Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado
Facultad de Ciencias Matemáticas
Unidad de Posgrado

Uso de los métodos multivariantes para el análisis del desempeño académico de los estudiantes de la educación superior. (Caso: estudiantes ingresantes en el primer curso de Matemática de la UNALM)

TESIS

Para optar el Grado Académico de Magíster en Estadística Matemática

AUTOR

Rocío Consuelo DELGADO AGUILAR

ASESOR

Antonio BRAVO QUIROZ

Lima, Perú

2020



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

https://creativecommons.org/licenses/by-nc-sa/4.0/

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Delgado, R. (2020). *Uso de los métodos multivariantes para el análisis del desempeño académico de los estudiantes de la educación superior. (Caso: estudiantes ingresantes en el primer curso de Matemática de la UNALM)*. Tesis para optar el grado de Magíster en Estadística Matemática. Unidad de Posgrado, Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos, Lima, Perú.

HOJA DE METADATOS COMPLEMENTARIOS

Código ORCID del autor	—
DNI o pasaporte del autor	09363606
Código ORCID del asesor	—
DNI o pasaporte del asesor	10130035
Grupo de investigación	—
Agencia financiadora	—
Ubicación geográfica donde se desarrolló la investigación	Perú, Lima, Lima, La Molina Longitud: O76°57'2.633" Latitud: S12°4'47.716"
Disciplinas OCDE	Estadísticas, Probabilidad http://purl.org/pe-repo/ocde/ford#1.01.03

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú, DECANA DE AMÉRICA

FACULTAD DE CIENCIAS MATEMÁTICAS
VICEDECANATO DE INVESTIGACIÓN Y POSGRADO
UNIDAD DE POSGRADO

ACTA DE SUSTENTACIÓN VIRTUAL DE TESIS PARA OPTAR EL GRADO ACADÉMICO DE MAGÍSTER EN ESTADÍSTICA MATEMÁTICA

Siendo las, 14:00 horas del día martes primero de setiembre del dos mil veinte, en la sala virtual https://meet.google.com/vzc-ptam-uxv el Jurado de Tesis conformado por los siguientes docentes:

PRESIDENTE:	Dr. Roger Pedro Norabuena Figueroa
MIEMBRO:	Mg. Rosa Ysabel Adriazola Cruz
MIEMBRO:	Mg. José Antonio Cárdenas Garro
ASESOR:	Mg. Antonio Bravo Quiroz

se reunieron para la sustentación de la tesis titulada: USO DE LOS MÉTODOS MULTIVARIANTES PARA EL ANÁLISIS DEL DESEMPEÑO DE LOS ESTUDIANTES DE LA EDUCACIÓN SUPERIOR (CASO: ESTUDIANTES INGRESANTES EN EL PRIMER CURSO DE MATEMÁTICA DE LA UNALM)» presentada por la Bachiller Rocío Consuelo Delgado Aguilar para optar el Grado Académico de Magíster en Estadística Matemática.

Concluida la exposición, los miembros del Jurado de Tesis procedieron a formular sus preguntas que fueron absueltas por la graduanda; acto seguido se procedió a la evaluación correspondiente, según tabla adjunta, habiendo obtenido la Bachiller Rocío Consuelo Delgado Aguilar el calificativo de BUENO (15).

Habiendo sido aprobada la sustentación de la Tesis, el Jurado Evaluador recomienda para que el Consejo de Facultad apruebe el otorgamiento del Grado Académico de **Magíster en Estadística Matemática** a la **Bachiller Rocío Consuelo Delgado Aguilar**.

Siendo las 15:30 horas, se levantó la sesión, firmando para constancia la presente Acta:

Mg. Rosa Ysabel Adriazola Cruz
MIEMBRO

Dr. Roger Pedro Norabuena Figueroa
PRESIDENTE

Mg. José Antonio Cárdenas Garro
MIEMBRO

Mg. Antonio Bravo Quiroz
ASESOR

Resumen

El propósito del presente estudio fue comparar tres modelos de clasificación: Modelo logístico binario, el análisis discriminante y redes neuronales tipo perceptrón multicapa para evaluar el desempeño de estudiantes en primer ciclo académico de la universidad. El desempeño académico fue medido desde dos aspectos: el promedio final en el curso de matemática y la cantidad de créditos aprobados al finalizar el ciclo académico.

Con este fin, se aplicaron tres técnicas estadísticas sobre un total de 553 estudiantes universitarios correspondientes a los ciclos académicos 2017 I y 2017 II. Se trabajó con dos grupos de datos, de entrenamiento (70%) y de validación (30%). Los modelos propuestos se compararon mediante los indicadores: sensibilidad, especificidad y curva ROC.

Concluyendo que, los modelos de regresión logística binaria clasificaron mejor el desempeño académico de los estudiantes en el ciclo 2017 I y 2017 II a excepción del resultado en el ciclo 2017 II que se clasificó mejor con el modelo de redes neuronales tipo perceptrón multicapa. La nota promedio en el área de matemática del examen de admisión de la universidad fue la más importante para la clasificación del desempeño académico de los estudiantes de primer ciclo de la universidad.

Palabras clave: Desempeño académico, Regresión Logística Binaria, Análisis Discriminante, Redes Neuronales tipo Perceptrón Multicapa.

Abstract

The purpose of the present study was to compare three classification models: Binary logistic model, discriminant analysis and multilayer perceptron neural networks to evaluate the performance of students in the first academic cycle of the university. Academic performance was measured from two aspects: the final average in the mathematics course and the number of approved credits at the end of the academic cycle.

For this reason, three statistical techniques were applied to a total of 553 university students corresponding to the 2017 I and 2017 II academic cycles. Two data groups were used, training (70%) and validation (30%). The proposed models were compared using the indicators: sensitivity, specificity and ROC curve.

Therefore, this study concludes that the binary logistic regression models better classified the academic performance of the students in the 2017 I and 2017 II cycles, except for the result in the 2017 II cycle, which was better classified with the multilayer perceptron-type neural network model. The average grade in the mathematics area of the university entrance exam was the most important for the classification of the academic performance of the first cycle students of the university.

Keywords: Academic performance, Binary Logistic Regression, Discriminant Analysis, Multilayer Perceptron Neural Networks.

ÍNDICE GENERAL

Capítulo 1: Introducción 1

1.1 Situación Problemática 3

1.2 Formulación del Problema 6

1.3 Justificación de la Investigación 8

1.4 Objetivos de la Investigación 9

Capítulo 2: Marco Teórico 10

2.1 Antecedentes del investigación 10

2.2 Bases Teóricas 16

2.2.1 Educación matemática 16

2.2.2 Rendimiento académico 17

2.2.3 Factores asociados al rendimiento 18

2.2.4 Regresión Logística Binaria 22

2.2.5 Análisis Discriminante 30

2.2.6 Redes Neuronales 37

2.2.7 Evaluación de la clasificación 54

Capítulo 3: Metodología 59

3.1 Hipótesis general 59

3.2 Hipótesis específicas 59

3.3 Identificación de variables 59

3.4 Tipo y diseño de investigación 63

3.5 Unidad de análisis 63

3.6 Población de estudio 64

3.7 Técnicas de recolección de datos 65

3.8 Análisis e interpretación de la información 66

Capítulo 4: Resultados y Discusión 67

4.1 Estimación de Modelos de Regresión Logística Binaria 69

4.1.1 Modelo de Regresión Logística Binaria para Resultado 2017 I 70

4.1.2 Modelo de Regresión Logística Binaria para Resultado 2017 II 73

4.1.3 Modelo de Regresión Logística Binaria para Créditos Aprobados 2017 I 76

4.1.4 Modelo de Regresión Logística Binaria para Créditos Aprobados 2017 II 79

4.2 Estimación de Modelos Discriminante 82

4.2.1 Cumplimiento de supuestos para Resultado 2017 I y 2017 II 82

4.2.2 Modelo Discriminante para Resultado 2017 I 88

4.2.3 Modelo Discriminante para Resultado 2017 II 94

4.2.4 Cumplimiento de supuestos para Créditos Aprobados 2017 I y 2017 II 100

4.2.5 Modelo Discriminante para Créditos Aprobados 2017 I 105

4.2.6 Modelo Discriminante para Créditos Aprobados 2017 II 111

4.3 Modelos de Redes Neuronales tipo Perceptrón Multicapa 117

4.3.1 Modelo de Redes Neuronales para Resultado 2017 I 117

4.3.2 Modelo de Redes Neuronales para Resultado 2017 II 120

4.3.3 Modelo de Redes Neuronales para Créditos Aprobados 2017 I 123

4.3.4 Modelo de Redes Neuronales para Créditos Aprobados 2017 II 126

4.4 Técnicas de evaluación de clasificadores 129

CONCLUSIONES 132

RECOMENDACIONES 135

REFERENCIAS BIBLIOGRÁFICAS 136

APÉNDICES 147

Índice de Tablas

Tabla 1. Tasa de Desaprobación del Curso Matemática Básica 3

Tabla 2. Estudiantes ingresantes matriculados según cantidad de cursos aprobados en los ciclos 2017 I y 2017 II 4

Tabla 3. Estudiantes ingresantes según cantidad de cursos desaprobados en los ciclos 2017 I y 2017 II 5

Tabla 4. Resultados a Nivel Nacional de la ECE 2016 – Segundo Año de Secundaria 12

Tabla 5. Frecuencias esperadas y observadas para C_g 28

Tabla 6. Matriz de Contingencia 54

Tabla 7. Tabla de Clasificación. Área bajo la curva ROC 58

Tabla 8. Descripción de Variables 60

Tabla 9. Resumen de la Información 62

Tabla 10. Tipo de variables intervinientes en el estudio 68

Tabla 11. Variables en el modelo logístico para Resultado. Ciclo 2017 I 70

Tabla 12. Prueba Ómnibus del Modelo Logístico para Resultado. Ciclo 2017 I 71

Tabla 13. Prueba Hosmer y Lemeshow. Modelo Logístico para Resultado. Ciclo 2017 I 71

Tabla 14. Resultados de clasificación para Resultado. Ciclo 2017 I 72

Tabla 15. Variables en el modelo logístico para Resultado. Ciclo 2017 II 73

Tabla 16. Prueba Ómnibus del Modelo Logístico para Resultado. Ciclo 2017 II 74

Tabla 17. Prueba Hosmer y Lemeshow. Modelo Logístico para Resultado. Ciclo 2017 II 74

Tabla 18. Resultados de clasificación para Resultado. Ciclo 2017 II 75

Tabla 19. Variables en el modelo logístico para Créditos Aprobados. Ciclo 2017 I 76

Tabla 20. Prueba Ómnibus del Modelo Logístico para Créditos Aprobados. Ciclo 2017 I 77

Tabla 21. Prueba Hosmer y Lemeshow. Modelo Logístico para Créditos Aprobados. Ciclo 2017 I 78

Tabla 22. Resultados de clasificación para Créditos Aprobados. Ciclo 2017 I 78

Tabla 23. Variables en el modelo logístico para Créditos Aprobados. Ciclo 2017 II 79

Tabla 24. Prueba Ómnibus del Modelo Logístico para Créditos Aprobados. Ciclo 2017 II 80

Tabla 25. Prueba Hosmer y Lemeshow. Modelo Logístico para Créditos Aprobados. Ciclo 2017 II 80
Tabla 26. Resultados de clasificación para Créditos Aprobados. Ciclo 2017 II 81
Tabla 27. Prueba de Normalidad para Resultado. Ciclos 2017 I y 2017 II 82
Tabla 28. Comparación de medias y varianza para Resultado. Ciclo 2017 I 84
Tabla 29. Comparación de medias y varianza para Resultado. Ciclo 2017 II 84
Tabla 30. Prueba M de Box para Resultado. Ciclos 2017 I y 2017 II 85
Tabla 31. Prueba de Multicolinealidad para Resultado. Ciclos 2017 I y 2017 II 86
Tabla 32. Media de variables métricas según Resultado y ciclo 86
Tabla 33. Tamaño de muestra para el análisis discriminante 87
Tabla 34. Prueba de igualdad de medias para Resultado. Ciclo 2017 I 88
Tabla 35. Prueba M de Box para Resultado. Ciclo 2017 I 89
Tabla 36. Variables incluidas en el análisis para Resultado. Ciclo 2017 I 90
Tabla 37. Autovalores para Resultado. Ciclo 2017 I 90
Tabla 38. Prueba Lambda de Wilks para Resultado. Ciclo 2017 I 91
Tabla 39. Coeficientes y Matriz de estructuras para Resultado. Ciclo 2017 I 91
Tabla 40. Coeficientes no estandarizados discriminante de Resultado. Ciclo 2017 I 92
Tabla 41. Coeficientes de función de clasificación para Resultado. Ciclo 2017 I 92
Tabla 42. Resultados de clasificación discriminante para Resultado. Ciclo 2017 I 93
Tabla 43. Prueba de igualdad de medias para Resultado. Ciclo 2017 II 94
Tabla 44. Prueba M de Box para Resultado. Ciclo 2017 II 95
Tabla 45. Variables incluidas en el análisis para Resultado. Ciclo 2017 II 95
Tabla 46. Autovalores para Resultado. Ciclo 2017 II 96
Tabla 47. Prueba Lambda de Wilks para Resultado. Ciclo 2017 II 96
Tabla 48. Coeficientes y Matriz de estructuras para Resultado. Ciclo 2017 II 97
Tabla 49. Coeficientes no estandarizados discriminante de Resultado. Ciclo 2017 II 97
Tabla 50. Coeficientes de función de clasificación para Resultado. Ciclo 2017 II 98
Tabla 51. Resultados de clasificación discriminante para Resultado. Ciclo 2017 II 99
Tabla 52. Prueba de Normalidad. Créditos Aprobados. Ciclos 2017 I y 2017 II 100
Tabla 53. Comparación de medias y varianza para Créditos Aprobados. Ciclo 2017 I 101
Tabla 54. Comparación de medias y varianza para Créditos Aprobados. Ciclo 2017 II 102

Tabla 55. Prueba M de Box para Créditos Aprobados. Ciclos 2017 I y 2017 II 103

Tabla 56. Prueba de Multicolinealidad para Créditos Aprobados. Ciclos 2017 I y 2017 II 103

Tabla 57. Media de variables métricas según Créditos Aprobados y ciclo 104

Tabla 58. Prueba de igualdad de medias para Créditos Aprobados. Ciclo 2017 I 105

Tabla 59. Prueba M de Box para Créditos Aprobados. Ciclo 2017 I 106

Tabla 60. Variables incluidas en el análisis para Créditos Aprobados. Ciclo 2017 I 106

Tabla 61. Autovalores para Créditos Aprobados. Ciclo 2017 I 107

Tabla 62. Prueba Lambda de Wilks para Créditos Aprobados. Ciclo 2017 I 107

Tabla 63. Coeficientes y Matriz de estructuras para Créditos Aprobados. Ciclo 2017 I 108

Tabla 64. Coeficientes no estandarizados discriminante para Créditos Aprobados. Ciclo 2017 I 108

Tabla 65. Coeficientes de función de clasificación para Créditos Aprobados. Ciclo 2017 I 109

Tabla 66. Resultados de clasificación discriminante para Créditos Aprobados. Ciclo 2017 I 110

Tabla 67. Prueba de igualdad de medias para Créditos Aprobados. Ciclo 2017 II 111

Tabla 68. Prueba M de Box para Créditos Aprobados. Ciclo 2017 II 112

Tabla 69. Variables incluidas en el análisis para Créditos Aprobados. Ciclo 2017 II 112

Tabla 70. Autovalores para Créditos Aprobados. Ciclo 2017 II 113

Tabla 71. Prueba Lambda de Wilks para Créditos Aprobados. Ciclo 2017 II 113

Tabla 72. Coeficientes y Matriz de estructuras para Créditos Aprobados. Ciclo 2017 II 114

Tabla 73. Coeficientes no estandarizados discriminante para Créditos Aprobados. Ciclo 2017 II 114

Tabla 74. Coeficientes de función de clasificación para Créditos Aprobados. Ciclo 2017 II 115

Tabla 75. Resultados de clasificación discriminante para Créditos Aprobados. Ciclo 2017 II 116

Tabla 76. Modelo de redes neuronales para Resultado. Ciclo 2017 I 117

Tabla 77. Clasificación con redes neuronales para Resultado. Ciclo 2017 I 119

Tabla 78. Modelo de redes neuronales para Resultado. Ciclo 2017 II 120

Tabla 79. Clasificación con redes neuronales para Resultado. Ciclo 2017 II 122

Tabla 80. Modelo de redes neuronales para Créditos Aprobados. Ciclo 2017 I 123

Tabla 81. Clasificación con redes neuronales para Créditos Aprobados. Ciclo 2017 I 125

Tabla 82. Modelo de redes neuronales para Créditos Aprobados. Ciclo 2017 II 126

Tabla 83. Clasificación con redes neuronales para Créditos Aprobados. Ciclo 2017 II 128

Tabla 84. Comparativo para Resultado en el primer curso de matemática. Ciclo 2017 I 129

Tabla 85. Comparativo para Resultado en el primer curso de matemática. Ciclo 2017 II 130

Tabla 86. Comparativo para Créditos Aprobados. Ciclo 2017 I 130

Tabla 87. Comparativo para Créditos Aprobados. Ciclo 2017 II 131

Índice de Figuras

Figura 1: Dominios de la competencia matemática evaluados en PISA 2015 11
Figura 2: Componentes de una neurona 38
Figura 3: Funcionamiento de una red neuronal artificial 41
Figura 4: Función Sigmoidal 42
Figura 5: Función Tangente Hiperbólica 43
Figura 6: Función Secante Hiperbólica 43
Figura 7: Función Identidad 44
Figura 8: Función Escalón 44
Figura 9: Componentes de una red neuronal artificial 45
Figura 10. Estructura de una red neuronal tipo perceptrón multicapa 48
Figura 11: Relación entre Sensibilidad y Especificidad 56
Figura 12: Curva ROC 58
Figura 13: Estructura de la Red Neuronal tipo Perceptrón Multicapa para Resultado. Ciclo 2017 I 118
Figura 14: Importancia normalizada de las variables para Resultado. Ciclo 2017 I 119
Figura 15: Estructura de la Red Neuronal tipo Perceptrón Multicapa para Resultado. Ciclo 2017 II 121
Figura 16: Importancia normalizada de las variables para Resultado. Ciclo 2017 II 122
Figura 17: Estructura de la Red Neuronal tipo Perceptrón Multicapa para Créditos Aprobados. Ciclo 2017 I 124
Figura 18: Importancia normalizada de las variables para Créditos Aprobados. Ciclo 2017 I 125
Figura 19: Estructura de la Red Neuronal tipo Perceptrón Multicapa para Créditos Aprobados. Ciclo 2017 II 127
Figura 20: Importancia normalizada de las variables para Créditos Aprobados. Ciclo 2017 II 128

Índice de Apéndices

Apéndice A: Tablas de Contingencia 147

Apéndice B: Análisis Exploratorio Univariado 152

Apéndice C. Análisis Exploratorio Bivariado 165

Apéndice D : Matriz de Consistencia 174

Apéndice E : Glosario de términos 175

CAPÍTULO 1: INTRODUCCIÓN

La Universidad Nacional Agraria La Molina (UNALM) ubicada en la ciudad de Lima – Perú es la primera universidad pública que obtuvo la licencia de funcionamiento institucional otorgada por la Superintendencia Nacional de Educación Superior Universitaria, esto significa que la UNALM ofrece servicios educativos con altos estándares de calidad. Esto implica que la UNALM debe continuar con la tarea de generar indicadores que mejoren la gestión educativa y administrativa de la universidad para mantenerse a la vanguardia de las exigencias nacionales e internacionales.

Uno de estos indicadores está relacionado con el desempeño de los estudiantes, es por ello que este trabajo de investigación está orientado a comparar tres métodos multivariantes y determinar el modelo que analice mejor el desempeño académico de los estudiantes ingresantes en el curso de matemática.

Para lograr ese propósito se propone la siguiente metodología:

1. Determinar el modelo de regresión logística binaria que clasifique mejor el desempeño académico de los estudiantes de primer ciclo en el curso de matemática.
2. Determinar el modelo discriminante que clasifique mejor el desempeño académico de los estudiantes de primer ciclo en el curso de matemática.
3. Determinar el modelo de redes neuronales tipo perceptrón multicapa que clasifique mejor el desempeño académico de los estudiantes de primer ciclo en el curso de matemática.
4. Comparar los modelos discriminante, de regresión logística binaria y redes neuronales tipo perceptrón multicapa mediante los indicadores:

tasa de clasificación correcta, sensibilidad, especificidad y área bajo la curva ROC.

La base de datos empleada procede de la Oficina de Estudios y Registros Académicos y del Centro de Admisión y Promoción de la UNALM.

El presente estudio comprende cuatro capítulos, en el capítulo 1 se formula el problema y la justificación y objetivos de la investigación. En el capítulo 2 se plantean los antecedentes de investigación y las bases teóricas. En el capítulo 3 se expone la metodología de investigación. En el capítulo 4 se presentan los resultados de la investigación y la discusión de los resultados. Finalmente, se exponen las conclusiones y recomendaciones.

1.1. Situación Problemática

La Universidad Nacional Agraria La Molina ofrece doce carreras profesionales en dos concursos de admisión que convoca anualmente, seleccionando a más de cuatrocientos nuevos estudiantes en cada convocatoria. Todos los estudiantes recién ingresados tienen diferente malla curricular por ende los cursos asignados son diferenciados de acuerdo a la carrera profesional. Sin embargo, todos coinciden en tres cursos que son matemática básica, lengua y proyecto universitario personalizado.

Desde hace más de dos décadas, los estudiantes de la UNALM presentan una seria dificultad para aprobar los cursos de matemática y más aún el primero de ellos: matemática básica, cuya tasa de desaprobación bordea el 30% en cada ciclo académico como se observa en la Tabla 1.

Tabla 1

Tasa de Desaprobación del Curso Matemática Básica

Ciclo	Número de Desaprobados	Estudiantes matriculados	Tasa de desaprobación
2010 I	138	433	31,9
2010 II	122	480	25,4
2011 I	134	449	29,8
2011 II	138	462	29,9
2012 I	102	423	24,1
2012 II	118	433	27,3
2013 I	115	408	28,2
2013 II	101	409	24,7
2014 I	93	436	21,3
2014 II	120	452	26,5
2015 I	126	434	29,0
2015 II	164	440	37,3
2016 I	142	478	29,7
2016 II	172	559	30,8

Nota. Fuente: Actas de notas del curso.

Esta situación afecta el proceso de matrícula de los estudiantes pues no alcanzan una vacante en los grupos horarios que se programan, lo

que genera el despliegue adicional de docentes, recursos logísticos y de infraestructura para el siguiente ciclo, a este problema se le añade un aspecto importante como es la poca motivación de los estudiantes desaprobados. Al respecto, Castro E. y Ávila J. (2013) en su estudio con estudiantes de primer año universitario matriculados en el curso de cálculo inicial, señalan que la motivación es un elemento crucial en el desempeño de los estudiantes que a su vez fortalece la resiliencia y con ellos se pueden superar los problemas emocionales y afectivos de los estudiantes. Lo mencionado por los autores tiene concordancia con Bazán J. y Sotero H. (1997) que, en su estudio sobre la actitud de los estudiantes de la UNALM, encontraron que los estudiantes presentan una actitud negativa hacia las matemáticas en el primer ciclo de estudios universitarios, además no hallaron diferencias por sexo pero sí las hubo, de acuerdo a la carrera profesional del estudiante.

