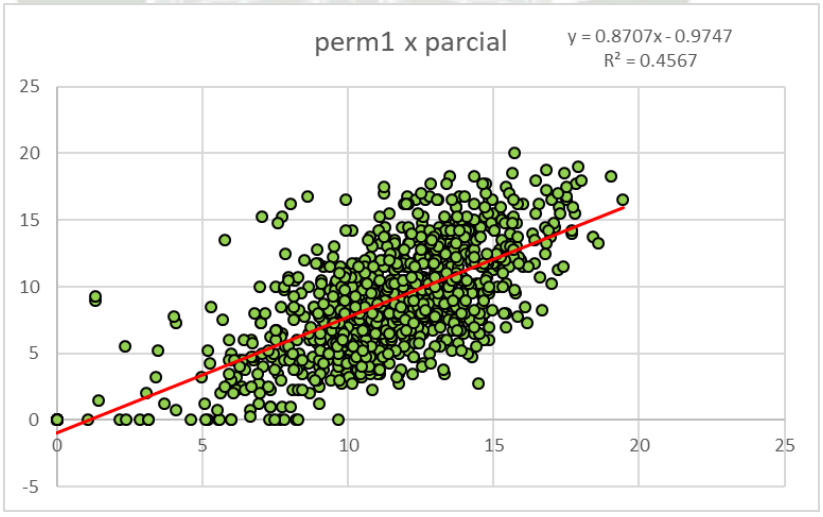

Figura 46: Correlación entre la nota permanente 1 y el examen parcial
Fuente: Elaboración propia.

Según el informe ANOVA, para probar la hipótesis que existe una relación lineal entre la nota permanente 1 y el examen parcial, para un nivel de significancia del 5% ($p < 0.05$) tenemos

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Dado que el $valor\ P = 3.874E - 126 < 0.05 = \alpha$ por lo cual aceptamos que existe una relación lineal entre la nota permanente 1 y el examen parcial

Estadísticas de la regresión	
Coeficiente (	0.675788367
Coeficiente (	0.456689918
R^2 ajustadc	0.456109458
Error típico	3.131522744
Observacion	938

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	7715.439118	7715.439118	786.773109	3.8742E-126
Residuos	936	9178.822876	9.806434696		
Total	937	16894.26199			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	-0.974675959	0.365892034	-2.663834873	0.00785832	-1.692739692	-0.256612226	-1.692739692	-0.256612226
perm1	0.870679433	0.031040845	28.0494761	3.874E-126	0.809761723	0.931597143	0.809761723	0.931597143

Figura 47: Informe ANOVA, correspondiente la nota permanente 1 y el examen parcial
Fuente: Elaboración propia.

Dado que el coeficiente de determinación es $r^2 = 0.6541$, su coeficiente de correlación es $r = 0.8088$, lo cual indica que existe una relación directa entre la nota permanente 2 y examen final es decir si aumenta la nota permanente 2 aumenta la nota del examen final, la correlación es fuerte, siendo que el 65.41% de las notas se describen a través de la recta de regresión $\hat{y} = 1.0216x - 2.5681$.

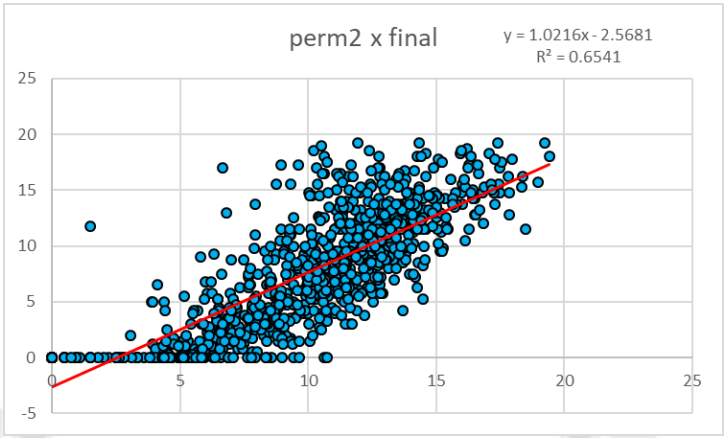

Figura 48: Correlación entre la nota permanente 2 y el examen final
Fuente: Elaboración propia.

Según el informe ANOVA, para probar la hipótesis que existe una relación lineal entre la nota permanente 2 y el examen final, para un nivel de significancia del 5% ($p < 0.05$) tenemos

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Dado que $valor\ P = 5.592E - 218 < 0.05 = \alpha$ por lo cual aceptamos que existe una relación lineal entre la nota permanente 2 y el examen final

Estadísticas de la regresión	
Coeficiente (	0.808753345
Coeficiente (	0.654081973
R^2 ajustadc	0.653712402
Error típico	3.160077558
Observacion	938

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	17673.81545	17673.81545	1769.84337	5.5922E-218
Residuos	936	9346.9804	9.986090171		
Total	937	27020.79585			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	-2.568132268	0.257838272	-9.960244629	2.7902E-22	-3.074140312	-2.062124223	-3.074140312	-2.062124223
perm2	1.021632378	0.024284392	42.06950638	5.592E-218	0.973974218	1.069290538	0.973974218	1.069290538

Figura 49: Informe ANOVA, correspondiente la nota permanente 2 y el examen final
Fuente: Elaboración propia.

Dado que el coeficiente de determinación es $r^2 = 0.3688$, su coeficiente de correlación es $r = 0.6073$, lo cual indica que existe una relación directa entre la nota del examen parcial y la nota del examen final es decir si aumenta la nota del examen parcial aumenta la nota del examen final, la correlación es fuerte, siendo que el 36.88% de las notas se describen a través de la recta de regresión $\hat{y} = 0.7681x + 0.5525$

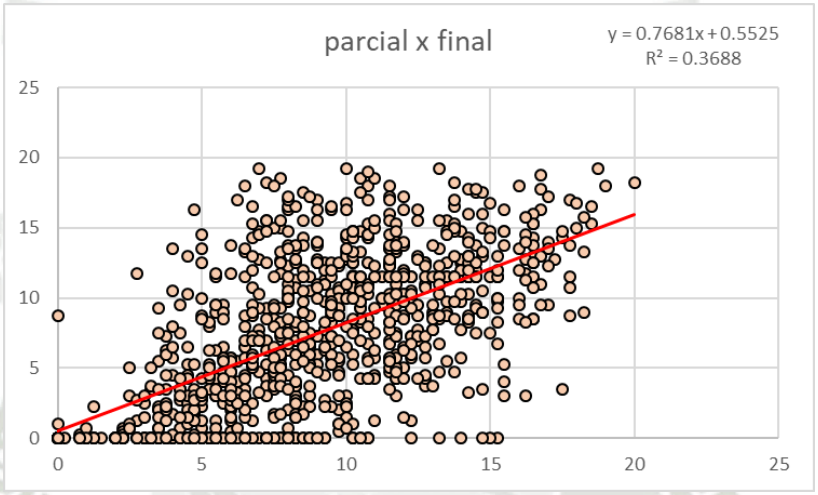

Figura 50: Correlación entre la nota del examen parcial y el examen final
Fuente: Elaboración propia.

Según el informe ANOVA, para probar la hipótesis que existe una relación lineal entre la nota del examen parcial y el examen final, para un nivel de significancia del 5% ($p < 0.05$) tenemos

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Dado que el $valor\ P = 1.2504E - 95 < 0.05 = \alpha$, por lo cual aceptamos que existe una relación lineal entre la nota del examen parcial y el examen final.

Estadísticas de la regresión	
Coeficiente (	0.607322484
Coeficiente (	0.3688406
R^2 ajustad(	0.368166284
Error típico	4.268552932
Observacion	938

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	9966.366547	9966.366547	546.985122	1.25042E-95
Residuos	936	17054.42931	18.22054413		
Total	937	27020.79585			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	0.55253117	0.323203896	1.709543657	0.08768174	-0.081757021	1.186819361	-0.081757021	1.186819361
parcial	0.768066486	0.032840598	23.38771305	1.2504E-95	0.703616757	0.832516215	0.703616757	0.832516215

Figura 51: Informe ANOVA, correspondiente la nota del examen parcial y el examen final
Fuente: Elaboración propia.

3.1.3.2 Análisis de correlación múltiple para el porcentaje de alumnos aprobados

Consideramos la Tabla 26, correspondiente a los alumnos que aprobaron la asignatura, durante los periodos lectivos desde el 2008-2 al 2019-2

Tabla 26: Porcentaje de estudiantes aprobados en los diferentes periodos lectivos

PERIODO	perm1	perm2	parcial	final	promedio
2008-2	83.33	41.67	25.00	16.67	33.33
2009-1	59.42	30.43	18.84	21.74	27.54
2009-2	26.09	26.09	8.70	15.22	19.57
2010-1	46.24	31.18	23.66	15.05	20.43
2010-2	41.27	26.98	7.94	15.87	12.70
2011-1	36.92	24.62	27.69	10.77	13.85
2011-2	44.44	28.89	31.11	11.11	17.78
2013-1	34.62	26.92	19.23	30.77	23.08
2013-2	70.00	60.00	30.00	40.00	60.00
2014-1	50.00	50.00	33.33	25.00	41.67
2014-2	58.33	36.11	25.00	27.78	25.00
2014-3	94.44	66.67	77.78	38.89	61.11
2015-1	46.67	53.33	46.67	36.67	36.67
2015-2	58.33	43.33	26.67	31.67	38.33
2015-R	87.10	87.10	35.48	54.84	58.06
2016-1	47.83	34.78	21.74	32.61	26.09
2016-2	58.46	47.69	41.54	43.08	41.54
2017-1	53.66	41.46	21.95	36.59	31.71
2017-2	62.96	48.15	29.63	55.56	55.56
2017-R	93.33	80.00	66.67	66.67	80.00
2018-1	65.22	52.17	39.13	56.52	52.17
2019-1	30.77	33.33	35.90	25.64	25.64
2019-2	56.92	38.46	63.08	32.31	46.15

Fuente: Elaboración propia.