En la UNALM, la proporción de estudiantes ingresantes del año 2017 que desaprobaron el curso de matemática fue superior al 55% del total de estudiantes matriculados. Además, del total de estudiantes que desaprobaron hasta tres cursos, el 39% desaprobó el curso de matemática, como se muestra en las Tablas 2 y 3.

Tabla 2

Estudiantes ingresantes matriculados según cantidad de cursos aprobados en los ciclos 2017 I y 2017 II

	2017 I		2017 II	
Número de estudiantes	N	%	N	%
Todos los cursos matriculados aprobados	203	44,2	137	37,8
Al menos un curso matriculado desaprobado	256	55,8	226	62,2
Total ingresantes matriculados	459	100,0	363	100,0

Nota. Fuente: Actas de notas del curso. OERA-UNALM

Tabla 3

Estudiantes ingresantes según cantidad de cursos desaprobados en los ciclos 2017 I y 2017 II

Número de cursos desaprobados	2017 I			2017 II		
	N	Desapr Matem	%	N	Desapr Matem	%
Uno	111	18	16,2	75	12	16,0
Dos	82	36	43,9	82	52	63,4
Tres	33	18	54,5	40	28	70,0
Cuatro	20	15	75,0	16	14	87,5
Cinco o más	10	8	80,0	13	9	69,0
Total estudiantes	256	95	37,1	226	115	50,9

Nota. N: Número total de estudiantes. Desapr Matem: Número de estudiantes desaprobados en matemática. Fuente: Actas de notas del curso. OERA-UNALM

Por ello, resulta prioritario identificar las principales características de los estudiantes con alto riesgo de desaprobar el primer curso de matemática en la UNALM. Es necesario encontrar un modelo que permita la clasificación de los estudiantes de acuerdo con su desempeño académico en el primer ciclo de estudios universitarios, esto facilitará la toma de decisiones desde el primer día que los estudiantes alcanzan una vacante en una de las carreras profesionales que ofrece la universidad. El modelo multivariante de clasificación de los estudiantes debe ser el ideal y preciso de tal forma que permita resolver la problemática existente en la UNALM.

El objetivo es comparar los métodos multivariantes que analicen mejor el desempeño académico de los estudiantes y determinar el modelo que clasifique mejor a los grupos de estudiantes de acuerdo con su desempeño académico reflejado en el promedio final del curso de matemática y los créditos aprobados en el primer ciclo académico.

1.2. Formulación del Problema

Los estudiantes que recién ingresan a la UNALM tienen asignado un grupo horario por cada curso matriculado bajo el criterio de la carrera profesional a la que pertenece, es decir, no se toman en cuenta otros criterios como el de la edad, lugar de procedencia, tipo y lugar de procedencia del colegio donde culminó sus estudios escolares, número de veces que postuló a la UNALM, tipo de preparación preuniversitaria, orden de mérito de ingreso a la UNALM, notas y puntaje general obtenidos en el examen de admisión, opción de ingreso, etc.

La información con la que cuenta la UNALM pasa desapercibida para los docentes que impartirán clases a los estudiantes que recién ingresan a la UNALM. Sin embargo, los resultados, al finalizar el ciclo académico, nos revelan la necesidad de conocer las variables más importantes que influyen sobre el desempeño académico de estudiantes de primer ciclo en el curso de matemática y el desempeño académico de los estudiantes al finalizar el primer ciclo de estudios universitarios. Así también se requiere conocer el modelo multivariante que permite optimizar la clasificación de los estudiantes de acuerdo a su desempeño académico.

Por todo ello se plantea la siguiente pregunta ¿Cuál de los métodos multivariantes analiza mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM?

Además, en este estudio se formulan los siguientes problemas específicos que requieren ser resueltos:

- ¿La regresión logística binaria analiza mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM?
- ¿El análisis discriminante analiza mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM?

- ¿Las redes neuronales tipo perceptrón multicapa analiza mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM?

1.3. Justificación de la Investigación

En la actualidad, en la bibliografía especializada existen pocos estudios que abordan el problema de establecer o determinar los factores asociados al rendimiento académico en cursos de matemática a nivel de la educación superior universitaria.

Este estudio cobra gran importancia debido a que responde a la necesidad de establecer mecanismos de detección temprana de aquellos estudiantes que siendo ingresantes y matriculados tienen una alta probabilidad de desaprobar el primer curso de matemática en la UNALM. En consecuencia, se podrá identificar el grupo de riesgo, es decir, a aquellos estudiantes ingresantes y matriculados que tienen una alta probabilidad de desaprobar el curso. Las autoridades universitarias tomarán las decisiones más adecuadas sobre este indicador, tales como, el dictado del ciclo cero antes del inicio del ciclo académico, formación de un grupo especial que cuente con asesorías académicas personalizadas durante el ciclo académico, asignación de un tutor para el seguimiento constante y oportuno de las condiciones en las que estos estudiantes rinden sus evaluaciones parciales y prácticas calificadas.

Se pretende que este trabajo de investigación de tesis tenga un impacto positivo tanto a nivel académico como de gestión ya que de conseguir la disminución en la tasa de desaprobación en el curso de matemática, la UNALM dejará de destinar un alto presupuesto a la atención de la demanda de profesores, aulas y recursos logísticos que en la actualidad generan debido a la existencia de un alto número de estudiantes en condición de repitentes de los cursos de matemática que se ofrece en la universidad.

1.4. Objetivos de la Investigación

1.4.1 Objetivo General

Comparar los métodos multivariantes que analizan mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM.

1.4.2 Objetivos Específicos

- Analizar el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM mediante la regresión logística binaria.
- Analizar el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM mediante el análisis discriminante.
- Analizar el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM mediante las redes neuronales tipo perceptrón multicapa.

CAPÍTULO 2: MARCO TEÓRICO

2.1 Antecedentes de investigación

La Organización para la Cooperación y el Desarrollo Económicos (OCDE) fue fundada en el año 1961 y está integrada por 36 países miembros de todo el mundo. Su misión es la de promover políticas que mejoren el bienestar económico y social de las personas alrededor del mundo. Si bien el Perú no es país miembro de la OCDE, está camino a lograrlo ya que participa de forma activa en diversos comités y colabora desde el año 2015 con el Programa País que tiene entre otros objetivos, incrementar la calidad de la educación en el Perú. (OCDE, 2019).

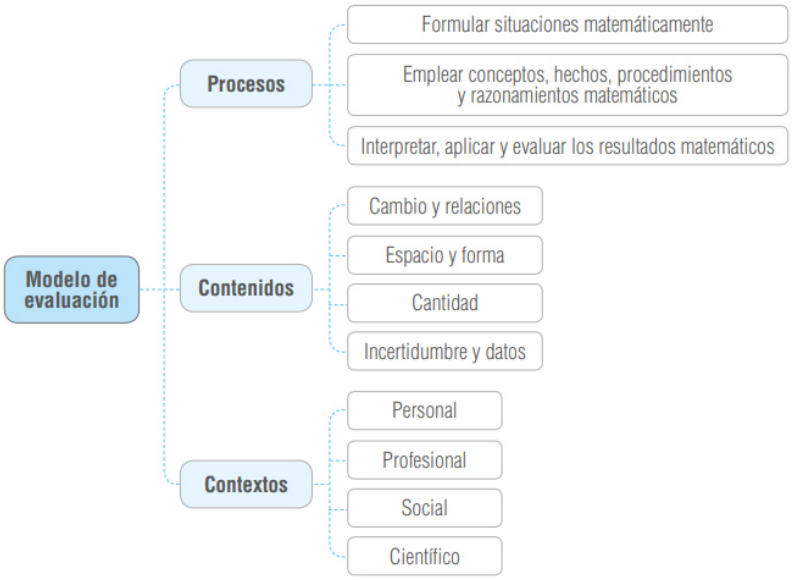
Uno de los programas de la OCDE es la evaluación internacional de alumnos mediante el cual se lleva a cabo la prueba PISA (Programme for International Student Assessment) desde el año 2000 y por cada 3 años. Es un estudio a nivel mundial que mide el rendimiento en el área de ciencia, lectura y matemática, de escolares entre 15 años 3 meses y 16 años dos meses de edad, son exámenes estandarizados que no miden los conocimientos adquiridos sino que más bien mide lo que son capaces de resolver los estudiantes con esos conocimientos. (OCDE).

El Perú participa desde el año 2000 en la prueba PISA obteniendo pésimos resultados, ocupando siempre los últimos puestos en el cuadro general. Los resultados de la prueba PISA del año 2015 aplicada en 70 países dan cuenta que en toda la región centro y sudamericana no estamos obteniendo los mejores puntajes. Según un

informe del Ministerio de Educación del año 2017, en el Perú participaron un total de 6971 estudiantes de 281 instituciones educativas de todo el país y de acuerdo al informe oficial de la OCDE 2016, el Perú ocupó en el área matemática, el puesto 62 con un puntaje promedio de 387 siendo la media de 490 y el puntaje más alto de 564 correspondiente a Singapur.

Figura 1.

Dominios de la competencia matemática evaluados en PISA 2015



Nota. Fuente: Organización para la Cooperación y el Desarrollo Económico (2016a) Tomado de *El Perú en PISA 2015. Informe nacional de resultados* por el Ministerio de Educación, 2017, p.75 (http://umc.minedu.gob.pe/wp-content/uploads/2017/04/ Libro_PISA.pdf).

Desde el año 2006 y a nivel nacional, el Ministerio de Educación (MINEDU) lleva a cabo la Evaluación Censal de Estudiantes (ECE) con la finalidad de conocer la calidad del aprendizaje de los estudiantes escolares en relación a lo propuesto en el Currículo Nacional de Educación Básica Regular. Es una prueba que se aplica anualmente a estudiantes de segundo y cuarto grado de primaria y segundo año de secundaria pertenecientes a instituciones educativas públicas y privadas de educación básica regular en todo el país. Esta prueba mide el aprendizaje de los estudiantes en las áreas de Comprensión Lectora

y Matemática, y a partir del año 2015 también se evalúan las áreas de Historia, Geografía y Economía en estudiantes de segundo año de secundaria. (Ministerio de Educación, 2017).

La ECE 2016 se aplicó a 503 841 estudiantes de segundo año de secundaria correspondientes a 12 875 instituciones educativas en todo el país, y se encontró que existe una brecha de 11,2% entre el rendimiento satisfactorio en matemática para estudiantes provenientes de colegios privados y públicos y una diferencia de 10,2% para el mismo nivel de logro entre los estudiantes del sector urbano y rural. En el siguiente cuadro se observan los resultados obtenidos en dicha evaluación, evidenciándose que sólo el 28,4% del total de estudiantes obtuvieron niveles de logro aceptables, es decir, los resultados más bajos (nivel de logro en inicio y previo al inicio) corresponden en mayor proporción al área de matemática (71,6%). Existe una leve mejora con respecto al año 2015, sin embargo, aún es necesaria la aplicación de nuevas y mejores estrategias de enseñanza aprendizaje acompañado de capacitación y actualización docente. (Ministerio de Educación, 2017). En la Tabla 4 se muestra el rendimiento promedio en las 3 áreas evaluadas en el año 2016.

Tabla 4

Resultados a Nivel Nacional de la ECE 2016 – Segundo Año de Secundaria

Niveles de logro	Áreas evaluadas		
	Matemática	Lectura	Historia, Geografía y Economía
Satisfactorio	11,5	14,3	15,0
En proceso	16,9	27,5	34,0
En inicio	39,3	37,7	28,1
Previo al inicio	32,3	20,5	22,9
Total	100,0	100,0	100,0

Nota. Adaptado de ¿Cuánto aprenden nuestros estudiantes? Resultados de la Evaluación Censal de Estudiantes ECE 2016, del Ministerio de Educación, 2017 (http://umc.minedu.gob.pe/wp-content/uploads/2017/04/Resultados-Nac ionales-2016.pdf).

Esta problemática nacional traspasa el nivel escolar ya que el rendimiento académico de los estudiantes de primeros ciclos en la universidad peruana no dista de los resultados observados, debido al alto índice de estudiantes desaprobados en materias básicas como las matemáticas.

Tomassini, M. (2018) analizó la influencia de los factores educativos y familiares que afectan el rendimiento académico de los estudiantes de agronomía de la Universidad Nacional de Huancavelica y la Universidad para el Desarrollo Andino. En este estudio, Tomassini señala que la experiencia escolar en el nivel primaria y secundaria influye directamente en el desempeño de los estudiantes de educación superior debido a que estos ofrecen al estudiante una mejor formación académica. Así mismo, indica que la edad es un factor que influye en el rendimiento académico de los estudiantes universitarios.

Clemente (1983) citado por Llerena (2017) señala que para las materias de lengua y matemática, las notas son un indicador fundamental del rendimiento académico y tiene en cuenta las calificaciones a lo largo del curso. Asimismo, menciona que Pérez-Serrano (1981) concluye que las calificaciones son el mejor criterio con que se cuenta que para definir el rendimiento.

Ramón, P. y Plasencia, S. (2010) señalan que son diversos los factores que afectan el aprendizaje de una disciplina. Existen algunos correspondientes al alumno como la inteligencia, ansiedad, motivación, antecedentes escolares, hábitos de estudio, conocimientos anteriores, estado de salud, actitudes frente a la disciplina. Otros factores están asociados al docente tales como la preparación académica, preparación metodológica, motivación, didáctica, satisfacción laboral, experiencia docente, etc. También existen factores que pueden atribuirse al currículo y la programación de la enseñanza como: la elaboración y organización de objetivos, la selección y dosificación de contenidos, las estrategias seleccionadas, los recursos disponibles,

criterios y procedimientos de evaluación, uso de medios y materiales didácticos, etc. En el estudio buscaron establecer la relación que existe entre el antecedente del proceso de admisión, la actitud para la matemática, la habilidad del razonamiento matemático, el desempeño global con el rendimiento en matemática de los estudiantes de la Facultad de Ciencias. Se halló que la asociación entre la variable, desempeño global y rendimiento en matemática es muy bajo, también una asociación muy baja con respecto a la actitud frente a las matemáticas y respecto a los conocimientos adquiridos sobre matemática en la educación secundaria.

En el año 2005, Gonzales realizó un estudio sobre el rendimiento académico en matemática en estudiantes preuniversitarios de la Universidad de Málaga, en el que señala que la educación matemática estimula la capacidad de análisis, síntesis, abstracción, el razonamiento lógico, el espíritu crítico y científico de quien la estudia. Es por ello que su conocimiento se vuelve necesario en una sociedad que evoluciona diariamente tanto en los métodos de enseñanza como en el empleo de las TIC's que complementan lo primero.

Vivo, Sánchez, & Franco (2004) realizaron un estudio del rendimiento académico en los curso de matemáticas a partir de información académica preuniversitaria y notas de ingreso de estudiantes de las carreras de Economía y Administración y Dirección de Empresas la Universidad de Murcia en los periodos académicos 98-99. Los datos fueron analizados mediante las curvas ROC (Receiver Operating Characteristic) que permitió discriminar entre los alumnos aprobados y desaprobados en los cursos de matemáticas. Para analizar el método de clasificación emplearon las medidas de Sensibilidad (probabilidad de que un alumno aprobado sea clasificado correctamente) y Especificidad (probabilidad de que un alumno desaprobado sea clasificado correctamente). Los resultados señalan que para el curso de Estadística I para alumnos de Economía y Administración son

buenos clasificadores las siguientes variables: nota media obtenida en secundaria y nota general de selección.

En relación con los métodos multivariantes empleados en este trabajo de investigación de tesis, Torrado y Berlanga (2013) afirman que “mediante el análisis discriminante se puede establecer el poder explicativo y discriminatorio de las características que diferencian a los alumnos según su rendimiento”. En tanto que Salazar (2005) señala que la regresión logística es una técnica apropiada cuando la variable dependiente es categórica. Finalmente, López (2012) menciona que las redes neuronales tipo perceptrón multicapa son modelos que constituyen un instrumento idóneo en la clasificación de alumnos ingresantes en los cursos de Álgebra y Geometría Analítica durante el primer ciclo académico de estudios de acuerdo con los altos porcentajes de clasificación correcta obtenidos.

2.2 Bases Teóricas

2.2.1 Educación matemática

La matemática es una ciencia formal y exacta y su enseñanza permite descubrir diversas habilidades en los estudiantes, de allí su importancia y estudio que se remonta a muchos años atrás. De acuerdo al Diccionario de la Real Academia Española (DRAE) la matemática se define como la ciencia deductiva que estudia las propiedades de los entes abstractos, como los números, figuras geométricas o símbolos, y sus relaciones. (Real Academia Española, s.f.).

La educación matemática cobra especial interés desde los años 60 y desde entonces ha sido objeto de numerosos estudios sin perder de vista al binomio enseñanza-aprendizaje, es así que es válido rescatar lo señalado por De Guzmán (2007) la educación ha de hacer, necesariamente, referencia a lo más profundo de la persona, una persona aún por conformar, a la sociedad en evolución en la que esta persona se ha de integrar, a la cultura en que esta sociedad se desarrolla, a los medios concretos personales y materiales de los que en el momento se puede o se quiere disponer, a las finalidades prioritarias que a esta educación se le quieran asignar y que pueden ser extraordinariamente variadas.

Diversos estudios relacionados con la educación matemática o matemática educativa tienen por objetivos explorar, evaluar, medir, clasificar o predecir el rendimiento académico de los estudiantes en los cursos de matemática, por ello iniciaremos con algunas definiciones halladas sobre este tema.

2.2.2 Rendimiento académico

El rendimiento es el producto o utilidad que rinde o da alguien o algo, asimismo, es la proporción entre el producto o el resultado obtenido y los medios utilizados. El desempeño es la acción y efecto de trabajar o dedicarse a una actividad. En tanto, académico se define como perteneciente o relativo a las academias, asimismo es propio y característico de las académicas, también se define como perteneciente o relativo a centros oficiales de enseñanza especialmente a los superiores. (Real Academia Española, s.f.). En el presente estudio nos referiremos al desempeño académico como el resultado obtenido por los estudiantes universitarios al finalizar el primer ciclo académico en la universidad a nivel general y especialmente en el curso de matemática.

Touron (1984) mencionado por Gonzales (2015) define el rendimiento académico universitario como un resultado del aprendizaje, suscitado por la actividad educativa del profesor producido en el alumno. Se expresa en una calificación cuantitativa y cualitativa en muchos casos, una nota que si es consistente y válida será el reflejo de un determinado aprendizaje o del logro de unos objetivos preestablecidos.

El problema del bajo rendimiento académico en matemática no solo se limita al nivel universitario, sino que de acuerdo a la literatura existente, este se inicia desde temprana edad en los niveles de primaria y secundaria, de allí que existen sobre él numerosas investigaciones, es así que Flores (2015) afirma que existen factores que influyen negativamente en la consolidación de un buen rendimiento académico tales como: ejercitación, repetición permanente, poca consolidación de nuevos con viejos conocimientos. A estos factores se precisan añadir aquellos que están vinculados con los exámenes de ingreso a la universidad.

2.2.3 Factores asociados al rendimiento

Diversos estudios relacionados a los factores asociados con el rendimiento académico en matemática se enfocan en los aspectos socioeducativos, demográficos, actitudinales, académicos (cursos, horarios de clase, carreras profesionales, infraestructura, profesorado, saberes previos), examen de selección o admisión a la universidad, entre otros. Se aborda el estudio de la transición a la universidad como un proceso multifactorial que vincula situaciones personales, familiares, sociales, institucionales, periodo preparatorio, entre otras.

Respecto a los estudiantes recién ingresantes a la universidad, Figueroa (2003) precisa que debe generar un nuevo espacio para las relaciones sociales que no siempre favorece la socialización y la creación de redes y grupo, constituyéndose en uno de los retos que debe asumir. La investigación comprendió una muestra de 1382 estudiantes que en un 71% provienen de colegios públicos y son mayoritariamente estudiantes mujeres que en un 84% ingresaron a la carrera profesional elegida como primera opción en la Universidad de Barcelona. Un tercio de los estudiantes manifestó una baja o insuficiente autoconfianza sobre las demandas que exigen los estudios universitarios. El rendimiento académico previo resultó un factor discriminativo importante en el proceso de transición académica.

Vargas y Montero (2016) en el estudio sobre el rendimiento académico de estudiantes de segundo año de la carrera de Ingeniería en la Universidad Nacional de Ingeniería de Nicaragua, matriculados en el curso Matemáticas III, tomaron en cuenta variables del estudiante tales como sexo, edad, lugar de origen, opción inicial de carrera, escala de motivación, escala de inteligencia emocional. Además, incluyeron otras variables como horario de estudio, tipo de colegio de procedencia, nota

promedio en la educación secundaria, escala socioeconómica, entre otros. De los resultados del estudio, se evidenció que la nota promedio de la educación secundaria no tiene relación directa con el rendimiento académico medido por la nota final del curso. Dadas las condiciones establecidas en el estudio, sus conclusiones se pueden generalizar para estudiantes de las carreras de ingeniería en las universidades de américa latina debido a las condiciones similares de desigualdad y falta de oportunidades educativas.

Unas de las variables que merecen la atención de investigadores del rendimiento académico en matemática es el sexo de los estudiantes, como el estudio de Aguiar (2011) de la Universidad de Guadalajara en el que evaluó el rendimiento de los estudiantes en 12 cursos sobre las diferencias en el rendimiento de estudiantes de la UNAM en el cual se evidenció que las estudiantes mujeres lograron un promedio más alto que los varones pese a optar por carreras que son mayoritariamente masculinos como las agropecuarias, ingeniería y tecnológicas. Independientemente del curso matriculado, son las mujeres quienes obtienen mejores notas que los varones. No obstante, propone profundizar el estudio a fin de tener mayores evidencias sobre los factores que influyen en este resultado.

El Ministerio de Educación (2016) realizó un estudio sobre los resultados de la evaluación PISA 2012 en el cual se emplearon diversas variables predictoras tales como género del estudiante, lengua materna, relación docente-estudiante, tipo de gestión y tamaño de la institución educativa y la proporción de mujeres en la institución educativa, entre otras. En el estudio se afirma que uno de los casos de inequidad existente entre estudiantes de 15 años de edad corresponde a las diferencias en el rendimiento académico según el género o sexo del estudiante. Según la Organization for Economic Cooperation and Development

(2013) señalado por el informe del Ministerio de Educación, en el Perú, la brecha no es tan amplia en el nivel de educación inicial, sin embargo, se hace notoria en el nivel secundario, encontrándose que las mujeres presentan menores logro de aprendizaje que los estudiantes varones.

Debido a la importancia del desarrollo de la competencia matemática de los estudiantes y los bajos resultados mostrados en las diferentes etapas de su preparación educativa, el Ministerio de Educación (2016) propone conocer y comprender todos los elementos que intervienen en el logro de los aprendizajes matemáticos de los estudiantes para realizar una intervención oportuna y de esta manera revertir la tendencia. Entre los resultados del estudio también destacan que los estudiantes provenientes de instituciones educativas particulares o privadas obtuvieron mejor rendimiento en matemática que aquellos estudiantes provenientes de colegios de gestión pública. No obstante, se incluye en el estudio una diferenciación entre colegios que albergan una población estudiantil grande o mediana de aquellos con pequeña población estudiantil, siendo el rendimiento académico en matemática más favorable en el primer grupo. No se evidenció diferencias significativas en el rendimiento académico entre estudiantes que provienen de colegios pequeños pudiendo ser de gestión privada o pública.