Dada la matriz de correlaciones de la Tabla 27 y la gráfica de correlaciones múltiple de la Figura 52, tenemos que:

- La mayor correlación lineal se da entre el porcentaje de aprobados de la nota promedio y de la nota permanente 2, $cor(perm2, promedio) = 0.9015$

- En las evaluaciones previas al porcentaje de aprobados de la nota promedio el porcentaje de aprobados de las notas con mayor correlación lineal son la nota permanente 1 y la nota permanente 2, $cor(perm1, perm2) = 0.8219$

- El porcentaje de aprobados de las notas menos correlacionadas son la nota de la evaluación parcial y la nota final, $cor(parcial, final) = 0.5284$

Tabla 27: Matriz de correlaciones

	perm1	*perm2*	*parcial*	*final*	*promedio*
perm1	1				
perm2	0.8219	1			
parcial	0.6005	0.6308	1		
final	0.6238	0.8039	0.5284	1	
promedio	0.8128	0.9014	0.7143	0.8701	1

Fuente: Elaboración propia.

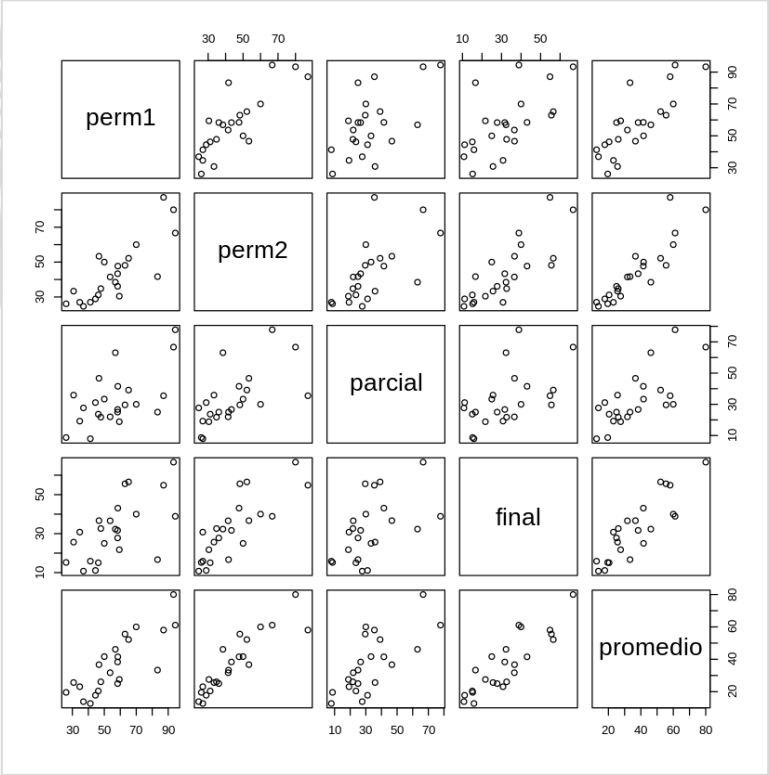

Figura 52: Gráfica de correlaciones múltiple
Fuente: Elaboración propia.

Al realizar un gráfico de líneas para la Tabla 27, es posible verificar que el porcentaje de estudiantes aprobados es superior al 20% aproximadamente

desde el periodo lectivo 2013-2 en todas las evaluaciones, algunos valores porcentuales elevados en los cursos de verano (2014-3, 2015-R y 2017-R)

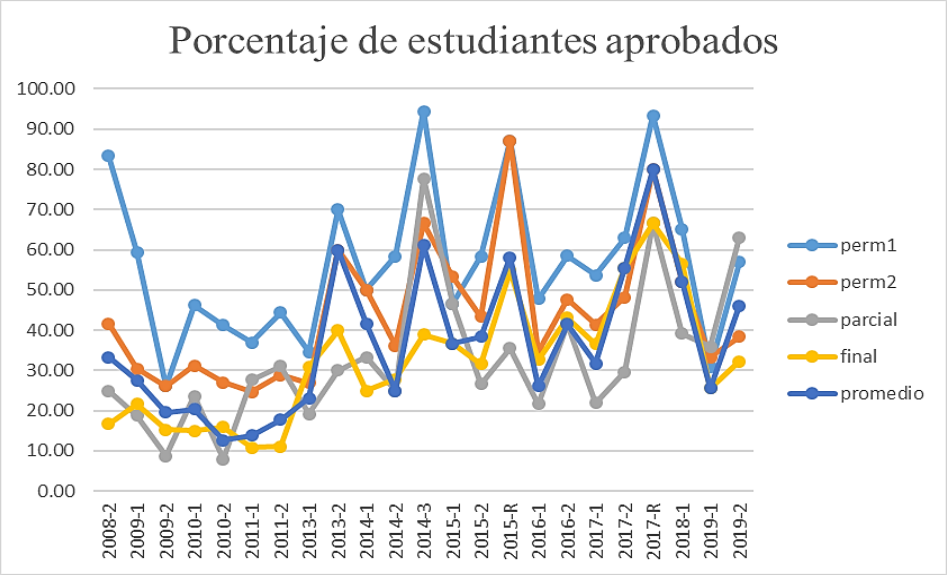

Figura 53: Porcentaje de estudiantes aprobados desde el 2008-2 al 2019-2
Fuente: Elaboración propia.

3.2 Análisis de regresión logística

Para elegir el tamaño de muestra se debe considerar que No todos los conjuntos de datos pueden ajustarse a un modelo logístico. Respecto al tamaño de muestra mínimo se recomienda una muestra diez veces mayor que el número de variables incluyendo a la dependiente que hay en el modelo (Cáceres, 2007)

Si k es el numero de variables independientes y n es el tamaño de muestra entonces

$$n \geq 10(k+1)$$

Por ejemplo, si existieran $k = 3$ variables independientes el tamaño de muestra mínimo debe ser $n = 40$

Considerando las dos notas permanentes y la variable dependiente tenemos que $k = 3$ en nuestra base de datos tenemos las evaluaciones permanentes desde el periodo 2008-02 al 2019-02, haciendo un total de $n = 938$ registros de notas, por lo cual se cumple con la condición mínima del tamaño de muestra

Donde

- **perm1:** Nota permanente 1
- **perm2**: Nota permanente 2
- **Y**: Condición del estudiante; 1 - aprobado, 0- desaprobado

INFORME03.sav [ConjuntoDatos1] - IBM SPSS Statistics Editor de datos

Archivo Editar Ver Datos Transformar Analizar Gráficos Utilidades Ampliaciones Ventana Ayuda

15 :

	perm1	perm2	Y	var	var	var	var	var
1	17,50	16,50	1					
2	13,20	11,25	0					
3	15,00	13,25	1					
4	13,40	10,92	0					
5	11,30	11,08	0					
6	12,25	8,92	0					
7	10,25	9,42	0					
8	13,25	10,33	0					
9	11,80	10,42	0					
10	12,50	12,92	0					
11	11,75	17,50	1					
12	16,00	13,17	1					
13	16,58	14,33	1					
14	11,00	8,06	0					
15	16,42	13,19	1					
16	15,42	11,58	1					
17	16,67	15,19	1					
18	16,33	14,44	1					
19	14,17	,00	0					
20	12,58	9,81	0					
21	9,08	6,33	0					
22	14,83	13,78	1					
23	15,75	11,69	0					
24	14,00	10,39	0					
25	12,42	9,83	0					
26	9,75	,00	0					
27	10,00	6,56	0					
28	10,33	5,94	0					
29	9,00	5,19	0					
30	13,00	10,19	0					
31	8,42	4,42	0					
32	13,42	12,44	1					

Figura 54: Base de datos para el análisis de las notas permanentes y la condición del estudiante en SPSS
Fuente: Elaboración propia.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	perm1	,476	,068	49,589	1	,000	1,609	1,409	1,837
	perm2	,739	,070	112,234	1	,000	2,094	1,826	2,401
	Constante	-15,100	1,115	183,301	1	,000	,000		

a. Variables especificadas en el paso 1: perm1, perm2.

Figura 55: Estimación de parámetros del modelo logístico, usando el programa SPSS
Fuente: Elaboración propia.

```
M=read_xlsx("INFORME02.xlsx")
M
```

A tibble: 938 × 3

perm1	perm2	Y
<dbl>	<dbl>	<dbl>
17.50	16.50	1
13.20	11.25	0
15.00	13.25	1
13.40	10.92	0
11.30	11.08	0
12.25	8.92	0
10.25	9.42	0
13.25	10.33	0
11.80	10.42	0
12.50	12.92	0
11.75	17.50	1
16.00	13.17	1
16.58	14.33	1

Figura 56: Base de datos para el análisis de las notas permanentes y la condición del estudiante usando el lenguaje R
Fuente: Elaboración propia.

Usando la base de datos de notas y condición del estudiante en el lenguaje de programación R, obtenemos de forma análoga la estimación de parámetros

```
Call:

glm(formula = Y1 ~ X1 + X2, family = binomial, data = M)

Deviance Residuals:

    Min      1Q  Median      3Q     Max
-2.2516 -0.3878 -0.0595  0.2586  2.7657

Coefficients:

            Estimate     Std. Error    z value     Pr(>|z|)
```

```
(Intercept) -15.1002      1.1153     -13.5390    < 2e-16 ***
X1            0.4756      0.0675       7.0420    1.9e-12 ***
X2            0.7390      0.0698      10.5940    < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1171.19 on 937 degrees of freedom
Residual deviance:  531.43 on 935 degrees of freedom

AIC: 537.43

Number of Fisher Scoring iterations: 7
```

El modelo logístico está dado por

$$p(x_1, x_2) = \frac{1}{1 + e^{-(-15.1000+0.4760x_1+0.7390x_2)}}$$

Al realizar algunos cálculos de probabilidades con el modelo obtenemos:

- Si un Estudiante obtiene en sus notas permanentes $x_1 = 14$ y $x_2 = 12$, la probabilidad de que apruebe la asignatura está dada por

$$P[Y = 1|x_1 = 14, x_2 = 12\ ] = p(14,12) = \frac{1}{1 + e^{-(-15.100+0.476(14)+0.739(12))}} = 0.6064$$

Es decir, existe una probabilidad de 60.64% de que un estudiante apruebe la asignatura al final del semestre.

- Si un Estudiante obtiene en sus notas permanentes $x_1 = 18$ y $x_2 = 16$, la probabilidad de que apruebe la asignatura está dada por

$$P[Y = 1|x_1 = 18, x_2 = 16\ ] = p(18,16) = \frac{1}{1 + e^{-(-15.1+0.476(18)+0.739(16))}} = 0.9950$$

Es decir, existe una probabilidad de 99.50% de que un estudiante apruebe la asignatura al final del semestre

- Si un Estudiante obtiene en sus notas permanentes $x_1 = 11$ y $x_2 = 10$, la probabilidad de que apruebe la asignatura está dada por

$$P[Y = 1|x_1 = 11, x_2 = 10\ ] = p(11,10) = \frac{1}{1 + e^{-(-15.1+0.476(11)+0.739(10))}} = 0.0777$$

Es decir, existe una probabilidad de 7.77% de que un estudiante apruebe la asignatura al final del semestre

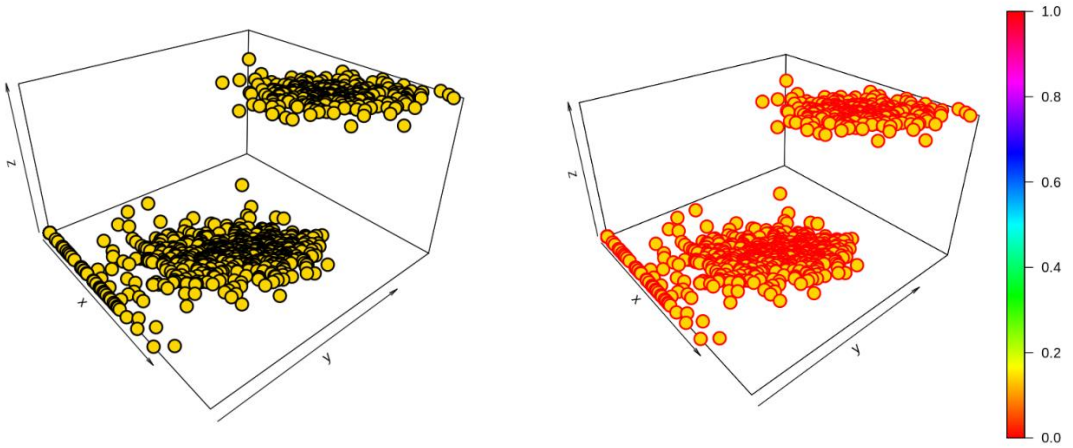

Figura 57: Base de datos de las notas permanentes considerando la condición final del estudiante
Fuente: Elaboración propia.

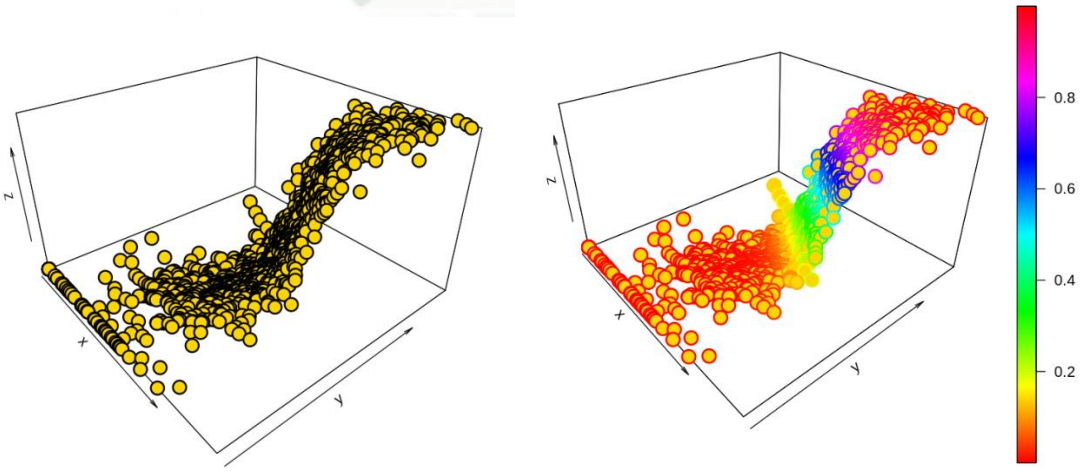

Figura 58: Base de datos de las notas permanentes ajustada al modelo logístico
Fuente: Elaboración propia.

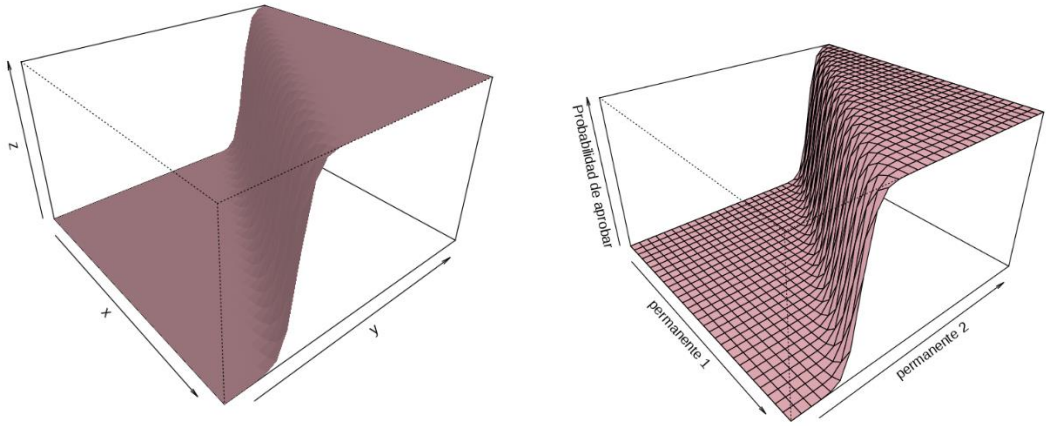

Figura 59: Modelo logístico para las notas permanentes
Fuente: Elaboración propia.

3.2.1 Prueba de hipótesis para los coeficientes del modelo de regresión logística

Para determinar la calidad del modelo usaremos la prueba de *Wald*, esta prueba se usa para evaluar la significancia estadística de cada variable explicativa o regresora. Según *Wald* los parámetros estimados por máxima verosimilitud para los modelos logísticos tienen una distribución normal cuando las muestras son grandes

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_1: para\ algun\ \ \beta_j \neq 0\,, \qquad j = 1{,}2$$

La prueba de H_0 determinará si las variables X_1: $Nota\ Permanente\ 1$ y X_2: $Nota\ Permanente\ 2$ no influyen significativamente sobre la condición de aprobar la asignatura para el estudiante. Con lo cual se dará también la validez del modelo ajustado con un nivel de significancia especificado que por lo general es del 5% ($p < 0.05$).

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	perm1	,476	,068	49,589	1	,000	1,609	1,409	1,837
	perm2	,739	,070	112,234	1	,000	2,094	1,826	2,401
	Constante	-15,100	1,115	183,301	1	,000	,000		

a. Variables especificadas en el paso 1: perm1, perm2.

Figura 60: Prueba de Wald del modelo logístico, usando el programa SPSS
Fuente: Elaboración propia.

Considerando $\alpha = 0.05$, la prueba de $Wald$ considera el estadístico

$$W^2 = \frac{\hat{\beta}_j^{\,2}}{\sigma^2(\hat{\beta}_j)}$$

$$Wald = W^2 \sim \chi_1^2$$

- Para β_1

$$W^2 = \frac{\hat{\beta}_1^{\,2}}{\sigma^2(\hat{\beta}_1)} = \frac{0.4756^2}{0.0675^2} = 49.5890$$

$$Valor\ p = 1 - P[\chi^2 < 49.589] = 1.8957E - 12 \approx 0.0000 < 0.05$$

- Para β_2

$$W^2 = \frac{\hat{\beta}_2^{\,2}}{\sigma^2(\hat{\beta}_2)} = \frac{0.7390^2}{0.0697^2} = 112.2340$$

$$Valor\ p = 1 - P[\chi^2 < 112.2340] \approx 0.0000 < 0.05$$

En ambos casos el $valor\ p\ < 0.05$, por lo cual aceptamos la hipótesis alterna, las variables X_1: $Nota\ Permanente\ 1$ y X_2: $Nota\ Permanente\ 2$ influyen significativamente sobre la condición de aprobar la asignatura para el estudiante. Con lo cual también se da la validez del modelo ajustado con un nivel de significancia del 5% ($p < 0.05$).

3.3 Métodos de clasificación de conglomerados

Se utilizará un método de clasificación jerárquico generando clases anidadas de un dendograma, posteriormente utilizaremos un método de clasificación no jerárquico mediante el uso de centroides.

3.3.1 Método de clasificación jerárquica usando un dendograma

Tabla 28: Porcentaje de estudiantes aprobados en los diferentes periodos lectivos

PERIODO	perm1	perm2	parcial	final	promedio
2008-2	83	42	25	17	33
2009-1	59	30	19	22	28
2009-2	26	26	9	15	20
2010-1	46	31	24	15	20
2010-2	41	27	8	16	13
2011-1	37	25	28	11	14
2011-2	44	29	31	11	18
2013-1	35	27	19	31	23
2013-2	70	60	30	40	60
2014-1	50	50	33	25	42
2014-2	58	36	25	28	25
2014-3	94	67	78	39	61
2015-1	47	53	47	37	37
2015-2	58	43	27	32	38
2015-R	87	87	35	55	58
2016-1	48	35	22	33	26
2016-2	58	48	42	43	42
2017-1	54	41	22	37	32
2017-2	63	48	30	56	56
2017-R	93	80	67	67	80
2018-1	65	52	39	57	52
2019-1	31	33	36	26	26
2019-2	57	38	63	32	46

Fuente: Elaboración propia.