Para Ramón y Plasencia (2010) uno de los factores asociados con el éxito en el rendimiento académico de los estudiantes es la vocación y la carrera profesional elegida. La vocación la definen como la disposición interior que se inclina hacia la profesión y el ejercicio de una determinada actividad. En tanto que la madurez vocacional es para los autores, la disposición para hacer frente a las tareas vocacionales o al desarrollo de la

carrera profesional con las que uno está realmente identificado o a punto de estarlo.

Gonzáles (2015) en su estudio sobre la relación entre el rendimiento académico en matemática y las variables afectivas y cognitivas en estudiantes preuniversitarios de la Universidad Católica Santo Toribio de Mogrovejo. Los estudiantes de los primeros ciclos presentaron un bajo rendimiento en los cursos de matemáticas y bajo la premisa que el rendimiento académico en el nivel precedente influye en el desempeño académico universitario, el autor encontró diferencias entre el rendimiento según el sexo del estudiante, siendo el más bajo el de las mujeres. Incluyó en su estudio otros factores que influyen el rendimiento académico como el coeficiente intelectual, las características del pensamiento formal, estrategias de aprendizaje, entre otros.

De acuerdo a la investigación realizada por Gonzáles (2017) en su estudio sobre la relación entre los puntajes del examen de admisión para ingreso a la universidad y el rendimiento académico de los estudiantes de la Facultad de ciencias Matemáticas de la UNMS, halló que los puntajes más altos obtenidos en el examen de admisión a la UNMSM no estaban necesariamente relacionados con los mejores rendimientos académicos en los dos primeros años de estudios pero sí tienen mayores posibilidades de clasificación como estudiantes de rendimiento medio o superior. Miljánovich (2005) mencionado por Gonzáles, señala que un examen de admisión se considera adecuado si es que guarda una correlación positiva y significativa con los rendimientos académicos de los estudiantes de nivel universitario. Además, Gonzáles encontró que en promedio, menos del 8% de la variación en el rendimiento académico de los estudiantes se explicó por el puntaje obtenido en el examen de admisión durante el periodo 2010 y 2013.

Los métodos multivariantes que se emplearon en el presente estudio de investigación del desempeño académico de los estudiantes ingresantes en el primer curso de matemática de la UNALM son: análisis discriminante, regresión logística binaria y redes neuronales tipo perceptrón multicapa que se desarrollará a continuación.

2.2.4 Regresión Logística Binaria

2.2.4.1 Concepto

El modelo logístico fue inicialmente elaborado por Berskon (1944) en contexto biométrico, basado en contribuciones previas de los estadísticos R.A. de Fisher, F. Yates y M.S. Bartlett. Cox (1970), en un libro sobre el análisis de datos.

López (2015) señala que si en un estudio, se tienen variables independientes cuantitativas o cualitativas y se desea encontrar el efecto de estas sobre una variable dependiente de tipo dicotómica, entonces, lo que se recomienda es el empleo de la regresión logística binaria.

Este método es muy útil para los casos en que se desea determinar la presencia o ausencia de una característica a partir de variables denominadas predictoras. Tiene mucha similitud al análisis de regresión lineal y tiene la ventaja que no se requiere probar el cumplimiento de supuestos tales como normalidad y homocedasticidad.

Otra de las ventajas que señala Hair (1999) es que la regresión logística se puede aplicar a muchas más situaciones que el análisis discriminante, además no tiene supuestos tan estrictos y es más robusta cuando estos supuestos no se cumplen.

2.2.4.2 Modelo de regresión logística

El modelo logístico o también llamado por Pérez (2004) como modelo de regresión logística o modelo de regresión con variable dependiente binomial es un modelo que permite estimar la probabilidad de un suceso que depende de los valores de ciertas variables llamadas covariables (Cuadras, 2018). Es pertinente el análisis de este modelo cuando la relación entre las variables dependientes e independientes es no lineal (Catena, 2003).

Según Cuadras (2018), si se considera una variable binaria y que toma dos valores: $y = 1$, si ocurre el evento A y $y = 0$, si no ocurre el evento A; además, sea la probabilidad p que depende de los valores de las variables $X_1, X_2, \dots, X_p$ de modo que si $\boldsymbol{x} = (x_1; x_2; \dots; x_p)'$ son las observaciones de un individuo ω respecto a las variables, entonces la probabilidad de que ocurra A dado $\boldsymbol{x}$ es $p(\boldsymbol{x}) = p(y = 1/\boldsymbol{x})$, nótese que p depende de $\boldsymbol{x}$. Del mismo modo, la probabilidad de que no ocurra el evento o suceso A dado $\boldsymbol{x}$ está dada por $1 - p(\boldsymbol{x}) = p(y = 0/\boldsymbol{x})$.

La función logística definida por Pérez (2004) es aquella que tiene la siguiente regla de correspondencia:

$$f(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}$$

A partir de esa función se puede modelar la probabilidad de un proceso binomial como una función logística de una combinación lineal de la variable dependiente. El modelo de regresión logística múltiple tendrá la expresión:

$$p(\boldsymbol{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}}$$

$$p(\boldsymbol{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta' X)}} = \frac{e^{\beta_0 + \beta' X}}{1 + e^{\beta_0 + \beta' X}}$$

En donde $\beta = (\beta_1, \dots, \beta_p)'$ son los parámetros de la regresión y también se puede observar que el objetivo es hallar todos los coeficientes $\beta_0, \beta_1, \dots, \beta_p$ que mejor se ajusten al modelo.

Además, se tiene que:

$$1 - p(\boldsymbol{x}) = \frac{e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)}}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)}}$$

De donde al dividir se obtiene:

$$\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}$$

Al aplicar logaritmo natural a ambos lados de la igualdad, se deduce:

$$ln\left[\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right] = ln\left[e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}\right]$$

$$ln\left[\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Como señala Catena (2003) con esta última igualdad se muestra la relación lineal que existe entre el cociente de probabilidades, es decir la probabilidad de que ocurra el evento A con respecto a la probabilidad de que no suceda tal evento. Ahora podemos, según Johnson (2007), considerar la razón de probabilidades denominada odds ratio ($odds$), en lugar de modelar directamente la probabilidad p mediante un modelo lineal.

$$odds = \frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}$$

Además, en regresión logística para una variable binaria, se modela el logaritmo natural de las odds ratio, denominada $logit(p)$. El $logit$ es una función de la probabilidad de p, tal forma que:

$$logit(p) = ln(odds) = ln\left[\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right]$$

$$logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

2.2.4.3 Estimación de los coeficientes del modelo

Hair (1999) señala que la regresión logística y la regresión múltiple son muy similares, pero difieren en la estimación de coeficientes. La estimación de los coeficientes del modelo de regresión logística por mínimos cuadrados no se aplica pues no se cumple la normalidad de la variable dependiente (Pérez, 2004). Entonces, la regresión logística no minimiza la desviación de los cuadrados, más bien maximiza la verosimilitud de la ocurrencia de un suceso.

Según señala Cuadras (2018), si Y es la variable binomial dependiente, entonces, la función de verosimilitud para una variable binomial puntual está dada por:

$$L = (p(\boldsymbol{x}))^y \, (1-p(\boldsymbol{x}))^{1-y}$$

Si se consideran n observaciones independientes, entonces, la función de verosimilitud estará dada por:

$$L = \prod_{i=1}^{n} (p(\boldsymbol{x}_i))^{y_i} (1-p(\boldsymbol{x}_i))^{1-y_i}$$

Al aplicar logaritmo natural a ambos lados de la igualdad, se obtiene:

$$\ln(L) = \sum_{i=1}^{n} y_i \, ln(p(\boldsymbol{x}_i)) + \sum_{i=1}^{n} (1-y_i)(1-p(\boldsymbol{x}_i))$$

Johnson (2007) señala que el valor de los parámetros que maximizan la probabilidad no se puede expresar como en el caso de los modelos lineales. La solución numérica se halla

asumiendo un valor inicial y luego se realizan iteraciones para el valor máximo de la función de probabilidad. Asimismo, Catena (2003) añade que el proceso finaliza cuando no se puede mejorar el ajuste.

2.2.4.4 Contraste de hipótesis sobre los coeficientes.

Cuadras (2018) menciona que según el teorema central del límite, los estimadores por máxima verosimilitud de los parámetros del modelo logístico son asintóticamente normales y su matriz de varianza covarianzas es calculable a partir del logaritmo de maximización de la funciones de verosimilitud.

Sean $\hat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p})'$ la estimación de los parámetros β y la matriz $I_\beta = X'VX$, donde:

$$V = \begin{bmatrix} p(x_1)\big(1 - p(x_1)\big) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p(x_n)\big(1 - p(x_n)\big) \end{bmatrix}$$

La distribución asintótica de $\hat{\beta}$ es normal multivariante $N_{p+1}(\beta, I_\beta^{-1})$. Así también, de forma particular, la distribución asintótica del parámetro $\widehat{\beta_i}$ es normal $N(\beta_i, var(\widehat{\beta_i}))$ donde $var(\widehat{\beta_i})$ es e correspondiente elemento diagonal de la matriz inversa I_β^{-1}. (Cuadras, 2018).

Pérez (2004) precisa que la hipótesis estadística a contrastar es:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

El estadístico para el contraste es: $Z = \frac{\widehat{\beta_i}}{\widehat{\sigma}(\widehat{\beta_i})} \sim N(0,1)$

La región crítica está dada por: $| Z | > Z_{\alpha/2}$.

Si se rechaza la hipótesis H_0 para algún valor de i, ello significa que la variable y no depende de la variable x_i con lo cual no aparecerá en el modelo.

Para el contraste, también se emplea el estadístico de Wald definido como $W = Z^2$ que se distribuye como una Chi-cuadrado con un grado de libertad. Se rechaza H_0 si: $W > \chi^2_\alpha$.

2.2.4.5 Bondad de ajuste del modelo.

Según Iglesias (2013), la bondad de ajuste global para comparar la diferencia entre los valores predichos y los valores observados en el modelo de regresión logística se puede evaluar mediante varias pruebas según estén basados en los patrones de las covariables, en las probabilidades estimadas por el modelo o en residuos suavizados. Aquí abordaremos dos de ellos, el estadístico de Hosmer-Lemeshow y el estadístico basado en la devianza D.

a. Estadístico de Hosmer-Lemeshow C_g.

Evalúa la bondad de ajuste del modelo construyendo una tabla de contingencia a la que se aplica un contraste de tipo Chi-cuadrado. El estadístico C_g se basa en la agrupación de las probabilidades estimadas $\widehat{p_j}$. Se forman grupos de tal forma que en el primer grupo están las observaciones con probabilidad predicha más pequeña, en el segundo grupo están los siguientes más pequeños y así sucesivamente. Los puntos de corte son los deciles de riesgo. La Tabla 5 muestra las frecuencias esperadas y observadas en cada grupo, en donde los valores d_i son los deciles de riesgo de las probabilidades estimadas (Iglesias, 2013).

Tabla 5

Frecuencias esperadas y observadas para C_g

Grupos	Respuesta			
	Y=1		Y=0	
	Observado	Esperado	Observado	Esperado
$\widehat{p_J} < d_1$	o_{11}	e_{11}	o_{01}	e_{01}
$d_1 \leq \widehat{p_J} < d_2$	o_{12}	e_{12}	o_{02}	e_{02}
…	…	…	…	…
$d_9 \leq \widehat{p_J} < d_{10}$	o_{1G}	e_{1G}	o_{0G}	e_{0G}
Total	o_1	e_1	o_0	e_0

Nota. Tomado de Iglesias (2013)

El estadístico C_g se obtiene comparando los valores observados y esperados, tal que:

$$C_g = \sum_{k=0}^{1} \sum_{g=1}^{G} \frac{(o_{kg} - e_{kg})^2}{e_{kg}}$$

Mediante estudios de simulación se demostró que cuando el número de patrones o valores distintos de la variable independiente (J) coincide con el número de observaciones, si el número de covariables más 1 es menor que el número de grupos G, bajo la hipótesis del modelo logístico, C_g tiene distribución asintótica χ^2_{G-2} (Hosmer-Lemeshow (1989) mencionado por Iglesias (2013)).

b. Estadístico basado en la Devianza D

Consideremos la función de verosimilitud en una regresión logística. Collete (1991) mencionado por Iglesias (2013).

$$L(\beta_0, \beta_1) = \prod_{j=1}^{J} \binom{n_j}{y_j} {p_j}^{y_j} (1 - p_j)^{n_j - y_j}$$

Y su log-verosimilitud

$$log\ [L(\beta_0, \beta_1)] = \sum_{j=1}^{J} \{\ log \binom{n_j}{y_j} + y_j \log p_j + (n_j - y_j) \log(1 - p_j)\}$$

Sea $\hat{L}_C = L(\hat{\beta}_0, \hat{\beta}_1)$ y $\hat{\beta}_0$, $\hat{\beta}_1$ son los estimadores máximos verosímiles de los parámetros. Bajo el modelo ajustado, la verosimilitud está dada por:

$$log\ \hat{L}_C = \sum_{j=1}^{J} \{\ log \binom{n_j}{y_j} + y_j \log \hat{p}_j + (n_j - y_j) \log(1 - \hat{p}_j)\}$$

Siendo: $\hat{p}_j = \frac{\hat{y}_j}{n_j}$ la probabilidad estimada de la variable dependiente y=1 para el j-ésimo patrón de covariables.

Un modelo saturado es aquel que se ajusta perfectamente a los datos, es decir, las frecuencias esperadas por el modelo coinciden con las observadas y tiene tantos parámetros libres o desconocidos como observaciones diferentes de las variables explicativas (Iglesias, 2013). Sea $\hat{L}_F$ la verosimilitud de este modelo, entonces su log-verosimilitud está dado por:

$$log\ \hat{L}_F = \sum_{j=1}^{J} \{\ log \binom{n_j}{y_j} + y_j\ log\ \tilde{p}_j + (n_j - y_j)\ log(1 - \tilde{p}_j)\}$$

Así, el estadístico de Wilks conocido como devianza, está dado por:

$$D = -2\ log\ \left(\frac{\hat{L}_C}{\hat{L}_F}\right) = -2(log\hat{L}_C - log\hat{L}_F)$$

Este estadístico se distribuye asintóticamente como una Chi-cuadrado.

La hipótesis a contrastar es:

H_0: No existen diferencias entre los valores observados y valores estimados.

H_1: Existen diferencias entre los valores observados y los valores estimados.

La hipótesis nula se rechaza si $D \geq \chi^2_{J-(p+1);\alpha}$, es decir, cuando el p-valor del contraste sea menor al nivel de significación fijado. (Iglesias, 2013).

Cuando se realiza el modelo de análisis de regresión logística con el programa de IBM denominado SPSS se muestran dos test o pruebas, de Ómnibus y de Hosmer-Lemeshow. En la prueba de ómnibus sobre los coeficientes del modelo aparece el valor de menos dos veces la diferencia entre la log-verosimilitud del modelo ajustado y el modelo nulo que solo incluye la constante (Chi-cuadrado), sus grados de libertad y su significación. En la prueba de Hosmer-Lemeshow aparece el valor del estadístico C_g Chi-cuadrado, sus grados de libertad y su significación (Iglesias, 2013).

2.2.5 Análisis Discriminante

2.2.5.1 Concepto

De la Fuente (2011) define el análisis discriminante como una técnica estadística multivariante que permite caracterizar, diferenciar o clasificar a grupos de individuos en base a un conjunto de variables medidas sobre los mismos, que puede ser importante en número. Uno de sus objetivos es determinar si ciertos grupos previamente definidos, difieren entre sí, considerando alguna función lineal o cuadrática de variables, posteriormente se busca emplear esa función para predecir la pertenencia de una nueva observación a alguno de los grupos.

Kovalevski, L. (2012, noviembre) señala que el análisis discriminante es óptimo cuando las variables se ajustan a una distribución normal multivariada en el que se cumple la homocedasticidad dentro de cada grupo. Cuando el análisis discriminante no es óptimo se proponen los métodos no paramétricos de clasificación tales como los k vecinos más cercanos, basado en los núcleos Kernel, árboles de clasificación y regresión (CART: Classification and Regression Trees) y Redes Neuronales Artificiales.

Johnson (2007) afirma que el análisis discriminante es una técnica multivariada relacionada con la separación de conjuntos distintos previamente definidos. El procedimiento de clasificación o asignación conlleva a reglas bien definidas que se puede usar para asignar nuevos objetos.

2.2.5.2 Definición

Sean Ω_1 y Ω_2 dos poblaciones y $X_1, X_2, \dots, X_p$ son variables observables. Sea $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ las observaciones de las variables sobre un individuo ω. Se trata de asignar ω a una de las dos poblaciones. (Cuadras, 2018).

El mismo autor señala que una regla discriminante es un criterio que permite asignar al individuo ω conociendo $(X_1, X_2, \dots, X_p)$ a una de las dos poblaciones. A menudo, esto se plantea mediante una función discriminante $D = (x_1, x_2, \dots, x_p)$. Entonces, la regla de clasificación es:

Si: $D = (x_1, x_2, \dots, x_p) \geq 0$, asignamos ω a Ω_1,

Si: $D = (x_1, x_2, \dots, x_p) < 0$, asignamos ω a Ω_2.

Esta regla divide el espacio p-dimensional R^p en dos regiones:

$$R_1 = \{x/D(x) > 0\}\ ,\ R_2 = \{x/D(x) < 0\}$$

La decisión de clasificar ω a uno de los dos grupos de interés podríamos cometer una equivocación o error, como por ejemplo, cuando asignamos ω a una población a la que no pertenece.

Según Cuadras (2018), la probabilidad de clasificación errónea (pce) está dada por:

$$pce = P(R_2/\Omega_1)\ P(\Omega_1)\ +\ P(R_1/\Omega_2)\ P(\Omega_2)$$

Hair (1999) señala que en el análisis discriminante se precisa el cumplimiento de los supuestos de normalidad multivariante, matriz de covarianzas iguales y ausencia de multicolinealidad.

2.2.5.3 Clasificación con dos grupos. Discriminador lineal

Cuadras (2018) señala lo siguiente, sean μ_1 y μ_2 los vectores de medias de las p variables predictoras en las poblaciones Ω_1 y Ω_2, respectivamente, y supongamos que la matriz de covarianzas Σ es común a ambas poblaciones, es decir, existe homocedasticidad, esto es,

$$\Sigma_1 = \Sigma_2 = \Sigma$$

Donde Σ_1 es la matriz de covarianza del primer grupo o población y Σ_2 es la matriz de covarianza del segundo grupo o población.

Las distancias de Mahalanobis de las observaciones $\boldsymbol{X} = (x_1, x_2, \dots, x_p)'$ de un individuo ω a las poblaciones son:

$$M^2(\boldsymbol{X}, \mu_i) = (\boldsymbol{X} - \mu_i)'\ \Sigma^{-1}\ (\boldsymbol{X} - \mu_i),\ \ i = 1, 2.$$

Se emplea la distancia de Mahalanobis en reemplazo de la distancia Euclidea cuando se observa la presencia de correlación entre las variables.

La clasificación consiste en asignar al individuo ω a la población más próxima, empleando el siguiente criterio:
Si: $M^2(\boldsymbol{X}, \mu_1) < M^2(\boldsymbol{X}, \mu_2)$, asignamos ω a Ω_1 ,
Si: $M^2(\boldsymbol{X}, \mu_1) \geq M^2(\boldsymbol{X}, \mu_2)$, asignamos ω a Ω_2.

Cuadras (2018) continúa, desarrollando esta regla como una función discriminante, obteniendo:

$$M^2(\boldsymbol{X}, \mu_2) - M^2(\boldsymbol{X}, \mu_1) = (\boldsymbol{X} - \mu_2)'\Sigma^{-1}(\boldsymbol{X} - \mu_2) - (\boldsymbol{X} - \mu_1)'\Sigma^{-1}(\boldsymbol{X} - \mu_1)$$

Al desarrollar la igualdad anterior y al aplicar las propiedades de matrices, se obtiene:

$$= \boldsymbol{X}'\Sigma^{-1}\boldsymbol{X} - \boldsymbol{X}'\Sigma^{-1}\mu_2 - \mu_2'\Sigma^{-1}\boldsymbol{X} + \mu_2'\Sigma^{-1}\mu_2 - \boldsymbol{X}'\Sigma^{-1}\boldsymbol{X} + \boldsymbol{X}'\Sigma^{-1}\mu_1 + \mu_1'\Sigma^{-1}\boldsymbol{X} - \mu_1'\Sigma^{-1}\mu_1$$
$$= \mu_2'\Sigma^{-1}\mu_2 - \mu_1'\Sigma^{-1}\mu_1 + (\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X} - \boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}) + (2\boldsymbol{X}'\Sigma^{-1}\mu_1 - 2\boldsymbol{X}'\Sigma^{-1}\mu_2)$$
$$= \mu_2'\Sigma^{-1}\mu_2 - \mu_1'\Sigma^{-1}\mu_1 + (2\boldsymbol{X}'\Sigma^{-1}\mu_1 - 2\boldsymbol{X}'\Sigma^{-1}\mu_2)$$
$$= \mu_2'\Sigma^{-1}\mu_2 - \mu_1'\Sigma^{-1}\mu_1 + \mu_2'\Sigma^{-1}\mu_1 - \mu_2'\Sigma^{-1}\mu_1 + 2\boldsymbol{X}'\Sigma^{-1}(\mu_1 - \mu_2)$$
$$= \mu_2'\Sigma^{-1}\mu_2 - \mu_1'\Sigma^{-1}\mu_1 + \mu_2'\Sigma^{-1}\mu_1 - \mu_1'\Sigma^{-1}\mu_2 + 2\boldsymbol{X}'\Sigma^{-1}(\mu_1 - \mu_2)$$

Al factorizar resulta:

$$= (\mu_2' - \mu_1')\Sigma^{-1}\mu_2 + (\mu_2' - \mu_1')\Sigma^{-1}\mu_1 + 2\boldsymbol{X}'\Sigma^{-1}(\mu_1 - \mu_2)$$
$$= (\mu_2 - \mu_1)'\Sigma^{-1}\mu_2 + (\mu_2 - \mu_1)'\Sigma^{-1}\mu_1 + 2\boldsymbol{X}'\Sigma^{-1}(\mu_1 - \mu_2)$$
$$= (\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 + \mu_1) + 2\boldsymbol{X}'\Sigma^{-1}(\mu_1 - \mu_2)$$

Ahora, se define la función discriminante:

$$L(\boldsymbol{X}) = [\boldsymbol{X} - \frac{1}{2}(\mu_2 + \mu_1)]'\Sigma^{-1}(\mu_2 - \mu_1)$$

Es una función lineal, en este caso es una función discriminante lineal conocida como el discriminador lineal de Fisher. (Cuadras, 2018). Entonces,

$$M^2(\boldsymbol{X}, \mu_2) - M^2(\boldsymbol{X}, \mu_1) = 2L(\boldsymbol{X}) - L[\frac{1}{2}(\mu_1 + \mu_2)]$$

y la regla ahora es:
Si: $L(\boldsymbol{X}) > 0$, asignamos ω a Ω_1,
Si: $L(\boldsymbol{X}) \leq 0$, asignamos ω a Ω_2.