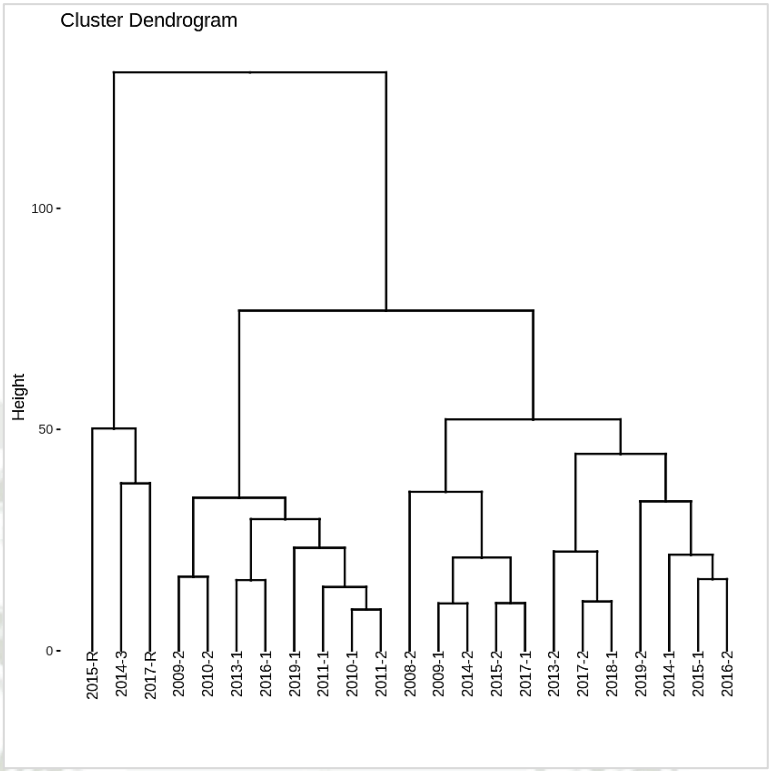

Figura 61: Dendograma para la clasificación del porcentaje de estudiantes aprobados en los diferentes periodos lectivos
Fuente: Elaboración propia.

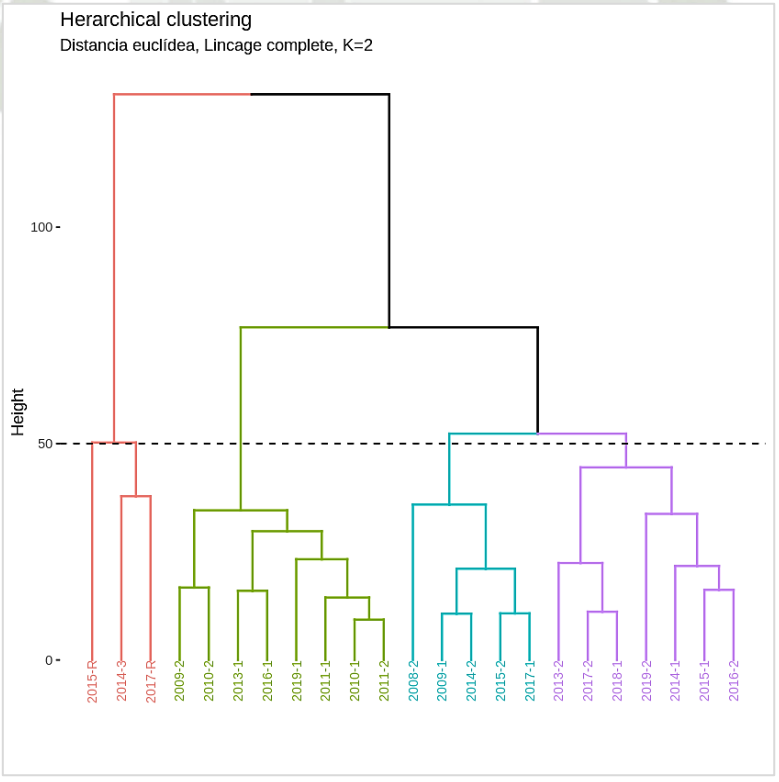

Figura 62: Clasificación a una distancia de 50, del porcentaje de estudiantes aprobados en los diferentes periodos lectivos
Fuente: Elaboración propia.

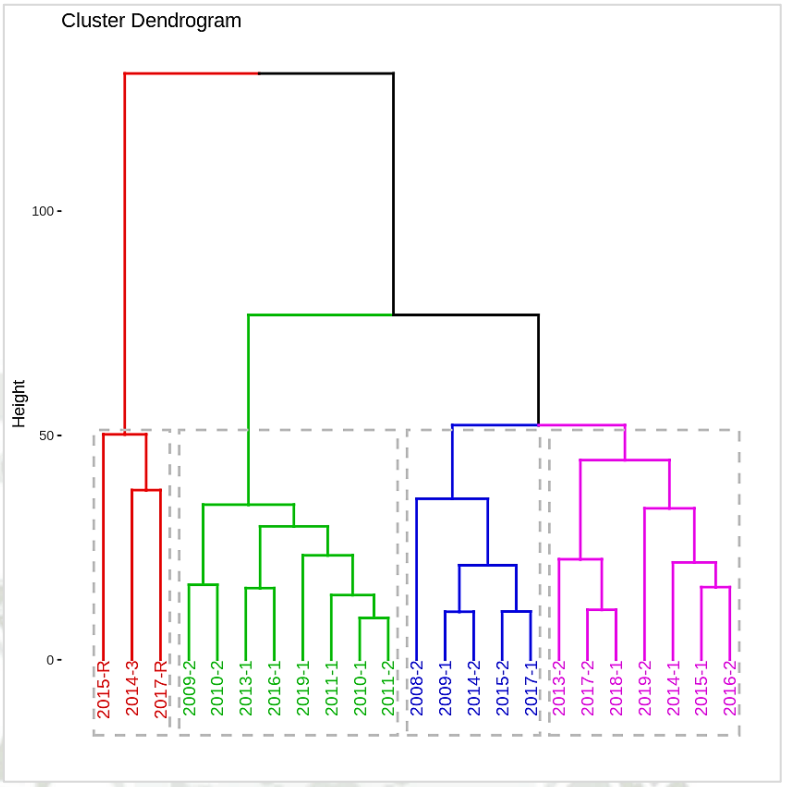

Figura 63: Grupos clasificados a una distancia de 50, del porcentaje de estudiantes aprobados en los diferentes periodos lectivos
Fuente: Elaboración propia.

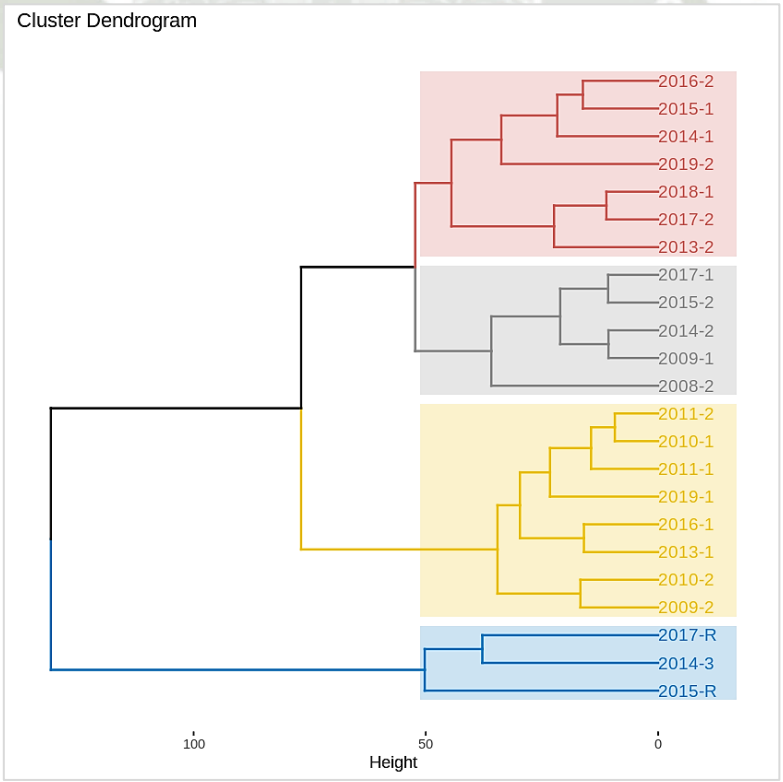

Figura 64: Cuatro grupos clasificados a una distancia de 50, porcentaje de estudiantes aprobados en los diferentes periodos lectivos
Fuente: Elaboración propia.

- **Grupo 1:** Periodos lectivos 2016-2, 2015-1, 2014-1, 2019-2, 2018-1, 2017-2, 2013-2
- **Grupo 2:** Periodos lectivos 2017-1, 2015-2, 2014-2, 2009-1, 2008-2
- **Grupo 3:** Periodos lectivos 2011-2, 2010-1, 2011-1, 2019-1, 2016-1, 2013-1, 2010-2, 2009-2
- **Grupo 4:** Periodos lectivos 2017-R, 2014-3, 2015-R

Tabla 29: Porcentajes de estudiantes aprobados por grupos

	Grupo1	Grupo 2	Grupo 3	Grupo 4
Permanente 1	59	63	39	92
Permanente 2	50	39	29	78
Examen parcial	40	23	22	60
Examen final	41	27	20	53
Promedio	48	31	20	66

Fuente: Elaboración propia.

- El grupo con mayor porcentaje de estudiantes aprobados corresponde al Grupo 4, donde en promedio el 92% de los estudiantes del grupo aprobaron la evaluación permanente 1. Este grupo corresponde a los cursos de verano del tercer semestre.
- En los periodos lectivos correspondientes al primer o segundo semestre el grupo con mayor porcentaje de estudiantes aprobados corresponde al Grupo 2, donde en promedio el 63% de los estudiantes del grupo aprobaron la evaluación permanente 1
- En los periodos lectivos correspondientes al primer o segundo semestre el grupo con menor porcentaje de estudiantes aprobados corresponde al Grupo 3, donde en promedio el 20% de los estudiantes del grupo aprobaron la evaluación final.
- En cada uno de los grupos el mayor porcentaje de estudiantes aprueba la evaluación permanente 1.

Los coeficientes de variación por grupo son los siguientes:

Tabla 30: Coeficientes de variación por grupos

	Grupo1	Grupo 2	Grupo 3	Grupo 4
CV	16.18%	43.44%	31.36%	22.11%

Fuente: Elaboración propia.

- El grupo más homogéneo en el porcentaje de aprobación de la asignatura es el Grupo1, dado que su $CV = 16.18\%$
- El grupo más heterogéneo en el porcentaje de aprobación de la asignatura es el Grupo2, dado que su $CV = 43.44\%$

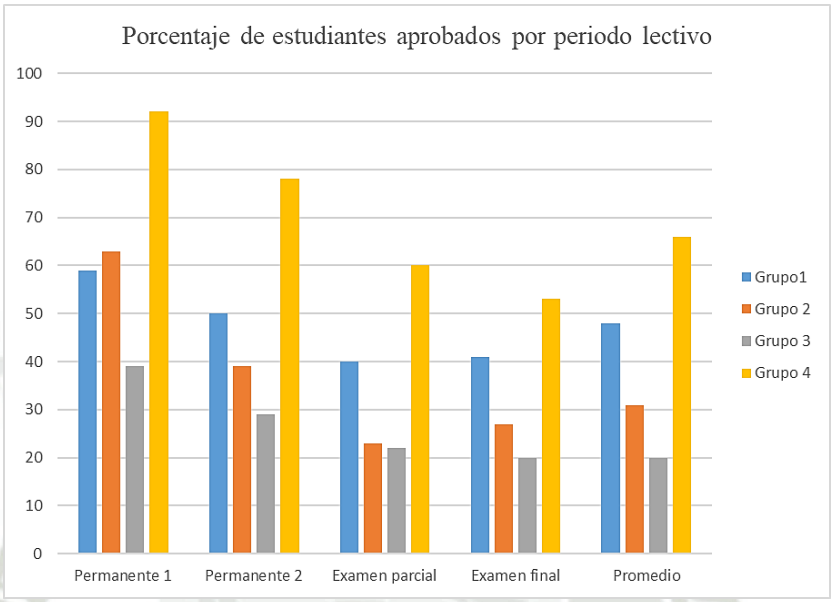

Figura 65: Porcentaje de aprobados por tipo de evaluación
Fuente: Elaboración propia

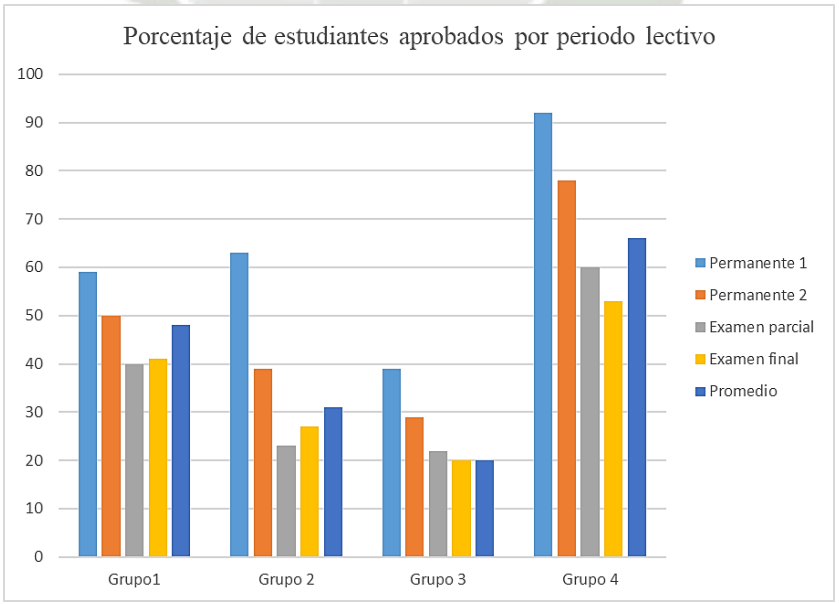

Figura 66: Porcentaje de aprobados por grupo
Fuente: Elaboración propia

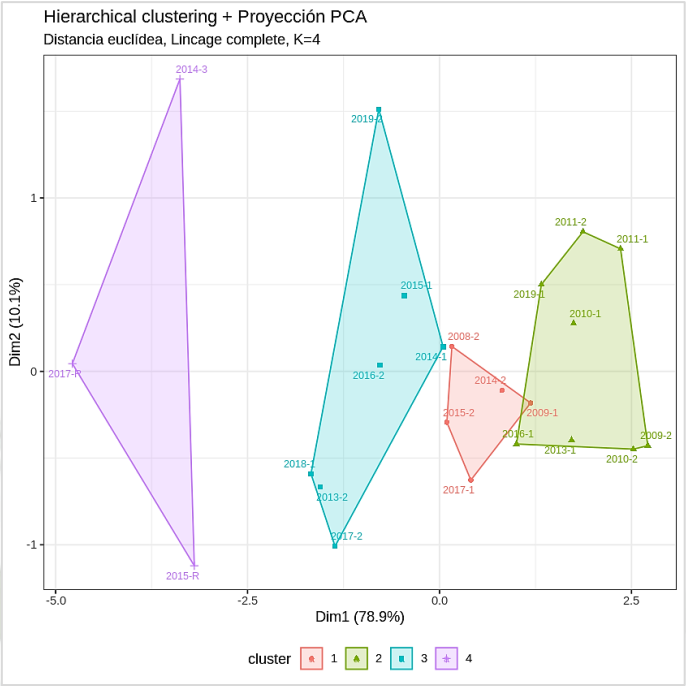

Figura 67: Clúster jerárquico del porcentaje de estudiantes aprobados en los diferentes periodos lectivos
Fuente: Elaboración propia.

3.3.2 Método de clasificación no jerárquico usando k-medias

La elección del número de clúster no se puede realizar al azar, por ello previamente utilizando la base de datos utilizando la técnica de análisis de componentes principales determinamos el número óptimo de clúster para posteriormente usar el método de k-medias para determinar los clústeres. Utilizando el lenguaje de programación R, obtenemos

```
K-means clustering with 4 clusters of sizes 6, 3, 7, 7
Cluster means:
   perm1     perm2   parcial    final promedio

1 60.0386  49.9682  41.6736  44.0214  48.6814
2 91.6248  77.9211  59.9761  53.4647  66.3918
3 37.1922  28.2877  22.0313  17.7764  19.0051
4 58.7007  39.6845  24.6472  27.4349  31.9519

Clustering vector:

2008-2 2009-1 2009-2 2010-1 2010-2 2011-1 2011-2 2013-1 2013-2 2014-1 2014-2
     4      4      3      3      3      3      3      3      1      4      4
2014-3 2015-1 2015-2 2015-R 2016-1 2016-2 2017-1 2017-2 2017-R 2018-1 2019-1
     2      1      4      2      4      1      4      1      2      1      3
2019-2
     1
```

```
Within cluster sum of squares by cluster:
[1] 2241.163 1879.129 1544.784 1727.469
(between_SS / total_SS = 77.6 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss"
[6] "betweenss" "size" "iter" "ifault"
```

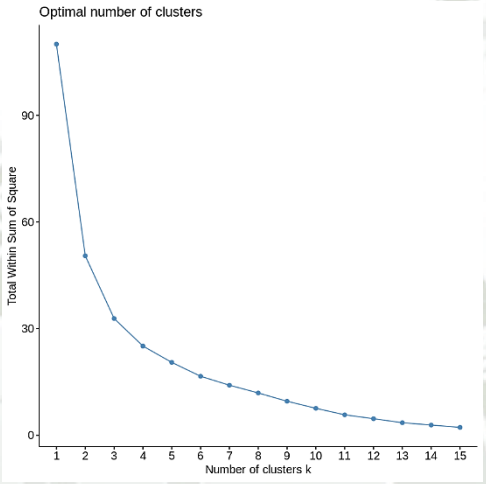

Figura 68: Grafica de sedimentación para el número de clúster óptimo
Fuente: Elaboración propia.

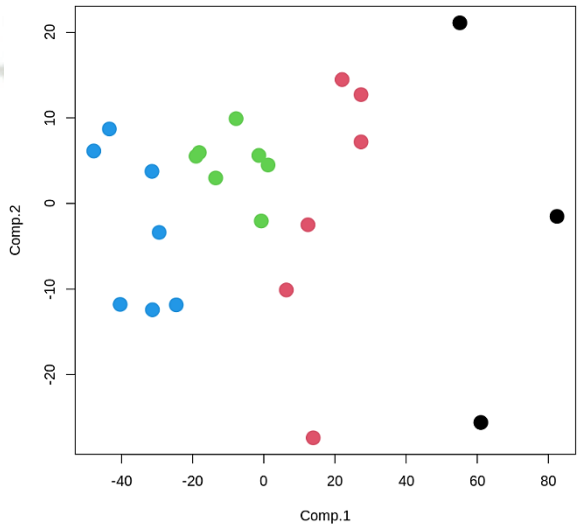

Figura 69: Conformación de cuatro clústeres para el porcentaje estudiantes aprobados en los diferentes periodos lectivos
Fuente: Elaboración propia.

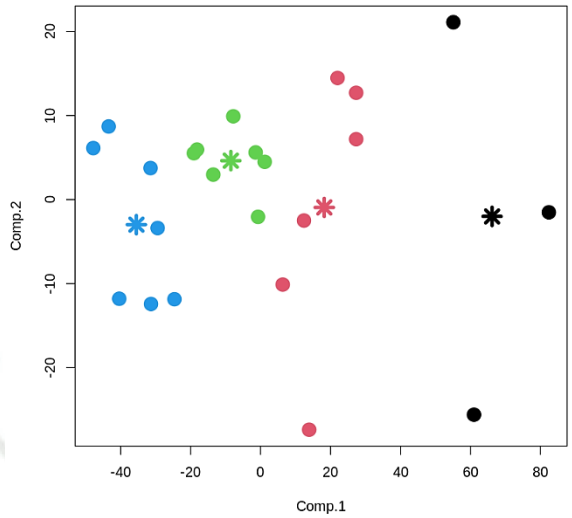

Figura 70: Porcentaje de estudiantes aprobados en los diferentes periodos lectivos
Fuente: Elaboración propia.