Pérez (2004) señala que existen dos enfoques en el análisis discriminante. El primero de ellos se basa en la obtención de las funciones discriminantes como se ha detallado en esta sección. El segundo enfoque emplea la técnica de correlación canónica y de componentes principales y se denomina análisis discriminante canónico. Cuando se tienen dos grupos, se obtiene una sola función discriminante canónica que será la ecuación de la componente principal:

$$C_1 = a_{11}x_1 + a_{12}\,x_2 + \ \ldots + a_{1p}x_p$$

y el valor propio asociado sería el poder discriminante. Además, los coeficientes de la ecuación (eje discriminante) muestra el peso que cada variable aporta a la discriminación. Se debe advertir que los coeficientes están afectados por las escalas de medida por lo que se precisa tener variables estandarizadas.

2.2.5.4 Regla de la máxima verosimilitud

Cuadras (2018) plantea la siguiente regla de clasificación: Si $f_1(\boldsymbol{X})$ y $f_2(\boldsymbol{X})$ son las densidades de $\boldsymbol{X}$ en Ω_1 y Ω_2 respectivamente. Una regla de clasificación consiste en asignar ω a la población donde la verosimilitud de las observaciones $\boldsymbol{X}$ es más grande, esto es:

Si: $f_1(\boldsymbol{X}) > f_2(\boldsymbol{X})$, asignamos ω a Ω_1,

Si: $f_1(\boldsymbol{X}) \leq f_2(\boldsymbol{X})$, asignamos ω a Ω_2.

La función discriminante estará dada por:

$$V(\boldsymbol{X}) = log\,[f_1(\boldsymbol{X})] - log[f_2(\boldsymbol{X})]$$

2.2.5.5 Regla de Bayes

Cuadras (2018) plantea las probabilidades a priori de que ω pertenezca a cada una de las poblaciones, esto es:

$$\mathrm{q}_1 = \mathrm{P}(\Omega_1) \text{ y } \mathrm{q}_2 = \mathrm{P}(\Omega_2)$$

donde: $\mathrm{q}_1 + \mathrm{q}_2 = 1$

Si se tienen las observaciones $\boldsymbol{X} = (x_1, x_2, \dots, x_p)$, por el Teorema de Bayes, las probabilidades a posteriori de que ω pertenezca a cada una de las poblaciones son:

$$P\left({}^{\Omega_i}/_{\boldsymbol{X}}\right) = \frac{\mathrm{q}_i\, f_i\,(\boldsymbol{X})}{\mathrm{q}_1\, f_1\,(\boldsymbol{X}) + \mathrm{q}_2\, f_2\,(\boldsymbol{X})}\,,\ \ i = 1{,}2$$

Siendo así, la regla de clasificación de Bayes estará dada por:

Si: $P(\Omega_1/\boldsymbol{X}) > P(\Omega_2/\boldsymbol{X})$, asignamos ω a Ω_1,

Si: $P(\Omega_1/\boldsymbol{X}) \leq P(\Omega_2/\boldsymbol{X})$, asignamos ω a Ω_2.

De este modo, el discriminador de Bayes es:

$$B(\boldsymbol{X}) = log\,[f_1(\boldsymbol{X})] - log[f_2(\boldsymbol{X})] + log\left(\frac{q_1}{q_2}\right)$$

La regla de Bayes minimiza la probabilidad de clasificación errónea. Este discriminador de Bayes es óptimo cuando $q_1 = q_2 = \frac{1}{2}$, en esta condición los discriminadores de máxima verosimilitud y de Bayes son iguales, es decir, $B(\boldsymbol{X}) = V(\boldsymbol{X})$.

2.2.5.6 Prueba de igualdad de medias

Rodríguez (2001) señala que para valorar la significación estadística de las funciones discriminantes halladas se debe emplear el estadístico Lambda de Wilks debido a que indica qué tan diferenciados se encuentran los grupos. Por lo tanto, una vez obtenida la función discriminante, se debe valorar su significancia estadística para probar la siguiente hipótesis nula:
H_0: la variable X_i no es buena para discriminar
H_1: la variable X_i sí es buena para discriminar

Para probar estas hipótesis se calcula la medida de Lambda de Wilks, Λ , también llamada U-estadístico, esto es:

$$\Lambda = \frac{|W|}{|T|}$$

donde:
W : es la matriz de varianza-covarianza combinada. Si W_1 y W_2 son las sumas de cuadrados y productos cruzados calculados para cada uno de los grupos, se tiene: $W = W_1 + W_2$.
T : es la matriz de varianza-covarianza total, es decir, es la suma de cuadrados total de las variables explicativas de la matriz $\boldsymbol{X}$. (Pérez 2004). Esta medida permite calcular el valor del estadístico F de Snedecor, mediante la siguiente igualdad:

$$F_c = \left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n-g}{g-1}\right) = \frac{(1-\Lambda)/(g-1)}{\Lambda/(n-g)}$$

donde:

n : es el número de datos. $n-1$: son los grados de libertad del numerador.

g : es el número de grupos en el análisis. $g-1$ son los grados de libertad del denominador.

El valor de F_c se compara con el valor de F tabular: $F_{t,\alpha,g-1,n-g}$

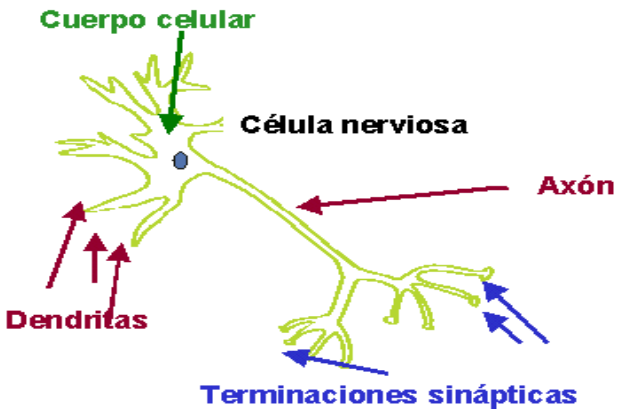
2.2.6 Redes Neuronales

2.2.6.1 Redes Neuronales Artificiales

El cerebro humano tiene uno o varios billones de neuronas que están conectadas. La neurona constituye una unidad del sistema nervioso que está compuesta de tres partes que se diferencian entre sí, que son: el soma o cuerpo celular, las dendritas y el axón. El soma o cuerpo celular es la parte que tiene mayor volumen y contiene al núcleo, desde allí se produce la energía necesaria para el buen funcionamiento de la neurona; las dendritas son las extensiones del cuerpo o soma que reciben los impulsos de otras neuronas que serán enviados al soma o cuerpo celular; finalmente, el axón es la prolongación de la neurona, como se observa en la Figura 2, y se conecta con las dendritas de otras neuronas a través de las sinapsis que vienen a ser las uniones entre las neuronas, su función es conducir el impulso nervioso desde el soma hacia otras neuronas o también hacia otras partes del cuerpo (Fernández, 2016).

Figura 2

Componentes de una neurona



Nota. Fuente: Bioquímica Ambiental por el Ministerio de Educación,2017, http://www3.uah.es/bioquimica/Tejedor/bioquimica_ambiental/tema12/tema%2012-estructura-neurona.htm

Tomando como base el comportamiento de las redes neuronales biológicas, se han logrado construir las denominadas redes neuronales artificiales.

Las redes neuronales artificiales (RNA) o en su defecto ANN por sus siglas en inglés que significan Artificial Neural Networks y de acuerdo a lo señalado por Picón (2011) han adquirido en la actualidad, un elevado interés debido a su empleo como herramientas útiles en la minería de datos debido a su potencia, flexibilidad y facilidad de uso.

Basogain (2009) señala que las Redes Neuronales Artificiales, están inspiradas en las redes neuronales biológicas del cerebro humano y están constituidas por elementos que se comportan de forma similar a la neurona biológica en sus funciones más comunes. Los elementos de una RNA se organizan de una forma muy parecida a la del cerebro humano.

Montaño (2002) señala que las RNA se basan en diversos algoritmos de aprendizaje a partir de datos ya existentes, es

decir, las RNA tienen una capacidad para aprender sobre un conjunto de datos que constituyen patrones de entrenamiento de la red. Los algoritmos de aprendizaje pueden ser supervisados o no supervisados y permiten que las neuronas se agrupen de acuerdo con las ponderaciones o pesos sinápticos a fin de minimizar la función error que surge del proceso de aprendizaje de la red.

Serrano (2009) menciona que la primera estructura neuronal que se usó para problemas de clasificación es la más sencilla. El funcionamiento del perceptrón es sencillo y se basa en comparar la salida del sistema con una señal deseada, de la función de activación se deduce que el perceptrón es un clasificador binario.

Con respecto al modelo estándar de una neurona artificial, Rumelhart y McClelland (1986) y McClelland y Rumelhart (1986) diseñaron algunos principios para describir la i-ésima neurona artificial estándar, mencionado en Larrañaga (1997).

Estos principios consisten en:

- Un conjunto de entradas x_j
- Pesos sinápticos w_{ij} con $j = 1, \dots, n$ son las conexiones entre las neuronas y se optimizan mediante el algoritmo de aprendizaje de la red (Serrano, 2009).
- Una regla de propagación h_i que se define a partir del conjunto de entradas y los pesos sinápticos. Siendo la combinación lineal, la más empleada, entonces la regla de propagación queda de la siguiente forma:

  $$h_i(x_1, \dots, x_n, w_{i1}, \dots, w_{in}) = \sum_{i=1}^{n} w_{ij} x_j$$

  A la sumatoria se le conoce como potencial pos-sináptico al que se le resta un parámetro θ_i denominado umbral. Luego, la regla de propagación resulta de la siguiente forma:

$$h_i(x_1, \dots, x_n, w_{i1}, \dots, w_{in}) = \sum_{i=1}^{n} w_{ij}x_j - \theta_i$$

Si hacemos $\theta_i = w_{i0}$ y $x_0 = -1$ la regla de propagación queda de la siguiente forma:

$$h_i(x_1, \dots, x_n, w_{i1}, \dots, w_{in}) = \sum_{j=0}^{n} w_{ij}x_j$$

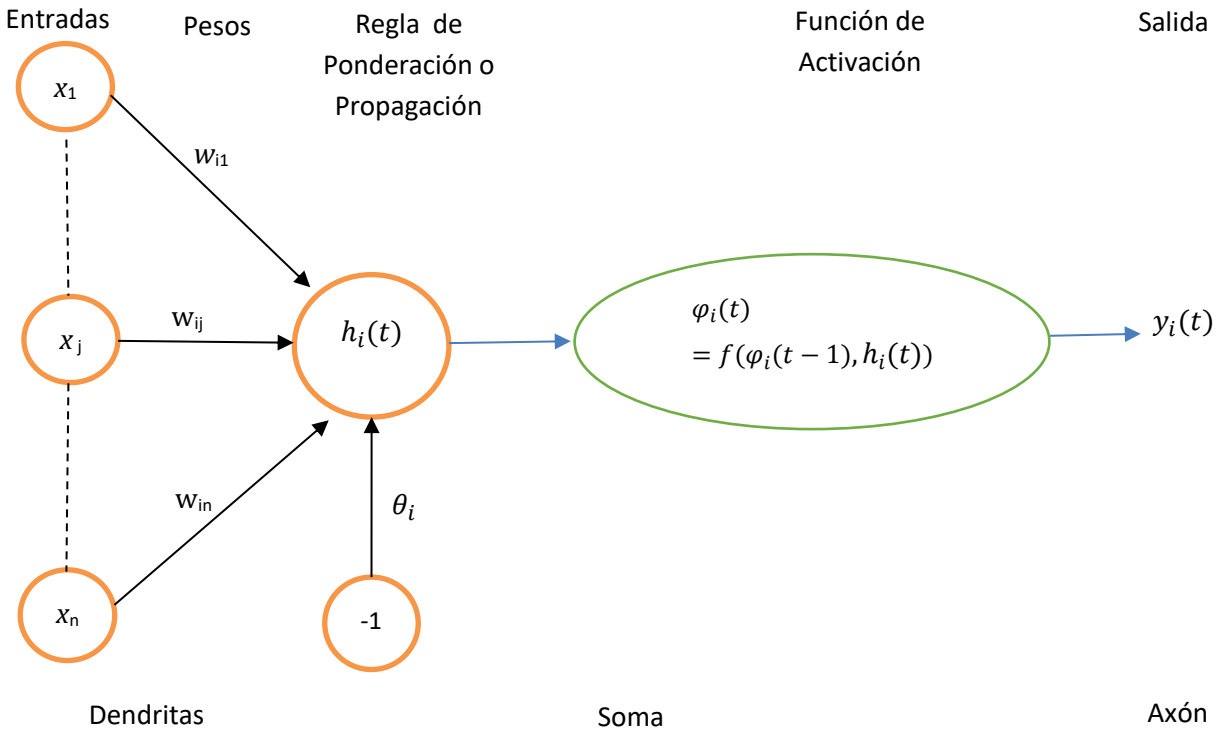
- Una función de activación, que representa la salida de la neurona y su estado de activación. Si la función de activación está dada por φ_i, entonces:

$$\varphi_i = f_i(h_i) = f_i\left(\sum_{j=0}^{n} w_{ij}x_j\right)$$

En la Figura 3, se muestra el funcionamiento de una red neuronal artificial, es decir, la arquitectura de una red neuronal artificial. Donde: $x_j,\ j = 1, 2, \dots, n$ es el conjunto de entradas $\boldsymbol{x}$ o vector de entradas con n componentes. w_{ij} es el conjunto de pesos sinápticos o ponderaciones o de propagación que corresponden a la neurona i, también llamado vector de pesos. θ_i es el umbral que se multiplica por (–1). $\varphi_i(t)$ es la función de activación de la neurona. Finalmente, $y_i(t)$ es el valor de salida de la neurona.

Figura 3

Funcionamiento de una red neuronal artificial.



2.2.6.2 Funciones de activación

Existen diversas funciones de activación o de transferencia que permiten obtener el valor de salida de una neurona. Castro (2006) y Rivera (2005) mencionan que entre las principales funciones de activación se tienen:

1. Función Sigmoidal:

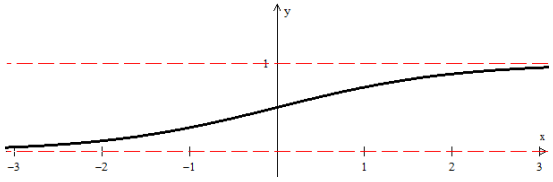
   Su regla de correspondencia está dada por:

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

   La función sigmoidal tiene un valor límite mínimo de cero (0) y máximo igual a uno (1). Su rango es un intervalo continuo dado por $[0;\ 1]$.

Figura 4

Función Sigmoidal



2. Función Tangente Hiperbólica:

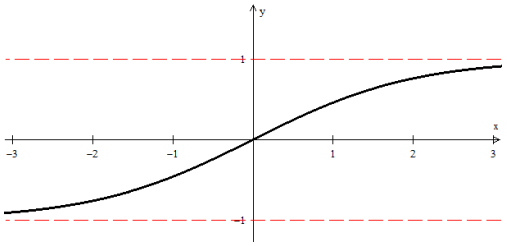
   Con regla de correspondencia dada por:

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

   En cambio, la función tangente hiperbólica tiene un valor límite mínimo de menos uno (–1) y máximo igual a uno (1). Su rango es un intervalo continuo dado por $[-1;\ 1]$.

Figura 5

Función Tangente Hiperbólica



3. Función Secante Hiperbólica:

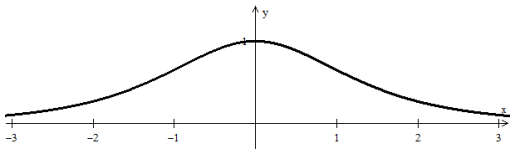
   Su regla de correspondencia está dada por:

$$f(x) = \frac{2}{e^x + e^{-x}}$$

   Cuando los valores de entrada son extremos, esta función asume valores cercanos a cero (0) y cuando los valores de entrada son cercanos a cero, la función asume valores próximos a la unidad (1). Su rango es un intervalo continuo dado por $\langle 0;\ 1]$.

Figura 6

Función Secante Hiperbólica



4. Función Identidad

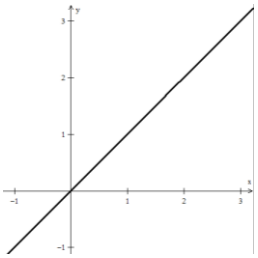
   Con regla de correspondencia dada por:

$$f(x) = x$$

   Esta función toma los mismos valores de la variable de entrada. Su dominio y rango están dados por todos los valores reales.

Figura 7

Función Identidad



5. Función Escalón

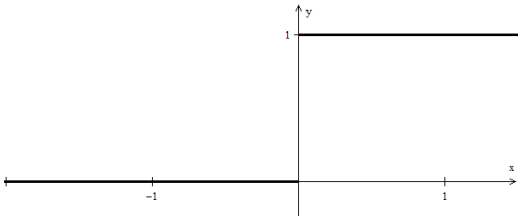
Su regla de correspondencia está dada por:

$$f(x) = \begin{cases} 1 \ ; \ x \geq 0 \\ 0 \ ; \ x < 0 \end{cases}$$

La función escalón asume solo dos valores, un valor de cero (0) y el otro igual a uno (1) dependiendo de los valores positivos o negativos de la variable de entrada. Su rango está dado por el conjunto discreto formado por $\{0, 1\}$.

Figura 8

Función Escalón



2.2.6.3 Arquitectura de una red neuronal artificial

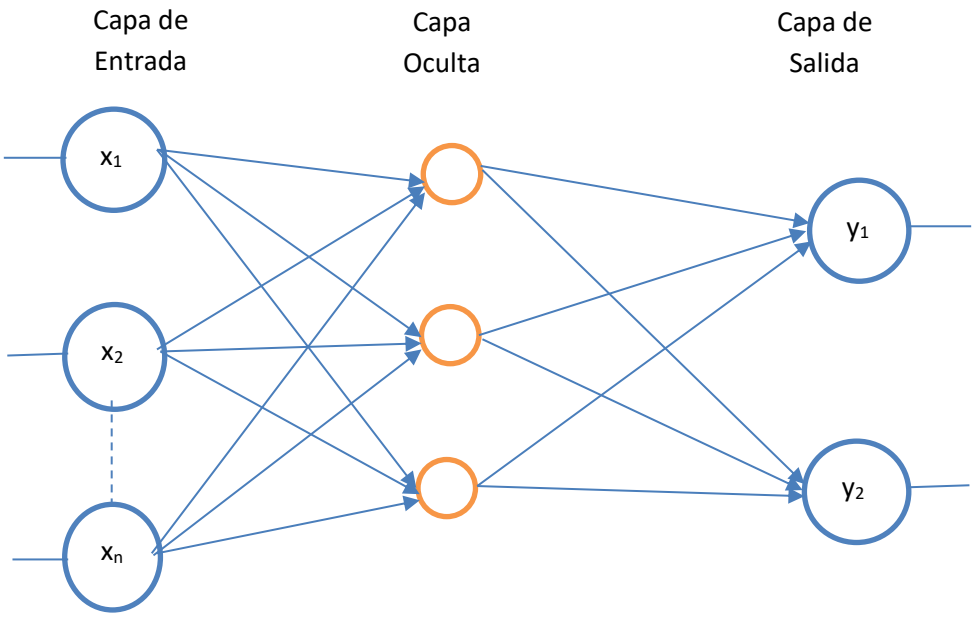
Larrañaga (1997) menciona precisa que la arquitectura es la estructura o patrón de conexión de una red neuronal, esta a su vez tiene nodos que al conectarse entre sí mediante la sinapsis determinan una estructura que determinará el comportamiento de la red.

Acevedo (2017) señala que las neuronas se interconectan formando una red neuronal, de esta forma las neuronas también quedan organizadas en tres capas, como las que se muestran en la Figura 9:

- Capa de entrada que son las que reciben las señales del entorno. Cumplen la tarea de las dendritas en una red biológica.
- Capas ocultas cuyas entradas y salidas se encuentran dentro de la red y se conectan mediante los pesos sinápticos. Se encargan de realizar el trabajo de la red. Hace la tarea del soma en una red biológica.
- Capa de salida que envían el resultado del trabajo realizado por la red hacia el exterior, enviando información a otras neuronas. En una red biológica esta tarea la realiza el axón.

Figura 9

Componentes de una red neuronal artificial.



2.2.6.4 Función de salida de una red neuronal artificial

Rivera (2005) menciona que las neuronas de salida son las que se encargan de dar la salida del sistema y de acuerdo con el aprendizaje que ha adquirido en la red, ofrecerá una respuesta correcta o incorrecta. Cabe señalar que el valor de salida de una neurona indica si está activa o no; además, en una capa oculta,el valor de salida se convierte en el valor de entrada de la siguiente capa oculta, caso contrario, será el valor final para la capa de salida.

2.2.6.5 Tipos de aprendizaje de una red neuronal artificial

Pérez (2007) precisa que el aprendizaje de una red neuronal artificial o también denominado proceso de entrenamiento es la etapa en la cual la red neuronal adquiere la capacidad de realizar las funciones necesarias para resolver un problema. Existen dos tipos principales de aprendizajes de una red neuronal artificial:

- Aprendizaje supervisado. Es el más usado y se caracteriza por que el valor de la salida obtenida o generada por la red es igual a la deseada, es decir, la red minimiza el error entre la salida calculada y la salida deseada. Se emplea en tareas de clasificación y predicción.

- Aprendizaje no supervisado. Se presenta cuando la red sabe la pregunta pero desconoce la respuesta o salida, es decir, la red desconoce el valor de la salida deseada y más bien conoce un conjunto de patrones, con lo cual se tiene clara la organización o estructura final de la red. Se emplea en tareas de cluster y segmentación jerárquica de datos.

2.2.6.6 Redes Neuronales tipo Perceptrón Multicapa

- **Definición**

  Martín (2007) define al perceptrón como un conjunto de neuronas no unidas entre sí, de manera que cada una de las entradas del sistema se conecta a cada neurona produciendo cada una de ellas su salida individual.