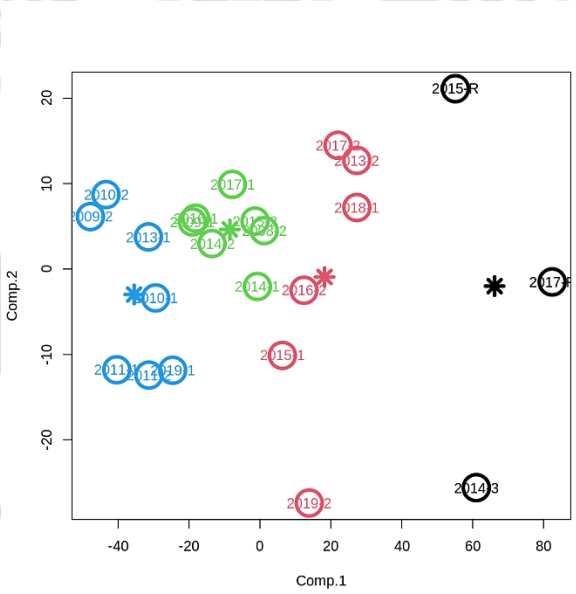

Figura 71: Centroides para los cuatro clústeres del porcentaje de estudiantes aprobados en los diferentes periodos lectivos
Fuente: Elaboración propia.

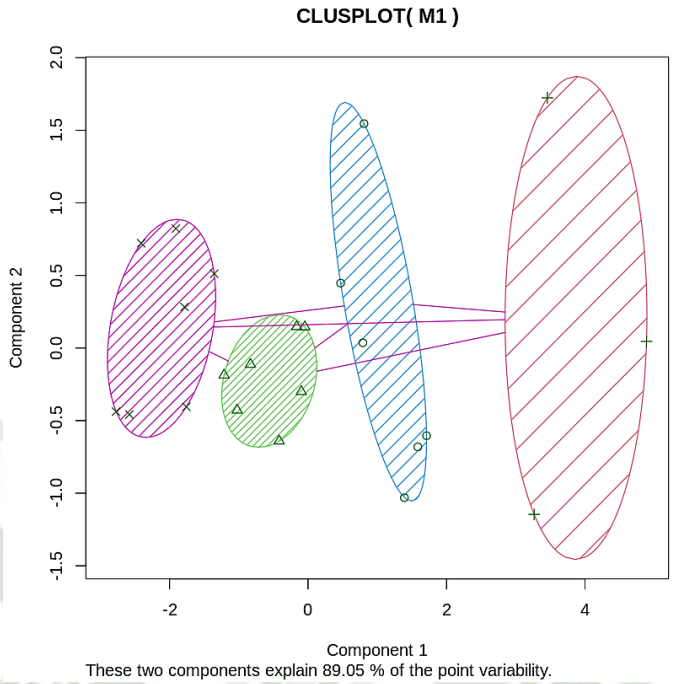

Figura 72: Porcentaje de estudiantes aprobados en los diferentes periodos lectivos
Fuente: Elaboración propia.

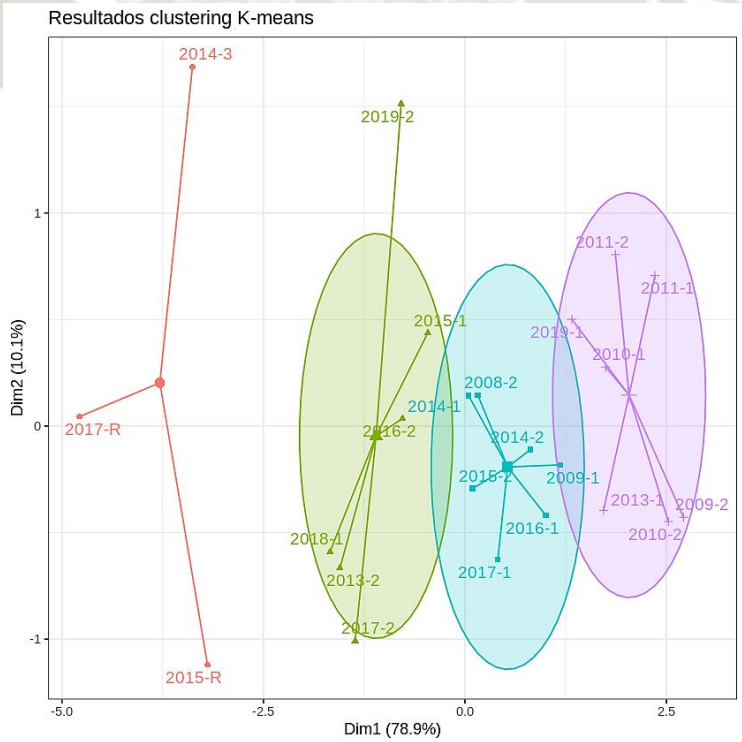

Figura 73: Gráfica de los cuatro grupos y sus centroides para el porcentaje de estudiantes aprobados en los diferentes periodos lectivos
Fuente: Elaboración propia.

- **Grupo 1:** Periodos lectivos 2019-2, 2015-1, 2016-2, 2018-1, 2013-2, 2017-2
- **Grupo 2:** Periodos lectivos 2008-2, 2014-1, 2014-2, 2015-2, 2009-1, 2016-1, 2017-1
- **Grupo 3:** Periodos lectivos 2011-2, 2011-1, 2019-1, 2010-1, 2013-1, 2009-2, 2010-2
- **Grupo 4:** Periodos lectivos 2017-R, 2014-3, 2015-R

Tabla 31: Porcentajes de estudiantes aprobados por grupos

	Grupo1	Grupo 2	Grupo 3	Grupo 4
Permanente 1	60	59	37	92
Permanente 2	50	40	28	78
Examen parcial	42	25	22	60
Examen final	44	27	18	53
Promedio	49	32	19	66

Fuente: Elaboración propia.

- El grupo con mayor porcentaje de estudiantes aprobados corresponde al Grupo 4, donde en promedio el 92% de los estudiantes del grupo aprobaron la evaluación permanente 1. Este grupo corresponde a los cursos de verano del tercer semestre.
- En los periodos lectivos correspondientes al primer o segundo semestre el grupo con mayor porcentaje de estudiantes aprobados corresponde al Grupo 1, donde en promedio el 60% de los estudiantes del grupo aprobaron la evaluación permanente 1
- En los periodos lectivos correspondientes al primer o segundo semestre el grupo con menor porcentaje de estudiantes aprobados corresponde al Grupo 3, donde en promedio el 18% de los estudiantes del grupo aprobaron la evaluación final.
- En cada uno de los grupos el mayor porcentaje de estudiantes aprueba la evaluación permanente 1.

Los coeficientes de variación por grupo son los siguientes:

Tabla 32: Coeficientes de variación por grupos

	Grupo1	Grupo 2	Grupo 3	Grupo 4
CV	14.29%	37.69%	31.67%	22.11%

Fuente: Elaboración propia.

- El grupo más homogéneo en el porcentaje de aprobación de la asignatura es el Grupo1, dado que su $CV = 14.29\%$

- El grupo más heterogéneo en el porcentaje de aprobación de la asignatura es el Grupo2, dado que su $CV = 37.69\%$

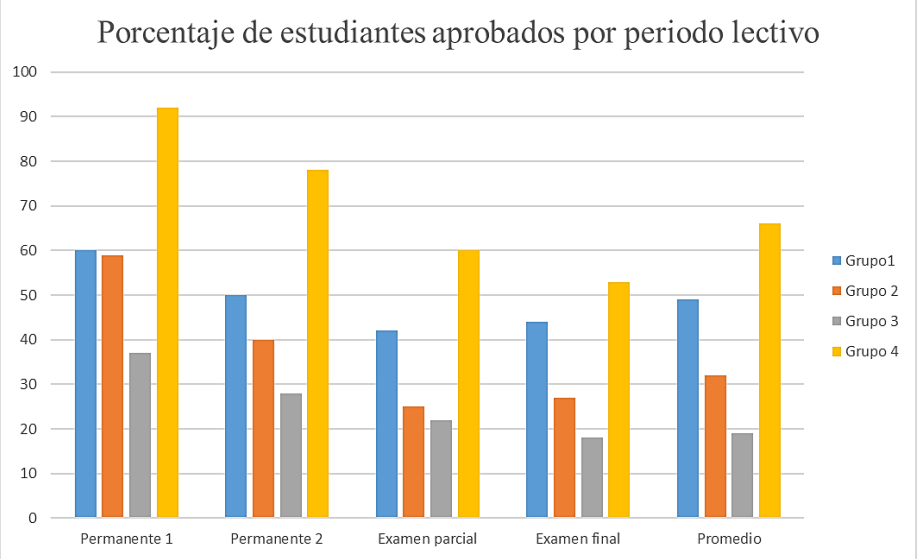

Figura 74: Porcentaje de aprobados por tipo de evaluación usando método de k-medias
Fuente: Elaboración propia

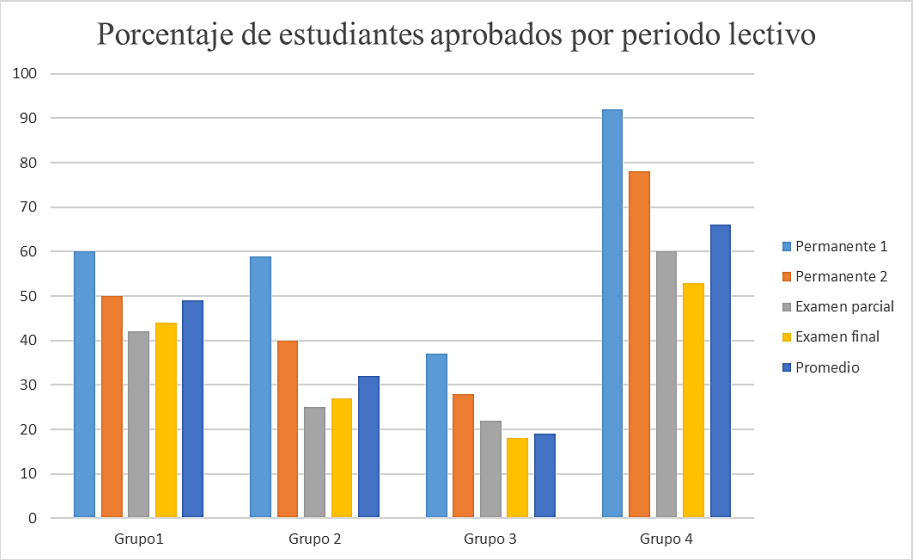

Figura 75: Porcentaje de aprobados por grupo usando método de k-medias
Fuente: Elaboración propia

CONCLUSIONES

i. Dadas las condiciones verificadas para el modelo logístico aplicado para contrastar las evaluaciones de *Permanente* 1 y *Permanente* 2 y su influencia sobre la condición de aprobar o desaprobar la asignatura motivo del estudio, se nota que la influencia es muy significativa y también se da validez al modelo ajustado, como propicio para determinar el rendimiento académico de los estudiantes de la asignatura de Algebra lineal y de Matemática II. Siendo así, para el modelo logístico determinado

$$p(x_1, x_2) = \frac{1}{1 + e^{-(-15.1000 + 0.4760x_1 + 0.7390x_2)}}$$

Según *Wald* los parámetros estimados por máxima verosimilitud $\beta_1 = 0.4760$; $\beta_2 = 0.7390$ para los modelos logísticos tienen una distribución normal cuando las muestras son grandes.

Si k es el número de variables independientes y n es el tamaño de muestra entonces, la muestra es grande si

$$n \geq 10(k + 1)$$

Por ejemplo, si existieran $k = 3$ variables independientes el tamaño de muestra mínimo debe ser $n = 40$, lo cual se verifica dado que como el tamaño de muestra tomado es $n = 938$ notas de estudiantes.

Siendo la hipótesis a verificar

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_1: para\ algun\ \beta_j \neq 0\ , \qquad j = 1{,}2$$

el estadístico de *Wald*

$$W^2 = \frac{\hat{\beta}_j^{\,2}}{\sigma^2(\hat{\beta}_j)}$$

verifica para $j = 1,2$ que su $valor\ p < 0.05$, por lo cual aceptamos la hipótesis alterna, las variables X_1: $Nota\ Permanente\ 1$ y X_2: $Nota\ Permanente\ 2$ influyen significativamente sobre la condición de aprobar la asignatura para el estudiante Y: condición del estudiante; $1 - aprobado$, $0 - desaprobado$. Con lo cual también se da la validez del modelo ajustado con un nivel de significancia del 5% ($p < 0.05$).

ii. En el hecho de tener una gran aproximación entre las soluciones obtenidas con la implementación del algoritmo que describe el modelo y los informes de respuestas de los softwares utilizados en el presente trabajo, se visualiza la incidencia que se da entre la matriz de evaluación y la metodología de enseñanza con el rendimiento académico de los estudiantes.

Los softwares estadísticos tienen implementados códigos que permiten estimar los parámetros de los modelos logísticos fundamentados en principios teóricos como la estimación de máxima verosimilitud e implementados en algoritmos, como por ejemplo el método de Newton. Para el caso de estudio revisado en este trabajo se describió el modelo logístico y construyó un algoritmo, al contrastar con el informe de respuestas del programa SPSS o del programa R, las soluciones son próximas, lo cual puede ser mejorado variando la tolerancia o cambiando por un método iterativo, por lo cual en este caso de estudio no es imprescindible su implementación.

El modelo logístico permite determinar la probabilidad de que el estudiante tenga la condición de aprobar o no aprobar en base a sus evaluaciones permanentes, así como la tasa de riesgo u oportunidad, mediante la razón $odss$.

iii. La incidencia que tiene la metodología de enseñanza y el uso de una matriz de evaluaciones en el rendimiento académico, se determinó a través de técnicas de clasificación.

iv. A través de un análisis exploratorio efectuado en las evaluaciones *permanentes*, se verifica con el empleo adecuado de técnicas de agrupación o clasificación que hay una gran influencia sobre las evaluaciones parciales y finales que a su vez determinarán en conjunto la condición de aprobación o desaprobación. Se determinó que para las evaluaciones previas a los exámenes parcial y final, existe una fuerte correlación entre el porcentaje de estudiantes aprobados en la notas permanentes, dado que $cor(perm1, perm2) = 0.8219$. Un gráfico de líneas permite visualizar que aproximadamente desde el periodo 2013-2, el porcentaje de estudiantes aprobados es superior al 20% en todas las evaluaciones permanentes y parciales, este hecho puede ser fundamentado con el uso del método de clasificación de $k - medias$, en el cual se determinó que dentro de los periodos lectivos correspondientes al primer o segundo semestre, el grupo con mayor porcentaje de estudiantes aprobados corresponde al Grupo 1, donde en promedio el 60% de los estudiantes del grupo aprobaron la evaluación permanente 1,

Grupo 1: Periodos lectivos 2019-2, 2015-1, 2016-2, 2018-1, 2013-2, 2017-2

El grupo más homogéneo en el porcentaje de aprobación de la asignatura es el Grupo1, dado que su $CV = 14.29\%$. Los resultados obtenidos de los periodos lectivos del 2013-2 en adelante concuerdan con el periodo de restructuración y cambios en la metodología de enseñanza y uso de matrices de evaluación, así como la implementación de programas de reforzamiento previos a las evaluaciones parciales en la Universidad Católica San Pablo, lo cual implica que los cambios dados fueron positivos para el rendimiento académico de los estudiantes.

RECOMENDACIONES

i. El modelo logístico descrito permite analizar el rendimiento académico de los estudiantes de la asignatura de Algebra lineal y Geometría de la Universidad Católica San Pablo (UCSP). Los parámetros pueden ser estimados mediante las librerías del CRAN del lenguaje de programación R, que es de distribución libre. Dado que verificamos la estimación de parámetros mediante un código en base al fundamento teórico, sería adecuado el usar directamente las librerías elaboradas con sus reportes más detallados y centrarse en el uso de las estimaciones para analizar el modelo logístico.

ii. Respecto a la asignatura de Algebra lineal y Geometría dado que se demostró que las variables X_1: $Nota\ Permanente\ 1$ y X_2: $Nota\ Permanente\ 2$ influyen significativamente en la condición de aprobar la asignatura para el estudiante, sería propicio comunicar a las instancias respectivas de la UCSP los resultados obtenidos, con el fin de innovar continuamente en las evaluaciones continuas, en cuanto a metodologías de enseñanza e innovaciones en las matrices de evaluación con el fin de elevar su rendimiento en sus evaluaciones parciales y finales. También puede sugerirse utilizar el modelo obtenido y aplicado en la asignatura motivo del presente trabajo en otras asignaturas.

iii. Revisar y reformular continuamente las matrices de evaluación de las notas permanentes, dado que dichas matrices determinan los criterios de evaluación. Solicitar a la UCSP proporcionar el acceso a algunos datos personales (de forma anónima) para considerar otras variables que influyan

en el rendimiento académico de los estudiantes, tales como variables socio familiares, hábitos de estudio, condiciones previas escolares, etc.

REFERENCIAS

Arévalo-Ascano, J. G., Pérez-González, S., Arévalo Ascanio, J. G., & Pérez González, S. (2018). El análisis de conglomerados como herramienta para evaluar el rendimiento académico: una experiencia en la universidad. *Revista Espacios*, *39*(46), 30–36. https://www.revistaespacios.com/a18v39n46/a18v39n46p30.pdf

Cabana, S. R., Cortés, F. H., Aguilera, M. I., & Vargas, F. A. (2018). Factores Determinantes para el Intraemprendimiento Social: El Caso de los Estudiantes de Ingeniería de la Universidad de La Serena, Chile TT - Determinant Factors for Social Intrapreneurism: The Case of the Engineering Students of the University of La Se. *Formación Universitaria*, *11*(2), 87–98. https://doi.org/10.4067/S0718-50062018000200087

Cáceres, R. Á. (2007). *Estadística aplicada a las ciencias de la salud*. Ediciones Díaz de Santos. https://www.ptonline.com/articles/how-to-get-better-mfi-results

Cantú Martínez, P. C., & Santoyo Stephano, M. A. (2019). Evaluación del rendimiento académico en bioestadística y la competencia disciplinar de pensamiento matemático en estudiantes universitarios. *Educación*, *28*(54), 45–60. https://doi.org/10.18800/educacion.201901.003

Enrique, M., & Pacheco, P. (2016). *Factores incidentes en el rendimiento académico estudiantil de Ingeniería en Sistemas de Información de la FAREM-Matagalpa , 2012-2016 Factors that affect students academic performance in Information*. 2012–2016.

Glonek, G. F. V., & McCullagh, P. (1995). Multivariate Logistic Models. In *Journal of the Royal Statistical Society: Series B (Methodological)* (Vol. 57, Issue 3, pp. 533–546). https://doi.org/10.1111/j.2517-6161.1995.tb02046.x

Levy Manquin, J. P., Varela Mallou, J., & A. G. (2008). *Análisis multivariable para las ciencias sociales*. Prentice-Hall.

McCullagh, P., & Nelder, J. A. (1989). Monographs on statistics and applied probability. In *Generalized linear models* (Vol. 37). Chapman & Hall.

Núñez, E., Steyerberg, E. W., & Núñez, J. (2011). Estrategias para la elaboración de modelos estadísticos de regresión. *Revista Espanola de Cardiologia*, *64*(6), 501–507. https://doi.org/10.1016/j.recesp.2011.01.019

Remesal, A. F., López, B. G., & Rodríguez, J. S. (2007). *Estrategias de Aprendizaje y*

Rendimiento Academico En Estudiantes Universitarios '.