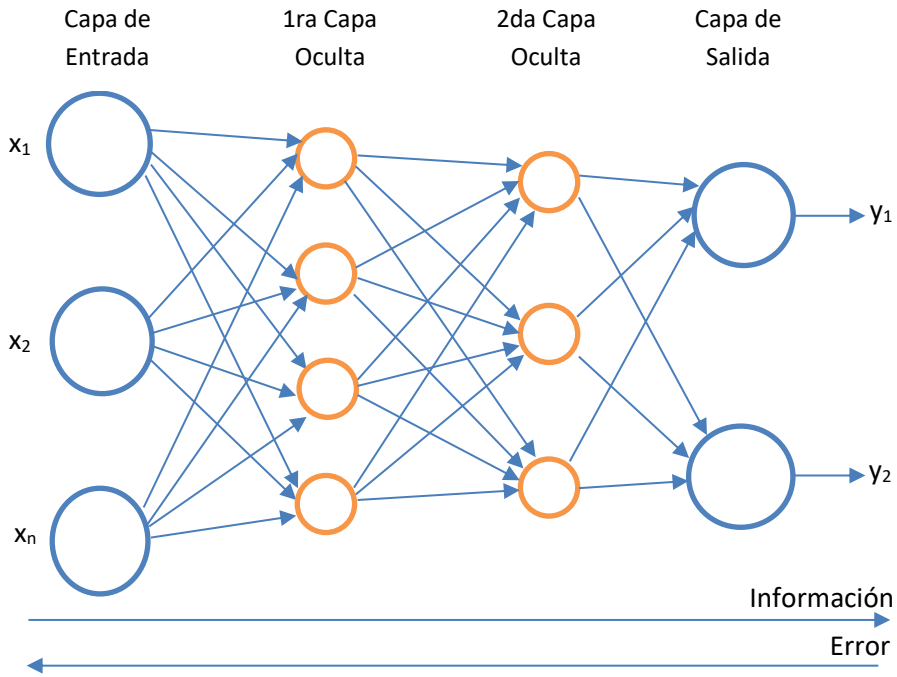
  Sánchez (2006) señala que el perceptrón multicapa, inicialmente desarrollado por P. Werbos (1974) permite resolver el problema de clasificación que no sean linealmente separables. Posee una estructura con al menos una capa oculta y su algoritmo de entrenamiento es del tipo corrección de error. Se basa en el cálculo del gradiente distribuido en los diferentes componentes de la red.

  El mismo autor señala que, las redes neuronales tipo perceptrón multicapa (MLP, por su nombre en inglés Multilayer Perceptron) tiene tres características: el modelo de cada neurona en la red incluye una función de activación no lineal, la red contiene una o más capas ocultas y la red presenta altos grados de conectividad que se determinan por las sinapsis de la propia red.

  En la Figura 10, se observa que la señal de información está constituida por el flujo de información que va desde la capa de entrada hacia la capa de salida. Así también, la señal de error está constituida por el flujo de error que va desde la capa de salida hacia la capa de entrada.

Figura 10

Estructura de una red neuronal tipo perceptrón multicapa



Nota. Tomado de Sánchez (2006).

El MLP se ha empleado satisfactoriamente para resolver problemas muy diversos y difíciles a través del algoritmo conocido como retropropagación que consta de dos etapas:

a) Etapa hacia adelante: en donde se fijan los parámetros de la red y se presenta una señal de entrada a la red, que se propaga hacia adelante para producir la salida.
b) Etapa hacia atrás: El error entre la salida deseada y la red se propaga hacia atrás. Los parámetros de la red se modifican para minimizar el cuadrado de dicho error.

Muñoz (2015) propone que para la construcción de la red neuronal se deben tomar en cuenta los siguientes aspectos: arquitectura, número de neuronas de la capa de entrada, número de neuronas de la capa de salida,

número de capas ocultas, número de neuronas de la capa oculta, valor neto del perceptrón, función de transferencia, factor de aprendizaje, producir regla de aprendizaje y algoritmo de aprendizaje propagación hacia atrás.

Teniendo como base lo descrito por Muñoz (2015) se presentan las siguientes características del modelo de redes neuronales tipo perceptrón multicapa:

- **Arquitectura de la red**

  La red está conformada por tres capas: la capa de entrada, la capa de salida y por lo general una sola capa oculta. Larrañaga, (1997) precisa que la arquitectura de una red neuronal tipo perceptrón multicapa se denomina red de retropropagación debido a que en esta red, el perceptrón multicapa suele entrenarse mediante el algoritmo de retropropagación de errores.

- **Número de neuronas**

  Marín (2012) precisa que el número de neuronas de entrada está en función al número de variables independientes que se incluyen en el estudio. El número de neuronas de la capa de salida está en función del número de categorías que tenga la variable de estudio. La capacidad de aprendizaje de una red está determinadas por el número de neuronas ocultas. Finalmente, se debe señalar que por lo general, las redes neuronales tipo perceptrón multicapa cuenta con solo una capa oculta. El número de neuronas en la capa oculta se puede calcular siguiendo la regla de la pirámide geométrica

presentado por Pano (2015) mediante la siguiente relación:

$$h = \sqrt{m.n}$$

donde h es el número de neuronas en la capa oculta, m es el número de neuronas en la capa de salida y n es el número de neuronas en la capa de entrada. Se debe cumplir con la siguiente condición:

$$n > h$$

- **Valor neto del perceptrón**

  Muñoz (2015) expresa el valor neto del perceptrón como la suma de los productos de cada conexión por los pesos aleatorios y el balance, es decir:

$$n = \sum_{i=0}^{i=q} (w_i p_i + b)$$

  Donde w_i son los pesos asociados a cada conexión en la red neuronal, p_i son las neuronas en la capa de entrada y b es el balance de la red neuronal.

- **Función de transferencia**

  La función de transferencia o de activación es la encargada de generar la señal excitadora (valores muy altos) o inhibidora (valores muy pequeños) en las capas ocultas y en la capa de salida de la red. Para Pérez (2013), con frecuencia, las redes neuronales tipo perceptrón multicapa emplea las funciones de activación sigmoidal y tangente hiperbólica. Como se muestran en las figuras 5 y 6, ambas funciones tienen un rango en un intervalo continuo de valores que son: [0;1] y [–1; 1] respectivamente. Cuando las neuronas de salida se diferencian del resto de neuronas, emplean otro tipo de función de activación, siendo las

más usadas la función identidad y la función escalón, mostradas en las figuras 7 y 8.

- **Factor de aprendizaje**

  Según señala Muñoz (2015) el factor de aprendizaje es la velocidad de aprendizaje de la red y su valor debe estar comprendido entre 0,4 y 1.

- **Regla de aprendizaje**

  Es la regla de corrección de error para el aprendizaje supervisado. La regla consiste en minimizar el error cuadrado total. Así también, Mateo (2012) citado por Muñoz (2015), señala que el aprendizaje de la red neuronal se basa en el algoritmo de retropropagación (backpropagation).

  Pérez (2013) señala que la regla o algoritmo de aprendizaje es el mecanismo que permite la modificación de los parámetros. Este algoritmo de aprendizaje es el que emplea el perceptrón multicapa, entonces, la regla de aprendizaje se formula como un problema de minimización del error, para tal fin se busca que la diferencia entre los valores de salida de la red y los valores de salida deseados, esto es:

  $$Min_W\ E$$

  Donde W es el conjunto de parámetros de la red, considerando los pesos y umbrales; así también E es la función error, tal que:

  $$E = \frac{1}{N}\sum_{n=1}^{N} e(n)$$

donde :

N = números de patrones o muestra

$e(n)$ = error cometido por la red para el patrón n dado por:

$$e(n) = \frac{1}{2}\sum_{i=1}^{n_c}(S_i(n) - y_i(n))^2$$

$S_i(n)$ es la salida obtenida y

$y_i(n)$ es la salida deseada

- **Algoritmo de aprendizaje propagación hacia atrás tomado de Muñoz (2015)**

  i. Inicialización: Se inicializan los pesos y balances de cada conexión con valores aleatorios que van en un rango de 0 y 1.
  ii. Muestra de entrenamiento: Elegir un patrón de entrada de los datos $X^{(\mu)}$ a ser analizados. Cada variable ingresará a una neurona en la capa de entrada.
  iii. Fase hacia adelante: propagar la señal hacia adelante a través de la red calculando las entradas netas y función de activación de cada perceptrón en la red neuronal. Calcular el error de aprendizaje δ_i^L, el error de aprendizaje en la capa resultante $o_1 = y_i^L$ que es la diferencia de los valores observados y los valores esperados.

$$\delta_i^L = a'(h_i^L)[\, d_i^\mu - y_i^L]$$

donde: h_i^L es el input a la red en la i-ésima unidad en la l-ésima capa y a' es la derivada de la función de activación a.

iv. Fase hacia atrás: el error de aprendizaje determina la actualización de los pesos que al final del entrenamiento producirá una respuesta satisfactoria. Se calculan los valores para las capas precedentes propagando los errores hacia atrás.

$$\delta_i^L = a'(h_i^l) \sum_i w_{ij}^{l+1} \delta_j^{l+1}$$

para: $l = (L-1)$ capas.

Cada peso se modifica utilizando la siguiente función:

$$\Delta w_{ji}^l = n\delta_i^l y_j^{l-1}$$

El método de aprendizaje es el gradiente descendiente que es la variación que se produce en el tiempo del valor de los pesos, propagando el error desde la capa final hasta la capa de entrada. Siendo n el factor de aprendizaje que determina la velocidad con la que aprende la red neuronal.

$$w_{ji}^l = w_{ji}^l + \Delta w_{ji}^l$$

v. Iteración: se repiten los pasos 3 y 4 con la muestra de entrenamiento hasta que el algoritmo ajuste los pesos de la red para minimizar el error cuadrático medio del error de aprendizaje hasta un decrecimiento menor a 10^{-6}.

2.2.7 Evaluación de la clasificación

Para establecer cuál de los métodos multivariantes estudiados, ofrece una mejor capacidad discriminatoria del desempeño académico, es decir, para determinar la mejor calidad de clasificación, se emplearán los siguientes indicadores: Sensibilidad, Especificidad y la Curva ROC.

2.2.7.1 Sensibilidad (S):

Marín (s.f.) define la Sensibilidad como la predicción de éxito cuando realmente lo es, es decir:

$$Sensibilidad = S = P(\hat{y} = 1 \,/ y = 1)$$

Si se considera al estudiante aprobado en el primer curso de matemática, entonces, se puede definir la Sensibilidad como la probabilidad de que el método lo clasifique aprobado cuando realmente aprobó. Madrid (2017), presentó una Matriz de Contingencia o de Confusión para evaluar el desempeño de los modelos multivariantes, que se muestra en la Tabla 6.

Tabla 6

Matriz de Contingencia

Clasificación		**Pronosticada**	
		Aprobado	Desaprobado
Observada	Aprobado	VP	FN
	Desaprobado	FP	VN

Nota. Adaptado de Madrid (2017).

donde:

VP es el número de observaciones para estudiantes aprobados clasificados como aprobados, es decir, es el valor Verdadero Positivo.

FN es el número de observaciones para estudiantes aprobados clasificados como desaprobados, es decir, es el valor Falso Negativo.

FP es el número de observaciones para estudiantes desaprobados clasificados como aprobados, es decir, es el valor Falso Positivo.

VN es el número de observaciones para estudiantes desaprobados clasificados como desaprobados, es decir, es el valor Verdadero Negativo.

Con la información de la Matriz de Contingencia mostrada en el Cuadro 4, se puede hallar la precisión en la clasificación de los estudiantes aprobados, es decir, la sensibilidad, que será igual a:

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

Cuanto más alto (cercano a uno) sea el valor de la Sensibilidad (S), mucho mejor será la calidad de la clasificación realizada con el modelo.

Además,

$$1 - S = 1 - \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} = \frac{\mathrm{VP} + \mathrm{FN} - \mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} = \frac{\mathrm{FN}}{\mathrm{VP} + \mathrm{FN}}$$

Es la fracción de Falsos Negativos o Falsos Desaprobados, es decir, de los estudiantes clasificados como desaprobados cuando estaban aprobados.

2.2.7.2 Especificidad (E):

Marín (s.f.) define la Especificidad como la predicción de un fracaso cuando realmente lo es, es decir:

$$Especificidad = E = P(\hat{y} = 0 \,/ y = 0)$$

Si se considera al estudiante desaprobado en el primer curso de matemática como la categoría común o específica, entonces se puede definir la Especificidad como la probabilidad de que la prueba lo clasifique como desaprobado cuando realmente lo es y de acuerdo a la Matriz de Contingencia; se obtiene:

$$Especificidad = E = \frac{VN}{\mathrm{VN} + \mathrm{FP}}$$

Es la fracción de Verdaderos Negativos o verdaderos desaprobados. Cuanto más alto (cercano a uno) sea el valor de la Especificidad (E), mucho mejor será la calidad de predicción del modelo. Además, los Falsos Negativos son lo opuesto a la Especificidad y se le denota por $(1 - E)$, donde:

$$1 - E = 1 - \frac{VN}{\mathrm{VN} + \mathrm{FP}} = \frac{VN + FP - VN}{\mathrm{VN} + \mathrm{FP}} = \frac{FP}{\mathrm{VN} + \mathrm{FP}}$$

En la Figura 11 se muestra la relación entre los dos indicadores.

Figura 11

Relación entre Sensibilidad y Especificidad



2.2.7.3 Curva ROC:

Según lo explica Cuadras (2018), el nombre de esta curva proviene de las iniciales Receiver Operating Characteristic y representa, en un mapa o en un sistema de coordenadas, la relación entre la Sensibilidad (S) y la Especificidad (E). En el eje horizontal o de abscisas de ubican los valores entre cero y la unidad, correspondientes a los valores de, uno menos la Especificidad, es decir, $(1 - E)$ y en el eje vertical o de ordenadas se ubican los valores entre cero y uno correspondientes a los valores de la Sensibilidad denotado por (S). Para diferentes criterios de decisión o puntos de corte o valores umbral, se obtendrán diferentes puntos de la forma $(1 - E;\ S)$, al unirse estos puntos se dará origen a la denominada curva ROC. El análisis de la curva ROC incluye el cálculo del área bajo la curva ROC denominada AUC por sus siglas en inglés de Area Under Curve. El área bajo la curva es un estadístico por excelencia que mide la capacidad discriminante de una prueba y compara pruebas entre sí y determinar cuál es el más eficaz (Del Valle, 2017). el valor del área también se encuentra entre cero y uno. Así entonces, cuanto más cercano a uno sea el área AUC, mejor será el método empleado, es decir, discriminará mejor entre estudiantes aprobados y desaprobados en el primer curso de matemática.

Empleando el método trapezoidal, se hallará el área AUC mediante la siguiente fórmula planteada por Del Valle (2017):

$$AUC = \frac{1}{2}\sum_{t=1}^{T}(FFP_t - FFP_{t-1})\ .\ (FVP_t + FFP_{t-1})$$

donde: FFP_t y FVP_t son las fracciones Falsos Positivos y Verdaderos Positivos respectivamente, calculadas para diferentes puntos de corte $t = 1,2,\dots,T$.

Para determinar la exactitud o eficacia del método empleado, se tomará en cuenta la clasificación de la Tabla 7.

Tabla 7

Tabla de Clasificación. Área bajo la curva ROC

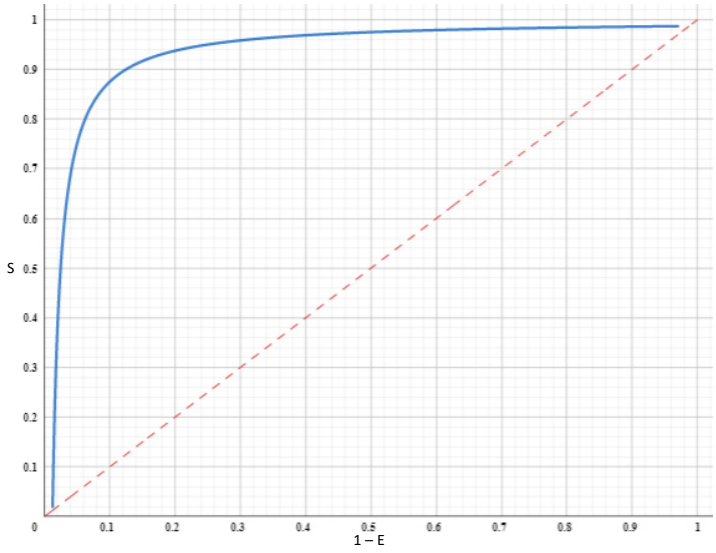
AUC	Descripción
$0{,}5 \leq AUC < 0{,}7$	Baja exactitud
$0{,}7 \leq AUC < 0{,}9$	Exactitud moderada
$0{,}9 \leq AUC \leq 1$	Alta exactitud

Nota. Adaptado de Del Valle (2017)

En la Figura 12, según señala Del Valle (2017), cuando la curva ROC está más cercana al punto (0;1), indica que los dos grupos están bien diferenciados. En cambio, si la curva ROC está más alejada del punto (0;1), indicará que la distribución de los grupos presenta solapamiento y el modelo no será muy útil para clasificarlos.

Figura 12

Curva ROC



CAPÍTULO 3: METODOLOGÍA

3.1 Hipótesis general

El método multivariante que analiza mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM es la regresión logística binaria.

3.2 Hipótesis específicas

- El método multivariante de regresión logística binaria es mejor que el análisis discriminante para analizar el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM.
- El método multivariante de regresión logística binaria es mejor que el de redes neuronales tipo perceptrón multicapa para analizar el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM.

3.3 Identificación de variables

Para el presente estudio de investigación se ha considerado como variable independiente a los métodos multivariantes: regresión logística binaria, análisis discriminante y redes neuronales tipo perceptrón multicapa. Las variables dependientes son: tasa de clasificación correcta, sensibilidad, especificidad y área bajo la curva ROC. En la Tabla 8 se muestran las variables intervinientes en este estudio.

Tabla 8.
Descripción de Variables intervinientes

Factores	Variable	Nombre	Descripción
Socio-demográfico	X1	Edad	Edad del estudiante al momento de la inscripción al Concurso de admisión de la UNALM.
	X2	Sexo	Con X2=1 si es mujer y X3=0 si es hombre.
	X3	Procedencia	Lugar de procedencia del estudiante, con X3=1 si es provincia de Lima y Callao y X3=0 si son otras provincias o departamentos.
Escolares	X4	Tipo de colegio	Tipo de colegio en donde el estudiante culminó sus estudios secundarios. Con X5=1 si es público y X5=0 si es privado.
	X5	Ubicación del colegio	Ubicación del colegio donde el estudiante culminó sus estudios secundarios. Con X6=1 si es provincia de Lima y Callao y X6=0 si son otras provincias o departamentos.
Admisión	X6	Carrera	Carrera profesional a la cual el estudiante ingresó. Con X7=1 si es Biología, Ingeniería Ambiental e Industrias Alimentarias; y X7=0 si el estudiante es de Agronomía, Meteorología, Ingeniería Forestal, Economía, Estadística Informática, Ingeniería en Gestión Empresarial, Ingeniería Agrícola, Zootecnia y Pesquería.
	X7	Modalidad	Modalidad de ingreso a la universidad Con X8=1 Concurso ordinario y X8=0 para otras modalidades de ingreso.

	X8	Opción	Orden de las preferencias u opciones de la carrera profesional elegidas al momento de postular y a la cual el estudiante ha ingresado y es una de las tres opciones marcadas en su postulación.
	X9	Matemática	Puntaje obtenido por el estudiante en el área de matemática del examen de admisión. Comprende las áreas de Aritmética, Álgebra, Geometría y Trigonometría.
Admisión	X10	Razonamiento Matemático	Puntaje obtenido por el estudiante en el área de Razonamiento Matemático del examen de admisión.
	X11	Razonamiento Verbal	Puntaje obtenido en el área de Razonamiento Verbal del examen de admisión.
	X12	Física	Puntaje que obtuvo el estudiante en el área de Física del examen de admisión.
	X13	Química	Puntaje obtenido por el estudiante en el área de Química del examen de admisión.
	X14	Biología	Puntaje que obtuvo el estudiante en el área de Biología del examen de admisión.
	X15	Puntaje general	Puntaje general de ingreso obtenido en el examen de admisión.
	X16	Orden de mérito	Orden de mérito de ingreso a la UNALM. Con X16=1 si corresponde a la mitad superior y X16=0 si corresponde a la mitad inferior.

	Y1	Resultado	Condición final en el curso de matemática. Con Y1=1, si es aprobado y Y1=0 si es desaprobado.
Desempeño académico	Y2	Créditos Aprobados	Condición final del estudiante de acuerdo a los créditos aprobados. Con Y2=1 si aprobó al menos 12 créditos cursados y Y2=0 si aprobó menos de 12 créditos cursados[a].*

[a] De acuerdo a la Ley Universitaria vigente, Artículo 99: un estudiante regular es aquel que se encuentra matriculado en un número mínimo de 12 créditos.

Tabla 9

Resumen de la Información

Registro de datos	Dimensiones	Variables independientes	Variables dependientes
553	3	16	3

El desempeño académico de los estudiantes en el curso de matemática durante el primer ciclo de estudios será medido a través de tres variables: resultado, créditos aprobados y cursos desaprobados.

La variable Resultado, se midió mediante el promedio final que se evidencia en las actas de notas. Esta variable dependiente será dicotomizada, es decir, 1: promedio final aprobado (mayor o igual a once) y 0: promedio final desaprobado (menor de once).

La variable Créditos Aprobados mide el porcentaje de créditos aprobados por los estudiantes en el primer ciclo universitario. También será dicotomizada, siendo 1: si aprueba doce (12) o más créditos cursados y se asigna el valor de 0: si aprueba menos de 12 créditos.

3.4 Tipo y Diseño de Investigación

De acuerdo a lo que señala Hernández, Fernández & Baptista (2006), el presente estudio es observacional, no experimental porque se realiza sin manipulación de variables, es decir observándolos tal como se presentan en su contexto para su posterior análisis de los datos.

Este estudio es de tipo descriptivo y correlacional debido a que los datos se analizan con la finalidad de obtener un patrón de características de los estudiantes ingresantes que aprueban el primer curso de matemática en la UNALM, asimismo este estudio busca encontrar el mejor modelo de clasificación de estudiantes ingresantes y que aprueban el primer curso de matemática.

De acuerdo a lo señalado por los mismos autores, este estudio es de corte transversal ya que se realizó en un solo momento específico de tiempo.

Los resultados encontrados se soportan en el acopio de la información obtenida a partir de los datos de la Oficina de Estudios y Registros Académicos de la UNALM, del Centro de Promoción y Admisión de la UNALM, así como de las actas de evaluación final del curso de matemática.

3.5 Unidad de Análisis

En el presente estudio la unidad de análisis está representada por un estudiante ingresante, es decir, de primer ciclo académico y matriculado en el curso de matemática en la UNALM en los ciclos académicos 2017-I y 2017-II.

3.6 Población de estudio

La población estuvo conformada por todos los estudiantes ingresantes y matriculados en el curso de matemática en el primer ciclo académico de la UNALM comprendidos entre los ciclos 2017-I y 2017-II. Se tomó como factor de inclusión a los estudiantes que ingresaron bajo las siguientes modalidades de ingreso Concurso de Admisión Ordinario, Primeros Puestos de Colegio, Colegio Mayor COAR, Ley 29973 Personas con discapacidad, Ley 28592 Plan Integral de Reparaciones Todos los postulantes de las modalidades mencionadas rinden la prueba de admisión que consta de 100 preguntas que corresponden a Razonamiento verbal (21), Razonamiento matemático (21), Matemática (16), Física (14), Química (14) y Biología (14), la prueba tiene una duración de 3 horas.

Se excluyen del presente estudio a los estudiantes que ingresaron a la UNALM bajo las modalidades de Traslado Externo y Graduados y Titulados universitarios, debido a que los primeros rinden una prueba especial de selección que consta de 42 preguntas de Razonamiento verbal (21) y Razonamiento matemático (21), la prueba tiene una duración de 100 minutos. En tanto, los Graduados y Titulados universitarios no rinden ninguna prueba de selección, a menos que existan más postulantes que vacantes, y en dicho caso, los postulantes rinden una prueba especial en la misma facultad a la cual postulan.

Para el presente estudio no se obtuvo una muestra, es decir, se tomó en cuenta a todos los estudiantes ingresantes en los ciclos 2017-I y 2017-II y matriculados en el curso de matemática, que pertenecían a una de las doce carreras profesionales de la universidad y que cumplieron con los criterios de inclusión antes descritos.

El presente estudio se contó con un total de 553 registros de datos correspondientes a 300 y 253 estudiantes ingresantes a la UNALM

mediante las modalidades mencionadas y matriculados en el curso de matemática en los ciclos 2017-I y 2017-II respectivamente.

Hair (1999) señala que se debe contar con una muestra de validación para que la evaluación del modelo sea independiente de la muestra de entrenamiento. En este estudio se construyeron tres modelos que fueron obtenidos y validados empleando método de división aleatoria de datos 70:30 propuesto por Fernández (2018), de esta forma, los datos fueron divididos en una muestra de entrenamiento de 210 y 177 estudiantes y para la de validación fue 90 y 76 estudiantes ingresantes a la UNALM en los ciclos académicos 2017 I y 2017 II, respectivamente.

3.7 Técnicas de recolección de datos

Se utilizaron los datos de los estudiantes ingresantes de primer ciclo académico de la UNALM que se matricularon en el curso de matemática. Los datos corresponden a: edad, lugar de procedencia, tipo de colegio, ubicación del colegio, carrera profesional a la cual ingresó, modalidad de ingreso, orden de preferencia (opciones) de la carrera profesional a la cual ingresó y los puntajes obtenidos en las diferentes áreas temáticas del examen de admisión y el orden de mérito de ingreso a la universidad. Los datos fueron obtenidos a partir del acopio de la información de los datos de la Oficina de Estudios y Registros Académicos y del Centro de Admisión y Promoción de la UNALM.

Los datos para las variables en estudio se obtuvieron de la Oficina de Estudios y Registros Académicos de la UNALM, así como de las Actas de evaluación final del curso de matemática. Los datos corresponden a los ciclos académicos 2017 I y 2017 II.

3.8 Análisis e interpretación de la información

3.8.1 Clasificación, registro y codificación de los datos

Los datos serán procesados a través del software estadístico IBM SPSS Statistics versión 25.0 para Windows (versión de prueba), empleando el Análisis de Regresión Logística Binaria, Análisis Discriminante y las Redes Neuronales tipo Perceptrón Multicapa.

3.8.2 Técnicas estadísticas

Para el análisis de la información y comprobación de las hipótesis se emplearon tres técnicas estadísticas: Regresión Logística Binaria, Análisis Discriminante y Redes Neuronales tipo Perceptrón Multicapa.

Para establecer cuál de los métodos multivariantes estudiados, ofrece una mejor capacidad discriminatoria del desempeño estudiantil, se emplearon los siguientes indicadores: Tasa de clasificación correcta, Sensibilidad, Especificidad y Área bajo la curva ROC.

Los tres modelos de clasificación se compararon llevando a cabo los siguientes pasos:

1. Se estableció una muestra de entrenamiento con el 70% de los datos.
2. Con la muestra de entrenamiento se construyeron los modelos de regresión logística binaria, discriminante y de redes neuronales tipo perceptrón multicapa.
3. Con la muestra de validación se realización las nuevas clasificaciones.
4. Se comparan las clasificaciones obtenidas empleando los indicadores o criterios de evaluación.

CAPÍTULO 4: RESULTADOS Y DISCUSIÓN

En el presente estudio se utilizó el paquete estadístico SPSS 25.0 para la construcción de los modelos de clasificación y también para la validación y comparación de los modelos obtenidos.

Con los datos de los estudiantes se construyeron tres modelos estadísticos que permitieron la clasificación del desempeño de los estudiantes del primer ciclo académico en la UNALM.

Los tres modelos fueron comparados con los indicadores que permitieron determinar el modelo que mejor clasifica el desempeño de estudiantes de los ciclos académicos 2017 I y 2017 II.

Para el desarrollo del este estudio se emplearon las variables intervinientes que se muestran en la Tabla 10.

Tabla 10

Tipo de variables intervinientes en el estudio

Factor	Variable	Nombre	Tipo
Socio-demográfico	X1	Edad	Cuantitativo
	X2	Sexo	Cualitativo
	X3	Procedencia	Cualitativo
Educativo	X4	Tipo de colegio	Cualitativo
	X5	Ubicación del colegio	Cualitativo
Admisión	X6	Carrera	Cualitativo
	X7	Modalidad	Cualitativo
	X8	Opción	Cualitativo
	X9	Matemática	Cuantitativo
	X10	Razonamiento Matemático	Cuantitativo
	X11	Razonamiento Verbal	Cuantitativo
	X12	Física	Cuantitativo
	X13	Química	Cuantitativo
	X14	Biología	Cuantitativo
	X15	Puntaje general	Cuantitativo
	X16	Orden de mérito	Cualitativo
Desempeño	Y1	Resultado	Binario
	Y2	Créditos Aprobados	Binario

En resumen, se analizaron 8 variables cuantitativas y 8 variables cualitativas. Las variables de estudio fueron dos del tipo binario.

4.1 Estimación de los modelos de Regresión Logística Binaria

Con los datos de los estudiantes se construyó un modelo de regresión logística binaria para cada una de las dos variables de estudio correspondientes al desempeño académico de los estudiantes ingresantes y matriculados en el primer ciclo académico en la UNALM. El tamaño de muestra de entrenamiento fue de 210 y 177 para los ciclos 2017 I y II respectivamente. La muestra de validación fue de 90 y 76 para los ciclos 2017 I y II respectivamente.

Si bien la regresión logística no exige el cumplimiento de supuestos, es conveniente que las variables presenten una distribución normal multivariante y ausencia de multicolinealidad. Para tal efecto se emplearon los resultados de la evaluación de supuestos del análisis discriminante.

Las variables dependientes categóricas binarias son: Resultado y Créditos Aprobados. Se eligieron las siguientes variables independientes métricas: Edad, Matemática, Razonamiento Matemático, Razonamiento Verbal, Física, Química, Biología. Las variables independientes categóricas son: Opción, Orden de Mérito, Sexo, Tipo de Colegio, Ubicación del Colegio, Carrera profesional, Modalidad de ingreso y Procedencia del estudiante.

Las variables se seleccionaron por el método de pasos hacia adelante (Wald). El modelo inicia con el término constante y se van agregando las variables de acuerdo con su relación con el Resultado.

4.1.1 Modelo de regresión logística binaria para Resultado 2017 I

Tabla 11

Variables en el modelo logístico para Resultado. Ciclo 2017 I

Variable	B	Error estándar	Wald	gl	Sig.	Exp(B)
Matemática	0,168	0,057	8,779	1	0,003	1,183
Física	0,261	0,060	19,120	1	0,000	1,298
Química	0,211	0,061	11,825	1	0,001	1,235
Constante	-8,058	1,437	31,439	1	0,000	0,000

De la Tabla 11 se observa que los coeficientes de las variables Matemática, Física y Química son significativos para el modelo propuesto. Así también de los coeficientes mostrados en la tabla, se puede deducir que por cada incremento en la nota obtenida en Matemática en el ciclo 2017 I, el logaritmo de los odds para la aprobación del curso de matemática en comparación con la desaprobación aumenta en 0,168. Asimismo, cada incremento en la nota obtenida en Física, el logaritmo de los odds para la aprobación del curso de matemática en comparación con la desaprobación aumenta en 0,261. Finalmente, por cada incremento en la nota obtenida en Química, el logaritmo de los odds para la aprobación del curso de matemática en comparación con la desaprobación aumenta en 0,211.

Así, el modelo logístico propuesto para la variable Resultado de los estudiantes en el primer curso de matemática en el ciclo 2017 I de la UNALM, está dado por:

$$Prob(Y_1 = 1) = \frac{1}{1 + e^{8{,}058 - 0{,}168 * \text{Matemática} - 0{,}261 * \text{Física} - 0{,}211 * \text{Química}}}$$

Se realizaron las pruebas Ómnibus para probar la hipótesis nula sobre la significancia de los coeficientes del modelo:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

Tabla 12

Prueba Ómnibus del Modelo Logístico para Resultado. Ciclo 2017 I

Descripción	Chi-cuadrado	gl	Sig.
Modelo	56,995	3	0,000

En la Tabla 12 se muestra que el valor Chi-cuadrado es 56,995 con p_valor igual a 0,000, esto permite concluir que existe una relación estadísticamente significativa entre la variable dependiente Resultado del primer curso de matemática y las variables independientes Matemática, Física y Química que corresponden a las notas que obtuvieron los estudiantes en el examen de admisión 2017 I en la UNALM.

Se realizaron las pruebas de Hosmer y Lemeshow para probar la hipótesis sobre la bondad de ajuste del modelo:

H_0: No existen diferencias entre los Resultados observados y pronosticados.

H_1: Existen diferencias entre los Resultados observados y pronosticados.

Tabla 13

Prueba Hosmer y Lemeshow. *Modelo Logístico para Resultado. Ciclo 2017 I*

Descripción	Chi-cuadrado	gl	Sig.
Modelo	12,257	8	0,140

De la Tabla 13 se observa que el valor Chi-cuadrado es 12,257con un p_valor igual a 0,140, esto nos indica que se acepta la hipótesis nula, en conclusión, no existen diferencias entre los Resultados observados y pronosticados para los dos grupos de estudiantes aprobados y desaprobados en el primer curso de matemática en la UNALM. Finalmente, el modelo ajustado es significativo.

Tabla 14

Resultados de clasificación para Resultado. Ciclo 2017 I

			Resultado Pronosticado		
			0	1	Total
Resultado observado	N	0	26	38	64
		1	20	126	146
	%	0	40,6	59,4	100,0
		1	13,7	86,3	100,0

De la Tabla 14 se observa que el modelo logístico propuesto aplicado en la muestra de entrenamiento genera un 72,4% de clasificación correcta de estudiantes, de los cuales se clasificó correctamente al 86,3% de los estudiantes que aprobaron y al 40,6% de los estudiantes que desaprobaron el primer curso de matemática en la UNALM en el ciclo 2017 I. Esto significa que el modelo logístico es bueno para clasificar a los estudiantes que aprueban o no el primer curso de matemática en la UNALM.

4.1.2 Modelo de regresión logística binaria para Resultado 2017 II

Tabla 15

Variables en el modelo logístico para Resultado. Ciclo 2017 II

Variable	B	Error estándar	Wald	gl	Sig.	Exp(B)
Matemática	0,184	0,046	15,641	1	0,000	1,202
Modalidad	21,088	14536,086	0	1	0,999	1439742371
Constante	-1,950	0,535	13,286	1	0,000	0,142

De la Tabla 15 se observa que los coeficientes de las variables Matemática y Modalidad de ingreso son significativos para el modelo propuesto. Así también de los coeficientes mostrados en la tabla, se puede deducir que por cada incremento en la nota obtenida en Matemática en el ciclo 2017 II, el logaritmo de los odds para la aprobación de al menos 12 créditos en comparación con la aprobación de menos de 12 créditos aumenta en 0,184. Asimismo, por cada cambio de modalidad de ingreso de concurso ordinario a otro tipo de modalidad, el logaritmo de los odds para la aprobación de al menos 12 créditos en comparación con la aprobación de menos de 12 créditos aumenta en 21,088.

Así, el modelo logístico propuesto para la variable Resultado de los estudiantes en el primer curso de matemática en el ciclo 2017 II de la UNALM, está dado por:

$$Prob(Y_1 = 1) = \frac{1}{1 + e^{1,950 - 0,184 * \text{Matemática} - 21,088 * \text{Modalidad}}}$$

Se realizaron las pruebas Ómnibus para probar la hipótesis nula sobre la significancia de los coeficientes del modelo:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

Tabla 16

Prueba Ómnibus del Modelo Logístico para Resultado. Ciclo 2017 II

Descripción	Chi-cuadrado	gl	Sig.
Modelo	26,914	2	0,000

En la Tabla 16 se muestra que el valor Chi-cuadrado es 26,914 con p_valor igual a 0,000, esto permite concluir que existe una relación estadísticamente significativa entre la variable dependiente Resultado del primer curso de matemática y las variables independientes Matemática y Modalidad de ingreso en el examen de admisión 2017 II en la UNALM.

Se realizaron las pruebas de Hosmer y Lemeshow para probar la hipótesis sobre la bondad de ajuste del modelo:

H_0: No existen diferencias entre los Resultados observados y pronosticados.

H_1: Existen diferencias entre los Resultados observados y pronosticados.

Tabla 17

Prueba Hosmer y Lemeshow. *Modelo Logístico para Resultado. Ciclo 2017 II*

Descripción	Chi-cuadrado	gl	Sig.
Modelo	8,646	7	0,279

De la Tabla 17 se observa que el valor Chi-cuadrado es 8,646 con un p_valor igual a 0,279, esto nos indica que se acepta la hipótesis nula,

en conclusión, no existen diferencias entre los Resultados observados y pronosticados para los dos grupos de estudiantes aprobados y desaprobados en el primer curso de matemática en la UNALM. Finalmente, el modelo ajustado es significativo.

Tabla 18

Resultados de la clasificación para Resultado. Ciclo 2017 II

			Resultado Pronosticado		
			0	1	Total
Resultado observado	N	0	46	36	82
		1	31	64	95
	%	0	56,1	43,9	100,0
		1	32,6	69,5	100,0

De la Tabla 18 se observa que el modelo logístico propuesto aplicado en la muestra de entrenamiento genera un 63,3% de clasificación correcta de estudiantes, de los cuales se clasificó correctamente al 69,5% de los estudiantes que aprobaron y al 56,1% de los estudiantes que desaprobaron el primer curso de matemática en la UNALM en el ciclo 2017 II. Esto significa que el modelo logístico es bueno para clasificar a los estudiantes que aprueban o no el primer curso de matemática en la UNALM.

4.1.3 Modelo de regresión logística binaria para Créditos Aprobados 2017 I

Tabla 19

Variables en el modelo logístico para Créditos Aprobados. Ciclo 2017 I

Variable	B	Error estándar	Wald	gl	Sig.	Exp(B)
Matemática	0,126	0,063	3,977	1	0,046	1,134
Raz. Verbal	-0,228	0,101	5,073	1	0,024	0,796
Física	0,147	0,062	5,672	1	0,017	1,159
Química	0,191	0,066	8,401		0,004	1,211
Constante	-2,487	1,879	1,752	1	0,186	0,083

De la Tabla 19 se observa que los coeficientes de las variables Matemática, Razonamiento Verbal, Física y Química son significativos para el modelo propuesto. Así también de los coeficientes mostrados en la tabla, se puede deducir que por cada incremento en la nota obtenida en Matemática en el ciclo 2017 I, el logaritmo de los odds para la aprobación de al menos 12 créditos matriculados en comparación con la aprobación de menos de 12 créditos, aumenta en 0,126. Asimismo, cada incremento en la nota obtenida en Razonamiento Verbal, el logaritmo de los odds para la aprobación de al menos 12 créditos matriculados en comparación con la aprobación de menos de 12 créditos disminuye en 0,2228. Además, cada incremento en la nota obtenida en Física, el logaritmo de los odds para la aprobación de al menos 12 créditos matriculados en comparación con la aprobación de menos de 12 créditos aumenta en 0,147. Finalmente, por cada incremento en la nota obtenida en Química, el logaritmo de los odds para la aprobación de al menos 12 créditos matriculados en comparación con la aprobación de menos de 12 créditos, aumenta en 0,191.

Así, el modelo logístico propuesto para la variable Créditos Aprobados en el primer ciclo 2017 I de la UNALM, está dado por:

$$Prob(Y_1 = 1) = \frac{1}{1 + e^{2{,}487 - 0{,}126 * \text{Matem} + 0{,}228 * \text{Raz.Verbal} - 0{,}147 * \text{Física} - 0{,}191 * \text{Química}}}$$

Se realizaron las pruebas Ómnibus para probar la hipótesis nula sobre la significancia de los coeficientes del modelo:

$$H_0: \ \beta_i = 0$$

$$H_1: \ \beta_i \neq 0$$

Tabla 20

Prueba Ómnibus del Modelo Logístico para Créditos Aprobados. Ciclo 2017 I

Descripción	Chi-cuadrado	gl	Sig.
Modelo	35,224	4	0,000

En la Tabla 20 se muestra que el valor Chi-cuadrado es 35,224 con p_valor igual a 0,000, esto permite concluir que existe una relación estadísticamente significativa entre la variable dependiente Créditos Aprobados en el primer ciclo y las variables independientes Matemática, Razonamiento Verbal, Física y Química que corresponden a las notas que obtuvieron los estudiantes en el examen de admisión 2017 I en la UNALM.

Se realizaron las pruebas de Hosmer y Lemeshow para probar la hipótesis sobre la bondad de ajuste del modelo:

H_0: No existen diferencias entre los Créditos Aprobados observados y pronosticados.

H_1: Existen diferencias entre los Créditos Aprobados observados y pronosticados.

Tabla 21

Prueba Hosmer y Lemeshow. *Modelo Logístico para Créditos Aprobados. Ciclo 2017 I*

Descripción	Chi-cuadrado	gl	Sig.
Modelo	2,501	8	0,962

De la Tabla 21 se observa que el valor Chi-cuadrado es 2,501 con un p_valor igual a 0,962, esto nos indica que se acepta la hipótesis nula, en conclusión, no existen diferencias entre los Créditos Aprobados observados y pronosticados para los dos grupos de estudiantes que aprobaron al menos 12 créditos y los que aprobaron menos de 12 créditos matriculados en el primer ciclo 2017 I de la UNALM. Finalmente, el modelo ajustado es significativo.

Tabla 22

Resultados de la clasificación para Créditos Aprobados. Ciclo 2017 I

			Créditos Aprobados Pronosticado		
			0	1	Total
Créditos Aprobados observado	N	0	9	30	39
		1	4	167	171
	%	0	23,1	76,9	100,0
		1	02,3	97,7	100,0

De la Tabla 22 se observa que el modelo logístico propuesto aplicado en la muestra de entrenamiento genera un 83,8% de clasificación correcta de estudiantes, de los cuales se clasificó correctamente al 97,7% de los estudiantes que aprobaron y al 23,1% de los estudiantes que desaprobaron el primer curso de matemática en la UNALM en el ciclo 2017 I. Esto significa que el modelo logístico es bueno para clasificar a los estudiantes que aprueban o no el primer curso de matemática en la UNALM.

4.1.4 Modelo de regresión logística binaria para Créditos Aprobados Resultado 2017 II

Tabla 23

Variables en el modelo logístico para Créditos Aprobados. Ciclo 2017 II

Variable	B	Error estándar	Wald	gl	Sig.	Exp(B)
Matemática	0,128	0,044	8,405	1	0,004	1,136
Modalidad	20,726	14843,273	0	1	0,999	1002326776
Constante	-0,982	0,498	3,883	1	0,000	0,375

De la Tabla 23 se observa que los coeficientes de las variables Matemática y Modalidad de ingreso son significativos para el modelo propuesto. Así también de los coeficientes mostrados en la tabla, se puede deducir que por cada incremento en la nota obtenida en Matemática en el ciclo 2017 II, el logaritmo de los odds para la aprobación de al menos 12 créditos en comparación con la aprobación de menos de 12 créditos aumenta en 0,128. Asimismo, por cada cambio de modalidad de ingreso de concurso ordinario a otro tipo de modalidad, el logaritmo de los odds para la aprobación de al menos 12 créditos en comparación con la aprobación de menos de 12 créditos aumenta en 20,726.

Así, el modelo logístico propuesto para la variable Créditos Aprobados en el primer ciclo 2017 II de la UNALM, está dado por:

$$Prob(Y_1 = 1) = \frac{1}{1 + e^{0,982 - 0,128 * \text{Matemática} - 20,726 * \text{Modalidad}}}$$

Se realizaron las pruebas Ómnibus para probar la hipótesis nula sobre la significancia de los coeficientes del modelo:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

Tabla 24

Prueba Ómnibus de coeficientes del Modelo Logístico para Créditos Aprobados. Ciclo 2017 II

Descripción	Chi-cuadrado	gl	Sig.
Modelo	15,964	2	0,000

En la Tabla 24 se muestra que el valor Chi-cuadrado es 15,964 con p_valor igual a 0,000, esto permite concluir que existe una relación estadísticamente significativa entre la variable dependiente Créditos Aprobados en el primer ciclo y las variables independientes Matemática y Modalidad de ingreso que corresponden a las notas que obtuvieron los estudiantes en el examen de admisión 2017 II en la UNALM.

Se realizaron las pruebas de Hosmer y Lemeshow para probar la hipótesis sobre la bondad de ajuste del modelo:

H_0: No existen diferencias entre los Créditos Aprobados observados y pronosticados.

H_1: Existen diferencias entre los Créditos Aprobados observados y pronosticados.

Tabla 25

Prueba Hosmer y Lemeshow. *Modelo Logístico para Créditos Aprobados. Ciclo 2017 II*

Descripción	Chi-cuadrado	gl	Sig.
Modelo	6,210	7	0,515

De la Tabla 25 se observa que el valor Chi-cuadrado es 6,210 con un p_valor igual a 0,515, esto nos indica que se acepta la hipótesis nula, en conclusión, no existen diferencias entre los Créditos Aprobados observados y pronosticados para los dos grupos de estudiantes que aprueban al menos 12 créditos y los que aprueban menos de 12 créditos matriculados en el primer ciclo 2017 II en la UNALM. Finalmente, el modelo ajustado es significativo.

Tabla 26

Resultados de clasificación para Créditos Aprobados. Ciclo 2017 II

			Resultado Pronosticado		
			0	1	Total
Resultado observado	N	0	22	46	68
		1	15	94	109
	%	0	32,4	67,6	100,0
		1	13,8	86,2	100,0

De la Tabla 26 se observa que el modelo logístico propuesto aplicado en la muestra de entrenamiento genera un 65,5% de clasificación correcta de estudiantes, de los cuales se clasificó correctamente al 86,2% de los estudiantes que aprobaron y al 32,4% de los estudiantes que desaprobaron el primer curso de matemática en la UNALM en el ciclo 2017 II. Esto significa que el modelo logístico es bueno para clasificar a los estudiantes que culminan el primer ciclo en la UNALM con 12 o más créditos aprobados.

4.2 Estimación de modelos del análisis discriminante

4.2.1 Cumplimiento de supuestos para Resultado 2017 I y 2017 II

En el presente estudio de investigación se buscó conocer a qué grupo pertenece un estudiante ingresante y matriculado en el primer ciclo académico de la universidad, según su desempeño académico, es decir, tomando en cuenta las variables categóricas Resultado y Créditos Aprobados. Se eligieron las siguientes variables independientes métricas: Edad, Matemática, Razonamiento Matemático, Razonamiento Verbal, Física, Química, Biología y Puntaje general.

Antes de estimar el modelo discriminante, Hair (1999) propone la comprobación del cumplimiento de los siguientes supuestos del análisis discriminante:

a) Normalidad multivariante de las variables independientes.
b) Matriz de covarianzas iguales.
c) No presencia de multicolinealidad.