Roux, R., & Anzures González, E. E. (2015). Estrategias de aprendizaje y su relación con el rendimiento académico en estudiantes de una escuela privada de educación media superior / Learning strategies and their relationship with academic achievement in students of a private high school. *Actualidades Investigativas En Educación*, *15*(1). https://doi.org/10.15517/aie.v15i1.17731

Tejedor, F. (2003). Poder explicativo de algunos determinantes del rendimiento en los estudios universitarios. *Enero-Abril*, *224*, 5–32. https://doi.org/10.1515/JPM.2010.021

Trigueros Ramos, R., & Navarro Gómez, N. (2019). The influence of the teacher on the motivation, learning strategies, critical thinking and academic performance of high school students in Physical Education. *Psychology, Society and Education*, *11*(1), 137–150. https://doi.org/10.25115/psye.v10i1.2230

Véliz Capuñay, C. (2017). *Análisis multivariante. Métodos estadísticos multivariantes para la investigación*. México: Cengage Learning.

Yolvi, O. F. (2011). Variables Académicas Que Influyen En El Rendimiento Académico De Los Estudiantes Universitarios. *Investigación Educativa*, *15*(27), 165–180.

ANEXOS

Los registros proporcionados por la Universidad Católica San Pablo, fueron otorgados de forma anónima, con fines de investigación y siendo un total de 5628 registros y solamente se exhiben en anexo algunos de los registros.

	PERIODO	CURSO	GRUPO	EVALUACION	NOTA
1	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	15.7500
2	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	9.3700
3	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	11.6100
4	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	8.7500
5	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	8.9100
6	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	8.5200
7	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	6.3000
8	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	7.6400
9	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	10.5600
10	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	8.9600
11	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	12.3900
12	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	11.5200
13	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	17.0000
14	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	12.2200
15	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	14.1200
16	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	12.1600
17	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	11.1900
18	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	10.5800
19	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	9.8300
20	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	11.7900
21	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	11.1100
22	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	12.7100
23	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	14.6200
24	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente	14.5800
25	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	17.5000
26	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	13.2000
27	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	15.0000
28	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	13.4000
29	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	11.3000
30	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	12.2500
31	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	10.2500
32	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	13.2500
33	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	11.8000
34	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	12.5000
35	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	11.7500
36	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 1	16.0000
37	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	16.5000
38	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	11.2500
39	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	13.2500
40	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	10.9200

41	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	11.0800
42	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	8.9200
43	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	9.4200
44	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	10.3300
45	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	10.4200
46	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	12.9200
47	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	17.5000
48	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Evaluación Permanente 2	13.1700
49	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	16.5000
50	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	8.0000
51	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	11.2500
52	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	9.0000
53	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	9.5000
54	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	11.5000
55	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	6.5000
56	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	6.0000
57	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	9.7500
58	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	9.5000
59	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	7.0000
60	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Parcial	11.5000
61	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	14.2500
62	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	8.2500
63	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	10.0000
64	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	6.0000
65	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	6.7500
66	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	4.7500
67	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	3.5000
68	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	5.7500
69	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	10.7500
70	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	5.7500
71	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	14.7500
72	2008-02	Álgebra Lineal y Geometría Analítica	IND1-7	Examen Final	9.2500
73	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	13.9600
74	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	7.1300
75	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	11.7700
76	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	12.9500
77	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	11.7500
78	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	13.8100
79	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	4.1500
80	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	8.9100
81	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	9.1600
82	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	12.1200
83	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	11.0700
84	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	11.0400
85	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	10.7100
86	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	3.9400
87	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	4.8800
88	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	3.5600
89	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	3.1800
90	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	8.3500
91	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	2.9800
92	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	11.8000
93	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	4.9100
94	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	14.6300
95	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	8.9100
96	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	8.5000
97	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	11.2800
98	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	3.7000
99	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	10.6500
100	2009-01	Álgebra Lineal y Geometría Analítica	IND1-7	Promedio	0.0000

5512	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	13.5500
5513	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	13.7300
5514	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	13.0800
5515	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	10.8300
5516	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	12.2600
5517	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	12.8300
5518	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	12.2400
5519	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	10.4900
5520	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	8.2800
5521	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	14.6700
5522	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	13.8100
5523	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	15.0300
5524	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	1.3000
5525	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	12.1700
5526	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	1.3000
5527	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	10.2600
5528	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	9.8700
5529	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	12.8600
5530	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	10.3500
5531	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	9.5300
5532	2019-02	Matemática II	IND2-8	Evaluación Permanente 1	7.2500
5533	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	5.1100
5534	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	15.5700
5535	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	12.0500
5536	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	11.3300
5537	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	13.4200
5538	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	9.5700
5539	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	17.8300
5540	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	12.1700
5541	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	11.9400
5542	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	11.4200
5543	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	11.4900
5544	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	13.2100
5545	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	11.5800
5546	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	12.4700
5547	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	11.3800
5548	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	13.5300
5549	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	14.5700
5550	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	12.3000
5551	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	12.8800
5552	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	0.0000
5553	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	11.2900
5554	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	9.5600
5555	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	11.2800
5556	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	5.8500
5557	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	11.3900
5558	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	7.7200
5559	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	4.1200
5560	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	10.3300
5561	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	10.0500
5562	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	14.0900
5563	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	10.3800
5564	2019-02	Matemática II	IND2-8	Evaluación Permanente 2	3.9300
5565	2019-02	Matemática II	IND2-8	Examen Parcial	8.0000
5566	2019-02	Matemática II	IND2-8	Examen Parcial	16.5000
5567	2019-02	Matemática II	IND2-8	Examen Parcial	16.5000
5568	2019-02	Matemática II	IND2-8	Examen Parcial	16.2500
5569	2019-02	Matemática II	IND2-8	Examen Parcial	11.7500

5570	2019-02	Matemática II	IND2-8	Examen Parcial	9.2500
5571	2019-02	Matemática II	IND2-8	Examen Parcial	12.5000
5572	2019-02	Matemática II	IND2-8	Examen Parcial	8.5000
5573	2019-02	Matemática II	IND2-8	Examen Parcial	9.0000
5574	2019-02	Matemática II	IND2-8	Examen Parcial	13.7500
5575	2019-02	Matemática II	IND2-8	Examen Parcial	12.0000
5576	2019-02	Matemática II	IND2-8	Examen Parcial	10.5000
5577	2019-02	Matemática II	IND2-8	Examen Parcial	10.7500
5578	2019-02	Matemática II	IND2-8	Examen Parcial	13.0000
5579	2019-02	Matemática II	IND2-8	Examen Parcial	8.0000
5580	2019-02	Matemática II	IND2-8	Examen Parcial	9.0000
5581	2019-02	Matemática II	IND2-8	Examen Parcial	13.2500
5582	2019-02	Matemática II	IND2-8	Examen Parcial	7.2500
5583	2019-02	Matemática II	IND2-8	Examen Parcial	6.7500
5584	2019-02	Matemática II	IND2-8	Examen Parcial	0.0000
5585	2019-02	Matemática II	IND2-8	Examen Parcial	9.0000
5586	2019-02	Matemática II	IND2-8	Examen Parcial	14.2500
5587	2019-02	Matemática II	IND2-8	Examen Parcial	14.2500
5588	2019-02	Matemática II	IND2-8	Examen Parcial	9.0000
5589	2019-02	Matemática II	IND2-8	Examen Parcial	10.7500
5590	2019-02	Matemática II	IND2-8	Examen Parcial	9.2500
5591	2019-02	Matemática II	IND2-8	Examen Parcial	8.5000
5592	2019-02	Matemática II	IND2-8	Examen Parcial	9.7500
5593	2019-02	Matemática II	IND2-8	Examen Parcial	13.5000
5594	2019-02	Matemática II	IND2-8	Examen Parcial	11.5000
5595	2019-02	Matemática II	IND2-8	Examen Parcial	13.0000
5596	2019-02	Matemática II	IND2-8	Examen Parcial	5.2500
5597	2019-02	Matemática II	IND2-8	Examen Final	0.0000
5598	2019-02	Matemática II	IND2-8	Examen Final	16.0000
5599	2019-02	Matemática II	IND2-8	Examen Final	15.2500
5600	2019-02	Matemática II	IND2-8	Examen Final	15.7500
5601	2019-02	Matemática II	IND2-8	Examen Final	13.7500
5602	2019-02	Matemática II	IND2-8	Examen Final	8.2500
5603	2019-02	Matemática II	IND2-8	Examen Final	16.2500
5604	2019-02	Matemática II	IND2-8	Examen Final	8.0000
5605	2019-02	Matemática II	IND2-8	Examen Final	10.7500
5606	2019-02	Matemática II	IND2-8	Examen Final	8.5000
5607	2019-02	Matemática II	IND2-8	Examen Final	6.2500
5608	2019-02	Matemática II	IND2-8	Examen Final	11.5000
5609	2019-02	Matemática II	IND2-8	Examen Final	10.7500
5610	2019-02	Matemática II	IND2-8	Examen Final	8.5000
5611	2019-02	Matemática II	IND2-8	Examen Final	13.5000
5612	2019-02	Matemática II	IND2-8	Examen Final	16.2500
5613	2019-02	Matemática II	IND2-8	Examen Final	8.7500
5614	2019-02	Matemática II	IND2-8	Examen Final	14.7500
5615	2019-02	Matemática II	IND2-8	Examen Final	9.0000
5616	2019-02	Matemática II	IND2-8	Examen Final	0.0000
5617	2019-02	Matemática II	IND2-8	Examen Final	14.0000
5618	2019-02	Matemática II	IND2-8	Examen Final	8.5000
5619	2019-02	Matemática II	IND2-8	Examen Final	12.0000
5620	2019-02	Matemática II	IND2-8	Examen Final	0.0000
5621	2019-02	Matemática II	IND2-8	Examen Final	8.0000
5622	2019-02	Matemática II	IND2-8	Examen Final	6.5000
5623	2019-02	Matemática II	IND2-8	Examen Final	6.5000
5624	2019-02	Matemática II	IND2-8	Examen Final	7.0000
5625	2019-02	Matemática II	IND2-8	Examen Final	14.5000
5626	2019-02	Matemática II	IND2-8	Examen Final	12.7500
5627	2019-02	Matemática II	IND2-8	Examen Final	7.7500
5628	2019-02	Matemática II	IND2-8	Examen Final	5.0000