Tabla 27

Prueba de Normalidad para Resultado. Ciclos 2017 I y 2017 II

Variables	2017 I		2017 II	
	K-S[a]	P_valor.	K-S[a]	P_valor.
Edad	2,451	0,000	1,779	0,004
Matemática	2,031	0,001	1,848	0,002
Razonamiento Matemático	1,580	0,014	1,618	0,011
Razonamiento Verbal	1,173	0,128	1,263	0,082
Física	2,811	0,000	1,940	0,001
Química	1,684	0,007	0,872	0,432
Biología	0,966	0,308	1,228	0,098
Puntaje general	3,072	0,000	1,825	0,003

[a] Prueba de Kolmogorov-Smirnov.

Con respecto al ciclo 2017 I, en la Tabla 27 se muestran los resultados de la prueba de Kolmogorov-Smirnov. Se verificó la presencia de normalidad de las variables Razonamiento Verbal y Biología. Se rechazó la normalidad para el resto de las variables a un nivel de significación del 5%. De forma similar, para el ciclo 2017 II, se observa la presencia de Normalidad en las variables Razonamiento Verbal, Biología y Química.

Al respecto, Torrado y Berlanga (2013) afirman que el análisis discriminante es una técnica robusta que no se ve gravemente afectada si no se cumplen los supuestos de normalidad e igualdad de matrices covarianzas, en este caso recomiendan aplicar el test de M. de Box. En ese mismo sentido, Tabachnick y Fidell (1983, citado en De Diego, 2015) señalan que la importancia del incumplimiento del supuesto de normalidad se ve reducida cuando la muestra es muy grande. Este es el caso del presente estudio por ello se continuó con la revisión de los demás supuestos para el análisis discriminante.

Para evaluar la homogeneidad de varianza de una variable se emplea usualmente la prueba de Levene. Si se contrasta más de una variable métrica, implicando la comparación de igualdad de las matrices de varianzas-covarianzas, se aplica la prueba M de Box (Hair, 1999).

Así, en la Tabla 28 se observa que para el ciclo 2017 I existe homogeneidad de varianzas entre los dos grupos de Resultados (aprobado-desaprobado) para todas las variables, a excepción de la nota obtenida en Física y el Puntaje general del examen de admisión. Adicionalmente, de la Prueba t, se observa que las medias de todas las variables difieren significativamente en ambos grupos de Resultados, a excepción de Biología para el cual no se observan diferencias en ambos grupos, es decir, las notas de Biología no discriminan bien entre los grupos de estudiantes aprobados y desaprobados en el primer curso de matemática.

Tabla 28

Comparación de medias y varianza para Resultado. Ciclo 2017 I

Variable independiente	Prueba de Levene		Prueba t	
	F	p_valor	t [a]	p_valor
Edad	0,206	0,651	2,938	0,004
Matemática	0,018	0,892	-4,612	0,000
Raz. Matemático	0,831	0,363	-3,551	0,000
Razonamiento Verbal	0,641	0,424	2,051	0,041
Física	4,173	0,042	-6,102	0,000
Química	1,259	0,263	-3,098	0,002
Biología	0,158	0,691	-0,435	0,664
Puntaje general	17,139	0,000	-6,685[b]	0,000

[a] gl=298

[b] gl=255,254. No se asumen varianzas iguales.

Tabla 29

Comparación de medias y varianza para Resultado. Ciclo 2017 II

Variable independiente	Prueba de Levene		Prueba t	
	F	p_valor	t [a]	p_valor
Edad	2,666	0,104	1,992	0,047
Matemática	0,014	0,905	-4,564	0,000
Raz. Matemático	2,326	0,128	-2,769	0,006
Razonamiento Verbal	1,193	0,276	2,720	0,007
Física	0,215	0,643	-4,100	0,000
Química	0,043	0,836	-1,132	0,259
Biología	2,686	0,102	1,597	0,112
Puntaje general	15,096	0,000	-4,259 [b]	0,000

[a] gl=251

[b] gl=236,689. No se asumen varianzas iguales.

En la Tabla 29, para el ciclo 2017 II, se observa una situación muy similar al ciclo 2017 I, esto es, existe homogeneidad de varianzas entre los dos grupos de Resultados (aprobado-desaprobado) para todas las variables, a excepción de Puntaje general del examen de admisión, en consecuencia se excluye del análisis para ambos ciclos académicos.

Adicionalmente, de la Prueba t, se evidencia que las medias de todas las variables difieren significativamente en ambos grupos de Resultados, a excepción de Biología y Química, para los cuales no se observan diferencias en la media, es decir, las notas de Biología y Química del examen de admisión 2017 II, no discriminan bien entre los grupos de estudiantes aprobados y desaprobados en el primer curso de matemática.

En la Tabla 30 se muestran los resultados de la prueba M de Box y se comprobó, a un nivel de significación del 5%, que existe homogeneidad de matriz de covarianzas en cada uno de los ciclos académicos.

Tabla 30

Prueba M de Box para Resultado. Ciclos 2017 I y 2017 II

Descripción	Valores	
	2017 I	2017 II
M de Box	6,558	10,282
F Aprox.	1,079	1,691
gl 1	6	6
gl 2	224310,583	418286,619
Sig.	0,372	0,119

Finalmente, en la Tabla 31 se verificó que no existe multicolinealidad entre las variables. Si los valores de Tolerancia son próximos a cero, indica que la variable se encuentra muy correlacionada con el resto de variables por lo que se deberá evaluar su eliminación ya que resulta redundante para los fines del análisis discriminante. El valor de VIF debe ser cercano a la unidad para que no exista multicolinealidad entre las variables.

Tabla 31

Prueba de Multicolinealidad para Resultado. Ciclos 2017 I y 2017 II

Variables	*Ciclo 2017 I*		*Ciclo 2017 II*	
	Tolerancia	VIF.	Tolerancia	VIF.
Edad	0,940	1,064	0,939	1,064
Matemática	0,795	1,259	0,773	1,294
Razonamiento Matemático	0,854	1,171	0,759	1,317
Razonamiento Verbal	0,934	1,071	0,888	1,126
Física	0,872	1,146	0,877	1,140
Química	0,831	1,203	0,793	1,262
Biología	0,792	1,262	0,734	1,361

Se tomaron en cuenta los resultados obtenidos en la revisión de supuestos y se continuó con el análisis.

Primero se observaron las medias en ambos grupos o categorías de la variable Resultado, pues se trata de encontrar diferencias significativas entre ellas.

Tabla 32

Media de variables métricas según Resultado y ciclo

Variables	2017 I		2017 II	
	Resultado		Resultado	
	0	1	0	1
Edad	21,9	21,2	21,2	20,7
Matemática	13,1	14,9	09,8	11,9
Razonamiento Matemático	13,3	14,3	11,7	12,5
Razonamiento Verbal	10,4	09,8	12,2	11,5
Física	13,3	15,6	12,1	13,6
Química	12,4	13,5	13,8	14,2
Biología	11,3	11,5	09,5	08,8

En la Tabla 32 se evidencia que para el ciclo 2017 I las variables que presentan mayores diferencias absolutas son Física (2,3) y Matemática (1,8). En tanto que Biología (0,2) y Razonamiento Verbal (0,6) presentan pequeñas diferencias absolutas. Así también, para el ciclo 2017 II, las variables que presentan mayores diferencias absolutas son Matemática (2,1) y Física (1,5), en cambio, Química (0,4) presenta la menor diferencia absoluta. Los resultados obtenidos son congruentes con los mostrados en las tablas 12 y 13.

El desarrollo del análisis discriminante para la variable Resultado en los dos ciclos académicos 2017 I y 2017 II tiene dos etapas. En la primera etapa, se formuló la función discriminante en base a una muestra aleatoria de entrenamiento correspondiente al 70% de la muestra total. En la segunda etapa, se validó el modelo propuesto en la primera etapa; para ello se empleó una muestra de validación correspondiente al 30% de la muestra total para realizar el análisis comparativo de los tres modelos de clasificación propuestos en el presente trabajo de investigación.

En la Tabla 33 se detalla el tamaño de muestra para ambos ciclos académicos.

Tabla 33

Tamaño de muestra para el análisis discriminante

Muestra	2017 I	2017 II
Entrenamiento	210	177
Validación	90	76
Total	300	253

4.2.2 Modelo Discriminante para Resultado 2017 I

Se inició con la prueba de igualdad de medias para las dos categorías de la variable Resultado en el ciclo 2017 I

Tabla 34

Prueba de igualdad de medias para Resultado. Ciclo 2017 I

Variables	Lambda de Wilks	F	gl1	gl2	P_valor.
Edad	0,976	5,066	1	208	0,025
Matemática	0,919	18,222	1	208	0,000
Raz. Matemático	0,957	9,394	1	208	0,002
Raz. Verbal	0,978	4,762	1	208	0,030
Física	0,830	42,564	1	208	0,000
Química	0,919	18,455	1	208	0,000
Biología	0,998	0,503	1	208	0,479

En este estudio de investigación, la selección de variables se realizó siguiendo el procedimiento por pasos que inicia con la selección de la primera variable, que se espera sea la mejor discriminadora entre todas las variables, para ello se prueba una hipótesis de igualdad de medias iguales; así la variable con el mayor valor de F en el análisis de varianza será la primera variable seleccionada. Antes de elegir una nueva variable, este procedimiento comprueba si todas las variables que fueron seleccionadas previamente, siguen siendo significativas (Johnson, 2004).

De acuerdo al valor más alto de F de la Tabla 34, la primera variable que ingresó al modelo discriminante es Física. Además, se observó que todas las variables estudiadas son buenas para discriminar entre los dos grupos de Resultados, es decir, discrimina bien entre los grupos de estudiantes aprobados y desaprobados en el primer curso de matemática a excepción de la variable Biología (p_valor=0,479). Sin

embargo, se debe considerar que una variable no significativa a nivel univariante podría aportar información discriminativa a nivel multivariante (De la Fuente, 2011). Por ello, se decidió mantener la variable Biología dentro del análisis discriminante.

Cabe recordar que uno de los supuestos del análisis discriminante es que las matrices de covarianzas correspondientes a cada uno de los grupos son iguales entre sí, es decir, que todos los grupos provienen de una misma población. La prueba M de Box carece de distribución muestral conocida y se puede transformar en un estadístico F (De la Fuente, 2011). Mediante la prueba M de Box se verifica la igualdad de matrices de covarianza. Los resultados se muestran en la Tabla 35.

Tabla 35

Prueba M de Box para Resultado. Ciclo 2017 I

Descripción	Resultado 2017 I
M de Box	5,989
F Aprox.	0,979
gl 1	6
gl 2	97528,087
Sig.	0,438

Para estimar los coeficientes de la función discriminante fue necesario efectuar 3 pasos en los que se logró determinar cuáles fueron las variables introducidas o eliminadas en cada paso. Se usó el estadístico lambda de Wilks global para la selección de variables. Los criterios fueron: se incorpora una variable si el valor de F es mayor de 3,84 y se elimina la variable si el valor de F es menor de 2,71.

En la Tabla 36 se muestran las variables incluidas en el modelo al finalizar el paso 3. Como se señaló, la Tolerancia de una variable mide el grado de asociación lineal con el resto de las variables y cuando su valor se aproxima a la unidad significa que la variable no se correlaciona con el resto de las variables por lo que se debe mantener

en el modelo. Siendo así, la primera variable incluida en el paso 1 fue Física, Química en el paso 2 y Matemática en el tercer y último paso.

Tabla 36

Variables incluidas en el análisis para Resultado. Ciclo 2017 I

Variables	Tolerancia	F para eliminar	Lambda de Wilks
Física	0,947	23,574	0,844
Química.	0,986	12,112	0,801
Matemática	0,953	8,298	0,787

En la Tabla 37 se muestra el resumen de las funciones discriminantes canónicas. Dado que hay dos grupos de Resultado, se forma una sola función discriminante y por tanto se establece un solo autovalor que corresponde al autovector en la dirección de la máxima dispersión de las medias de los grupos. Así también, la correlación canónica expresa la asociación entre las puntuaciones discriminantes y los grupos de la variable categórica. Cuanto más alto es su valor, entonces la función discrimina mejor entre los grupos ya que existe mayor dispersión entre ellos. Se evidencia que la función discrimina bien los dos grupos establecidos, además, solo hay una función discriminante que explica el 100% de la varianza total.

Tabla 37

Autovalores para Resultado. Ciclo 2017 I

Función	Autovalor	% de varianza	Correlación Canónica
1	0,321	100	0,493

En la Tabla 38 se muestran los resultados de la prueba de igualdad de medias entre los dos grupos de las funciones discriminantes obtenidos. A un nivel de significación del 5%, se rechaza la hipótesis nula de la igualdad de medias, es decir, la función discriminante es significativa y tiene un alto poder para discriminar entre las dos categorías de Resultado, esto es, permite clasificar bien a los estudiantes que

aprueban y desaprueban el primer curso de matemática en la UNALM en el ciclo 2017 I.

Tabla 38

Prueba Lambda de Wilks para Resultado. Ciclo 2017 I

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	0,757	57,500	3	0,000

En la Tabla 39 se evidencia la capacidad discriminante que tienen las variables estudiadas. Los coeficientes de la función discriminante canónica estandarizada muestran la importancia de las variables que tienen mayor influencia en la función discriminante, al ser estandarizada se evitan posibles problemas de escala entre las variables. La matriz de estructuras está ordenada según el poder discriminante de la variable. Se concluye que la nota obtenida en el área de Física tiene mayor capacidad discriminante, seguido de las notas obtenidas en el área de Química y Matemática del examen de admisión 2017 I.

Tabla 39

Coeficientes y Matriz de estructuras para Resultado. Ciclo 2017 I

Variables	Función 1	
	Coef. de func. discrim. canónica estandarizada	Matriz de Estructuras
Física	0,668	0,798
Química	0,481	0,526
Matemática	0,409	0,522
Raz. Matemático[a]	-	0,193
Raz. Verbal[a]	-	-0,143
Biología[a]	-	0,067
Edad[a]	-	-0,031

[a] la variable no se utiliza en el análisis

En la Tabla 40 se muestran los coeficientes no estandarizados de la función discriminante canónica que son proporcionales a los coeficientes de la función discriminante lineal de Fisher. Los coeficientes no estandarizados son importantes porque allí se evalúan las medias de los grupos a fin de obtener los correspondientes centroides. A su vez, con estos se obtiene el punto de corte discriminante. En este caso, los centroides son -0,852 y 0,373 para el grupo de desaprobados y aprobados en el primer curso de matemática en el ciclo 2017 I, respectivamente.

Tabla 40

Coeficientes no estandarizados discriminante de Resultado. Ciclo 2017 I

Variables	Función 1
Matemática	0,131
Física	0,233
Química	0,161
Constante	-7,542

En la Tabla 41 se observan las variables con mayor poder discriminante y sus respectivos coeficientes, con los cuales se formulan las funciones de clasificación para los dos grupos de Resultado: aprobados y desaprobados.

Tabla 41

Coeficientes de función de clasificación para Resultado. Ciclo 2017 I

Variables	Resultado		Diferencia
	0	1	
Matemática	1,170	1,330	0,160
Física	1,191	1,476	0,285
Química	1,306	1,504	0,198
Constante	-24,476	-32,598	-8,122

Las funciones de clasificación para los dos grupos de Resultado en el ciclo 2017 I son Z_0 y Z_1 para los estudiantes desaprobados y aprobados en el primer curso de matemática, respectivamente.

$$Z_0 = -24{,}476 + 1{,}170 * \text{Matemática} + 1{,}191 * \text{Física} + 1{,}306 * \text{Química}$$

$$Z_1 = -32{,}598 + 1{,}330 * \text{Matemática} + 1{,}476 * \text{Física} + 1{,}504 * \text{Química}$$

La función discriminante lineal de Fisher para la variable Resultado en el ciclo 2017 I, está dada por:

$$Z_i = -8{,}122 + 0{,}160 * \text{Matemática}_i + 0{,}285 * \text{Física}_i + 0{,}198 * \text{Química}_i$$

Criterio de Decisión:
Si $Z_i < 0$, se clasifica al estudiante en el grupo 0: desaprobado.
Si $Z_i > 0$, se clasifica al estudiante en el grupo 1: aprobado.

Tabla 42

Resultados de clasificación discriminante para Resultado. Ciclo 2017 I

			Resultado Pronosticado		
			0	1	Total
Resultado observado	N	0	26	38	64
		1	16	130	146
	%	0	40,6	59,4	100,0
		1	11,0	89,0	100,0

De la Tabla 42 se observa que la función discriminante lineal de Fisher permitió la clasificación correcta del 74,3% de estudiantes, de los cuales se clasificó correctamente al 89% de los estudiantes que aprobaron y al 40,6% de los estudiantes que desaprobaron el primer curso de matemática en la UNALM en el ciclo 2017 I.

4.2.3 Modelo Discriminante para Resultado Ciclo 2017 II

Se probó la prueba de igualdad de medias para las dos categorías de la variable Resultado en el ciclo 2017 II.

Tabla 43

Prueba de igualdad de medias para Resultado. Ciclo 2017 II

Variables	Lambda de Wilks	F	gl1	gl2	P_valor.
Edad	0,995	0,801	1	175	0,372
Matemática	0,902	19,066	1	175	0,000
Raz. Matemático	0,951	9,110	1	175	0,003
Raz. Verbal	0,954	8,376	1	175	0,004
Física	0,970	5,421	1	175	0,021
Química	0,997	0,583	1	175	0,446
Biología	0,993	1,257	1	175	0,264

La selección de variables se realizó con el procedimiento de inclusión por pasos. De la Tabla 43, se observa que el mayor valor de F corresponde a la variable Matemática, lo cual significa que será la primera variable en ingresar al modelo discriminante. Además, las variables que discriminan mejor a los dos grupos de Resultados en el primer curso de matemática son Puntaje general, Razonamiento Matemático, Razonamiento Verbal y Física.

La Tabla 44 muestra los resultados de la prueba de igualdad de matrices de covarianza mediante la prueba M de Box, por lo cual existe suficiente evidencia estadística para no rechazar la hipótesis.

Tabla 44

Prueba M de Box para Resultado. Ciclo 2017 II

Descripción	Resultado 2017 II
M de Box	1,350
F Aprox.	0,445
gl 1	3
gl 2	18256093,99
Sig.	0,721

Los coeficientes de la función discriminante se estimaron en 2 pasos. Se usó el estadístico lambda de Wilks global para la selección de variables. Los criterios fueron: se incorpora una variable si el valor de F es mayor de 3,84 y se elimina la variable si el valor de F es menor de 2,71.

En la Tabla 45 se muestran las variables incluidas en el modelo al finalizar el paso 3. Como ya señaló, la Tolerancia mide el grado de asociación lineal de una variable con respecto al resto; y será mejor cuando se aproxima a la unidad ya que no existe correlación con el resto de las variables y por lo tanto, la variable se mantiene en el modelo. Siendo así, la primera variable incluida en el paso 1 fue Matemática y Razonamiento Verbal en el paso 2.

Tabla 45

Variables incluidas en el análisis para Resultado. Ciclo 2017 II

Variables	Tolerancia	F para eliminar	Lambda de Wilks
Matemática	0,971	14,671	0,954
Raz. Verbal.	0,971	4,279	0,902

En la Tabla 46 se muestra el resumen de las funciones discriminantes canónicas. Hay dos grupos de Resultado y es por ello que se forma una sola función discriminante y por tanto se establece un solo autovalor. Así también, cuanto más alto sea el valor de la correlación canónica, entonces la función discriminará mejor entre los grupos ya que habrá

mayor dispersión entre ellos. De los resultados se evidencia que la función discrimina bien los dos grupos establecidos, además, solo hay una función discriminante que explica el 100% de la varianza total.

Tabla 46

Autovalores para Resultado. Ciclo 2017 II

Función	Autovalor	% de varianza	Correlación Canónica
1	0,136	100	0,346

En la Tabla 47 se muestran los resultados de la prueba de igualdad de medias entre los dos grupos de Resultado. A un nivel de significación del 5%, se rechaza la hipótesis nula de igualdad de medias, con lo cual la función discriminante es significativa y tiene un alto poder para discriminar entre las dos categorías de Resultado, es decir, la función permite clasificar bien a los estudiantes que aprueban y desaprueban el primer curso de matemática en la UNALM en el ciclo 2017 II.

Tabla 47

Prueba Lambda de Wilks para Resultado. Ciclo 2017 II

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	0,880	22,221	2	0,000

En la Tabla 48 se observa la capacidad discriminante que tienen las variables estudiadas. Los coeficientes de la función discriminante canónica estandarizada muestran la importancia de las variables Matemática y Razonamiento Verbal ya que sus puntajes son los más altos, lo cual indica que tienen mayor influencia en la función discriminante. La matriz de estructuras está ordenada según el poder o capacidad discriminante que tiene la variable.

Tabla 48

Coeficientes y Matriz de estructuras para Resultado. Ciclo 2017 II

Variables	Función 1	
	Coef. de func. discrim. canónica estandarizada	Matriz de Estructuras
Matemática	0,817	0,894
Raz. Verbal	-0,454	-0,593
Física[a]	-	0,254
Raz. Matemático[a]	-	-0,253
Química[a]	-	0,248
Edad[a]	-	-0,131
Biología[a]	-	0,054

[a] la variable no se utiliza en el análisis

En la Tabla 49 se muestran los coeficientes no estandarizados de la función discriminante canónica que son proporcionales a los coeficientes de la función discriminante lineal de Fisher. Estos coeficientes son importantes porque al evaluarlo para las medias de los grupos se obtienen los correspondientes centroides que permiten hallar el punto de corte discriminante. En este caso, los centroides son -0,395 y 0,341 para el grupo de desaprobados y aprobados en el primer curso de matemática en el ciclo 2017 II, respectivamente.

Tabla 49

Coeficientes no estandarizados discriminante de Resultado. Ciclo 2017 II

Variables	Función 1
Matemática	0,225
Raz. Verbal	-0,212
Constante	0,031

En la Tabla 50 se observan las variables con mayor poder discriminante y con sus coeficientes se obtuvieron las funciones de clasificación para los dos grupos de Resultado: aprobados y desaprobados.

Tabla 50

Coeficientes de función de clasificación para Resultado. Ciclo 2017 II

Variables	Resultado		Diferencia
	0	1	
Matemática	1,040	1,205	0,165
Raz. Verbal	2,992	2,837	-0,155
Constante	-24,407	-24,218	0,189

Las funciones de clasificación para los dos grupos de Resultado en el ciclo 2017 II son Z_0 y Z_1 para los estudiantes desaprobados y aprobados en el primer curso de matemática, respectivamente.

$$Z_0 = -24{,}407 + 1{,}040 * \text{Matemática} + 2{,}992 * \text{Raz. Verbal}$$

$$Z_1 = -24{,}218 + 1{,}205 * \text{Matemática} + 2{,}837 * \text{Raz. Verbal}$$

La función discriminante lineal de Fisher para la variable Resultado en el ciclo 2017 II que está dada por:

$$Z_i = 0{,}189 + 0{,}165 * \text{Matemática}_i - 0{,}155 * \text{Raz. Verbal}_i$$

Criterio de Decisión:

Si $Z_i < 0$, se clasifica al estudiante en el grupo 0: desaprobado.

Si $Z_i > 0$, se clasifica al estudiante en el grupo 1: aprobado.

Con la función discriminante hallada se procedió a la evaluación para cada estudiante de la muestra y se encontró el número de casos asignados correctamente a ambos grupos de Resultado, es decir aprobado o desaprobado en el primer curso de matemática de la UNALM para el ciclo 2017 II.

Tabla 51

Resultados de la clasificación discriminante para Resultado. Ciclo 2017 II

			Resultado Pronosticado		
			0	1	Total
Resultado observado	N	0	42	40	82
		1	27	68	95
	%	0	51,2	48,8	100,0
		1	28,4	71,6	100,0

En la Tabla 51 se observa que la función discriminante lineal de Fisher permitió la clasificación correcta del 62,1% de estudiantes de primer ciclo, de los cuales se clasificó correctamente al 71,6% de estudiantes aprobados y al 51,2% de los estudiantes que desaprobaron el primer curso de matemática en la UNALM en el ciclo 2017 II.

4.2.4 Cumplimiento de supuestos para Créditos Aprobados 2017 i y 2017 II

Antes de hallar el modelo discriminante, se evaluó el cumplimiento de los supuestos del análisis discriminante: Normalidad multivariante de las variables independientes, Matriz de covarianzas iguales y ausencia de multicolinealidad.

Tabla 52

Prueba de Normalidad. Créditos Aprobados. Ciclos 2017 I y 2017 II

Variables	2017 I		2017 II	
	K-S[a]	P_valor.	K-S[a]	P_valor.
Edad	2,295	0,000	1,790	0,003
Matemática	1,594	0,012	1,251	0,087
Razonamiento Matemático	1,257	0,085	1,149	0,142
Razonamiento Verbal	1,173	0,128	0,629	0,824
Física	1,939	0,001	1,376	0,045
Química	1,650	0,009	1,454	0,029
Biología	0,896	0,399	0,680	0,745
Puntaje general	1,864	0,002	1,930	0,001

[a] Prueba de Kolmogorov-Smirnov.

Con respecto al ciclo 2017 I, en la Tabla 52 se observa que al aplicar la prueba de Kolmogorov-Smirnov las variables Biología, Razonamiento Verbal y Razonamiento Matemático presentan normalidad. En cambio, se rechazó la normalidad para el resto de las variables a un nivel de significación del 5%. Así también, para el ciclo 2017 II, se observa la presencia de normalidad en las variables Razonamiento Verbal, Biología, Razonamiento Matemático y Matemática; este es un resultado muy similar al observado en el ciclo 2017 I.

Tal como se procedió con la variable Resultado, se continuó con la revisión de los demás supuestos para el análisis discriminante.

Para evaluar la homogeneidad de varianza de una variable se empleó la prueba de Levene para el caso univariado y la prueba M de Box para el caso multivariado. En la Tabla 53 se observa la presencia de homogeneidad de varianzas entre los dos grupos de Créditos Aprobados para el ciclo 2017 I a excepción del Puntaje general del examen de admisión. Adicionalmente, de la Prueba t, se observa que las medias de todas las variables difieren significativamente en ambos grupos de Créditos Aprobados, a excepción de Biología y Razonamiento Matemático para los cuales no se observan diferencias en la media para ambos grupos de Créditos Aprobados, es decir, las notas de Biología y Razonamiento Matemático no discriminarían bien entre los grupos de estudiantes que aprobaron al menos 12 créditos o menos de 12 créditos matriculados en el primer ciclo 2017 I de la UNALM.

Tabla 53

Comparación de medias y varianza para Créditos Aprobados. Ciclo 2017 I

Variable independiente	Prueba de Levene		Prueba t	
	F	p_valor	t [a]	p_valor
Edad	1,274	0,260	3,268	0,001
Matemática	1,164	0,282	-2,790	0,006
Raz. Matemático	2,681	0,103	-1,558	0,120
Razonamiento Verbal	0,074	0,785	2,116	0,035
Física	0,026	0,872	-3,379	0,001
Química	0,778	0,378	-3,474	0,001
Biología	0,980	0,323	-1,275	0,203
Puntaje general	10,543	0,001	-4,706 [b]	0,000

[a] gl=298

[b] gl=120,383. No se asumen varianzas iguales.

Tabla 54

Comparación de medias y varianza para Créditos Aprobados. Ciclo 2017 II

Variable independiente	Prueba de Levene		Prueba t	
	F	p_valor	t [a]	p_valor
Edad	0,634	0,427	2,351	0,020
Matemática	0,147	0,702	-3,020	0,003
Raz. Matemático	0,394	0,531	-2,153	0,032
Razonamiento Verbal	0,943	0,333	0,929	0,354
Física	0,887	0,347	-2,880	0,004
Química	2,894	0,090	-2,150	0,033
Biología	0,912	0,341	0,711	0,478
Puntaje general	17,834	0,000	-4,050	0,000

[a] gl=251

En la Tabla 54, para el ciclo 2017 II, se observa una situación muy similar, esto es, existe homogeneidad de varianzas entre los dos grupos de Créditos Aprobados para todas las variables, a excepción de Puntaje general del examen de admisión. Adicionalmente, de la Prueba t, se evidencia que las medias de todas las variables difieren significativamente en ambos grupos de Créditos Aprobados, a excepción de Biología y Razonamiento Verbal, para los cuales no se observan diferencias en la media, es decir, las notas de Biología y Razonamiento Verbal no discriminarían bien entre los grupos de Créditos Aprobados, es decir, entre los estudiantes que aprueban al menos 12 créditos y los estudiantes que aprueban menos de 12 créditos matriculados en el primer ciclo académico 2017 II en la UNALM.

En la Tabla 55 se muestran los resultados de la prueba M de Box. Para el ciclo 2017 I si existe homogeneidad, en tanto, que para el ciclo académico 2017 II se rechazó la hipótesis de homogeneidad de matriz de covarianzas, a un nivel de significación del 5%.

Tabla 55

Prueba M de Box para Créditos Aprobados. Ciclos 2017 I y 2017 II

Descripción	Valores	
	2017 I	2017 II
M de Box	3,967	20,204
F Aprox.	0,386	1,983
gl 1	10	10
gl 2	47924,699	194399,695
Sig.	0,953	0,031

Finalmente, en la Tabla 56 se verificó que no existe multicolinealidad entre las variables. Si los valores de Tolerancia son próximos a cero, indica que la variable se encuentra muy correlacionada con el resto de las variables por lo que se deberá evaluar su eliminación ya que resulta redundante para los fines del análisis discriminante. El valor de VIF debe ser cercano a la unidad para que no exista multicolinealidad entre las variables.

Tabla 56

Prueba de Multicolinealidad para Créditos Aprobados. Ciclos 2017 I y 2017 II

Variables	*Ciclo 2017 I*		*Ciclo 2017 II*	
	Tolerancia	VIF.	Tolerancia	VIF.
Edad	0,940	1,064	0,939	1,064
Matemática	0,795	1,259	0,773	1,294
Razonamiento Matemático	0,854	1,171	0,759	1,317
Razonamiento Verbal	0,934	1,071	0,888	1,126
Física	0,872	1,146	0,877	1,140
Química	0,831	1,203	0,793	1,262
Biología	0,792	1,262	0,734	1,361

Se tomaron en cuenta los resultados obtenidos en la revisión de supuestos y se continuó con el análisis.

Primero se calcularon las medias de los grupos de la variable Créditos Aprobados, para encontrar diferencias significativas entre ellas.

Tabla 57

Media de variables métricas según Créditos Aprobados y ciclo

Variables	2017 I		2017 II	
	Créditos Aprobados		Créditos Aprobados	
	0	1	0	1
Edad	22,224	21,211	21,309	20,853
Matemática	13,294	14,595	09,967	11,742
Raz. Matemático	13,533	14,067	11,470	12,626
Raz. Verbal	10,545	09,836	12,225	11,697
Física	13,647	15,188	12,451	13,204
Química	11,948	13,463	13,515	14,123
Biología	10,898	11,571	09,274	09,009

En la Tabla 57 se evidencia que para el ciclo 2017 I las variables que presentan mayores diferencias absolutas son Física (1,541), Química (1,515) y Matemática (1,301). En tanto que Razonamiento Matemático (0,534) y Biología (0,673) presentan pequeñas diferencias absolutas. Así también, para el ciclo 2017 II las variables que presentan mayores diferencias absolutas son Matemática (1,775) y Razonamiento Matemático (1,156), en cambio, las notas de Biología (0,265) y Edad (-0,456) presentan las menores diferencias absolutas. Los resultados obtenidos son congruentes con los mostrados en las tablas 37 y 38.

Se procedió con el análisis discriminante para la variable Créditos Aprobados en los ciclos académicos 2017 I y 2017 II con una muestra aleatoria de entrenamiento de 210 estudiantes correspondiente al 70% de la muestra total.

4.2.5 Modelo Discriminante para Créditos Aprobados 2017 I

Se inició con la prueba de igualdad de medias para las dos categorías de la variable Créditos Aprobados en el ciclo 2017 I.

Tabla 58

Prueba de igualdad de medias para Créditos Aprobados. Ciclo 2017 I

Variables	Lambda de Wilks	F	gl1	gl2	P_valor.
Edad	0,974	5,472	1	208	0,020
Matemática	0,951	10,648	1	208	0,001
Raz. Matemático	0,979	4,445	1	208	0,036
Raz. Verbal	0,951	10,793	1	208	0,001
Física	0,927	16,312	1	208	0,000
Química	0,940	13,178	1	208	0,000
Biología	0,987	2,676	1	208	0,103

La selección de variables se realizó con el procedimiento de inclusión por pasos. De acuerdo con el valor de F obtenido (el más alto valor) en la Tabla 58, la primera variable que ingresó al modelo discriminante lineal es Física. Además, por la significancia de la prueba, se observa que todas las variables discriminan bien a los dos grupos de Créditos Aprobados en el primer curso de matemática, a excepción de Biología.

En la Tabla 59 se muestran los resultados de la prueba de igualdad de matrices de covarianza mediante la prueba M de Box, por lo cual existe suficiente evidencia estadística para no rechazar la hipótesis a un nivel de significación del 5%..

Tabla 59

Prueba M de Box para Créditos Aprobados. Ciclo 2017 I

Descripción	Créditos Aprobados 2017 II
M de Box	4,601
F Aprox.	0,442
gl 1	10
gl 2	21026,548
Sig.	0,927

Los coeficientes de la función discriminante se estimaron en 4 pasos. Se usó el estadístico lambda de Wilks global para la selección de variables. Los criterios fueron: se incorpora una variable si el valor de F es mayor de 3,84 y se elimina la variable si el valor de F es menor de 2,71.

En la Tabla 60 se muestran las variables incluidas en el modelo al finalizar el paso 4. La Tolerancia de una variable mide el grado de asociación lineal con el resto de las variables y cuando su valor se aproxima a la unidad significa que la variable no se correlaciona con el resto de variables por lo que se debe mantener en el modelo. Siendo así, la primera variable incluida en el paso 1 fue Física, Química en el paso 2, Razonamiento Verbal en el paso 3 y Matemática en el cuarto paso.

Tabla 60

Variables incluidas en el análisis para Créditos Aprobados. Ciclo 2017 I

Variables	Tolerancia	F para eliminar	Lambda de Wilks
Física	0,909	5,745	0,866
Química	0,975	9,102	0,880
Raz. Verbal	0,972	5,880	0,867
Matemática	0,917	3,858	0,858

En la Tabla 61 se muestra el resumen de las funciones discriminantes canónicas. Hay dos grupos de Créditos Aprobados y es por ello que se

forma una sola función discriminante y por tanto se establece un solo autovalor. Así también, cuanto más alto sea el valor de la correlación canónica, entonces la función discriminará mejor entre los grupos ya que existe mayor dispersión entre ellos. De los resultados se evidencia que la función discrimina bien los dos grupos de Créditos Aprobados, además, solo hay una función discriminante que explica el 100% de la varianza total.

Tabla 61

Autovalores para Créditos Aprobados. Ciclo 2017 I

Función	Autovalor	% de varianza	Correlación Canónica
1	0,187	100	0,397

En la Tabla 62 se muestran los resultados de la prueba de igualdad de medias entre los dos grupos de las funciones discriminantes. A un nivel de significación del 5% se rechaza la hipótesis nula de la igualdad de medias, es decir, la función discriminante es significativa y tiene un alto poder para discriminar entre las dos categorías de Créditos Aprobados, es decir, clasifica bien a los estudiantes que aprueban al menos 12 créditos y a los que aprueban menos de 12 créditos matriculados en el primer ciclo 2017 I en la UNALM.

Tabla 62

Prueba Lambda de Wilks para Créditos Aprobados. Ciclo 2017 I

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	0,842	35,333	4	0,000

De la Tabla 63 se observa la capacidad discriminante que tienen las variables estudiadas. Los coeficientes de la función discriminante canónica estandarizada muestran la importancia de las variables Física, Química, Razonamiento Verbal y Matemática ya que tienen mayor influencia en la función discriminante. La matriz de estructuras está ordenada según el poder discriminante de la variable.

Tabla 63

Coeficientes y Matriz de estructuras para Créditos Aprobados. Ciclo 2017 I

Variables	Función 1	
	Coef. de func. discrim. canónica estandarizada	Matriz de Estructuras
Física	0,436	0,647
Química	0,526	0,582
Raz. Verbal	-0,427	-0,527
Matemática	0,357	0,523
Raz. Matemático[a]	-	0,149
Biología[a]	-	0,063
Edad[a]	-	-0,062

[a] la variable no se utiliza en el análisis

En la Tabla 64 se muestran los coeficientes que son proporcionales a los coeficientes de la función discriminante lineal de Fisher. Son importantes porque al evaluarlo para las medias de los grupos se obtienen los correspondientes centroides que son necesarios para hallar el punto de corte discriminante. En este caso, los centroides son -0,901 y 0,206 para el grupo de Créditos Aprobados en el primer ciclo 2017 I de la UNALM.

Tabla 64

Coeficientes no estandarizados discriminante para Créditos Aprobados. Ciclo 2017 I

Variables	Función 1
Matemática	0,113
Raz. Verbal	-0,195
Puntaje General	0,144
Constante	-4,195

De la Tabla 65 se observan las variables con mayor poder discriminante y con los coeficientes se obtuvieron las funciones de clasificación para los dos grupos de Créditos Aprobados.

Tabla 65

Coeficientes de función de clasificación para Créditos Aprobados. Ciclo 2017 I

Variables	Créditos Aprobados		Diferencia
	0	1	
Matemática	1,316	1,441	0,125
Raz. Verbal	2,723	2,508	-0,215
Física	1,128	1,287	0,159
Química	1,156	1,349	0,193
Constante	-39,335	-42,116	-2,781

Las funciones de clasificación para Créditos Aprobados de ciclo 2017 I son:

$$Z_0 = -39{,}335 + 1{,}316 * \text{Matemática} + 2{,}723 * \text{Razonamiento Verbal} + 1{,}28 * \text{Física} + 1{,}156 * \text{Química}$$

$$Z_1 = -42{,}116 + 1{,}441 * \text{Matemática} + 2{,}508 * \text{Razonamiento Verbal} + 1{,}287 * \text{Física} + 1{,}349 * \text{Química}$$

La función discriminante lineal de Fisher para la variable Créditos Aprobados del ciclo 2017 I que está dada por:

$$Z_i = -2{,}781 + 0{,}125 * \text{Matemática} - 0{,}215 * \text{Razonamiento Verbal} + 0{,}159 * \text{Física} + 0{,}193 * \text{Química}$$

Criterio de Decisión:

Si $Z_i < 0$, se clasifica al estudiante en el grupo Aprobó menos de 12 créditos matriculados.

Si $Z_i > 0$, se clasifica al estudiante en el grupo Aprobó al menos 12 créditos matriculados.

Con la función discriminante hallada se procedió a la evaluación de cada uno de los casos y se encontró el número de casos asignados correctamente a ambos grupos de Créditos Aprobados, es decir, aprobó al menos 12 o menos de 12 créditos matriculados en el primer ciclo 2017 I de la UNALM.

Tabla 66

Resultados de clasificación discriminante para Créditos Aprobados. Ciclo 2017 I

			Créditos Aprobados Pronosticado		
			0	1	Total
Créditos Aprobados observado	N	0	9	30	39
		1	4	167	171
	%	0	23,1	76,9	100,0
		1	2,3	97,7	100,0

En la Tabla 66 se observa que la función discriminante lineal de Fisher permitió la clasificación correcta del 83,8% de estudiantes de primer ciclo, de los cuales se clasificó correctamente en un 23,1% y 97,7% para los estudiantes que aprobaron menos de 12 créditos y los que aprobaron al menos 12 créditos matriculados en el primer académico 2017 I en la UNALM.

4.2.6 Modelo Discriminante para Créditos Aprobados 2017 II

Se inició con la prueba de igualdad de medias para las dos categorías de la variable Créditos Aprobados en el ciclo 2017 II, para ello se retiró del análisis a la variable Puntaje general por presentar falta de homogeneidad de varianza univariada.

Tabla 67

Prueba de igualdad de medias para Créditos Aprobados. Ciclo 2017 II

Variables	Lambda de Wilks	F	gl1	gl2	P_valor.
Edad	0,987	2,263	1	175	0,134
Matemática	0,949	9,496	1	175	0,002
Raz. Matemático	0,955	8,266	1	175	0,005
Raz. Verbal	0,986	2,453	1	175	0,119
Física	0,985	2,611	1	175	0,108
Química	0,989	1,946	1	175	0,165
Biología	0,999	0,245	1	175	0,621

La selección de variables se realizó con el procedimiento de inclusión por pasos. De acuerdo al valor de F obtenido (el más alto valor) en la Tabla 67, la primera variable que ingresó al modelo discriminante lineal es Matemática. Además, las variables que discriminan mejor a los dos grupos de Créditos Aprobados en el primer curso de matemática son Matemática y Razonamiento Matemático.

La Tabla 68 muestra los resultados para la variable Créditos Aprobados con respecto a la prueba de igualdad de matrices de covarianza mediante la prueba M de Box, por lo cual no existe suficiente evidencia estadística para rechazar la hipótesis.

Tabla 68

Prueba M de Box para Créditos Aprobados. Ciclo 2017 II

Descripción	Créditos Aprobados 2017 II
M de Box	7,452
F Aprox.	1,217
gl 1	6
gl 2	134761,371
Sig.	0,294

Los coeficientes de la función discriminante se estimaron en 3 pasos. Se usó el estadístico lambda de Wilks global para la selección de variables. Los criterios fueron: se incorpora una variable si el valor de F es mayor de 3,84 y se elimina la variable si el valor de F es menor de 2,71.

En la Tabla 69 se muestran las variables incluidas en el modelo al finalizar el paso 3. La Tolerancia de una variable mide el grado de asociación lineal con el resto de las variables y cuando su valor se aproxima a la unidad significa que la variable no se correlaciona con el resto de las variables por lo que se debe mantener en el modelo. Siendo así, la primera variable incluida en el paso 1 fue Matemática, Química en el paso 2 y Razonamiento Matemático en el tercer paso.

Tabla 69

Variables incluidas en el análisis para Créditos Aprobados. Ciclo 2017 II

Variables	Tolerancia	F para eliminar	Lambda de Wilks
Matemática	0,888	6,220	0,926
Química	0,890	6,709	0,928
Raz. Matemático	0,839	6,444	0,927

En la Tabla 70 se muestra el resumen de las funciones discriminantes canónicas. Hay dos grupos de Créditos Aprobados y es por ello que se forma una sola función discriminante y por tanto se establece un solo

autovalor. Así también, cuanto más alto sea el valor de la correlación canónica, entonces la función discriminará mejor entre los grupos ya que existe mayor dispersión entre ellos. De los resultados se evidencia que la función discrimina bien los dos grupos establecidos, además, solo hay una función discriminante que explica el 100% de la varianza total.

Tabla 70

Autovalores para Créditos Aprobados. Ciclo 2017 II

Función	Autovalor	% de varianza	Correlación Canónica
1	0,119	100	0,326

En la Tabla 71 se muestran los resultados de la prueba de igualdad de medias entre los dos grupos de las funciones discriminantes. A un nivel de significación del 5% se rechaza la hipótesis nula de la igualdad de medias, es decir, la función discriminante es significativa y tiene un alto poder para discriminar entre las dos categorías de Créditos Aprobados, es decir, clasifica bien a los estudiantes que aprueban al menos 12 créditos y a los estudiantes que aprueban menos de 12 créditos en el primer ciclo académico 2017 II en la UNALM.

Tabla 71

Prueba Lambda de Wilks para Créditos Aprobados. Ciclo 2017 II

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	0,894	19,527	3	0,000

De la Tabla 72 se observa la capacidad discriminante que tienen las variables estudiadas, y cuanto más alto, será mejor. Los coeficientes de la función discriminante canónica estandarizada muestran la importancia de las variables Matemática, Razonamiento Matemático y Química ya que tienen mayor influencia en la función discriminante. La matriz de estructuras está ordenada según el poder discriminante de la

variable, es decir, las variables mencionadas tienen una mayor capacidad discriminante.

Tabla 72

Coeficientes y Matriz de estructuras para Créditos Aprobados. Ciclo 2017 II

Variables	Función 1	
	Coef. de func. discrim. canónica estandarizada	Matriz de Estructuras
Matemática	0,606	0,675
Raz. Matemático	0,634	0,630
Física[a]	-	0,315
Química	0,628	0,306
Biología[a]	-	-0,219
Raz. Verbal[a]	-	-0,194
Edad[a]	-	0,003

[a] la variable no se utiliza en el análisis

En la Tabla 73 se muestran los coeficientes no estandarizados de la función discriminante canónica para Créditos Aprobados que son proporcionales a los coeficientes de la función discriminante lineal de Fisher. Son importantes porque al evaluarlo para las medias de los grupos se obtienen los centroides con los que se halla el punto de corte discriminante. En este caso, los centroides son -0,435 y 0,271 para los grupos de Créditos Aprobados en el primer ciclo 2017 II de la UNALM.

Tabla 73

Coeficientes no estandarizados discriminante para Créditos Aprobados. Ciclo 2017 II

Variables	Función 1
Matemática	0,163
Raz. Matemático	0,244
Química	0,222
Constante	-7,856

De la Tabla 74 se observan las variables con mayor poder discriminante y con los coeficientes se obtuvieron las funciones de clasificación para los dos grupos de Créditos Aprobados: aprueba al menos 12 créditos o menos de 12 créditos matriculados.

Tabla 74

Coeficientes de función de clasificación para Créditos Aprobados. Ciclo 2017 II

Variables	Créditos Aprobados		Diferencia
	0	1	
Matemática	0,616	0,730	0,114
Raz. Matemático	2,261	2,433	0,172
Química	2,514	2,670	0,156
Constante	-33,977	-38,991	-5,014

Las funciones de clasificación para Créditos Aprobados de ciclo 2017 II son:

$$Z_0 = -33{,}977 + 0{,}616 * \text{Matemática} + 2{,}261 * \text{Raz. Mat} + 2{,}514 * \text{Química}$$

$$Z_1 = -38{,}991 + 0{,}730 * \text{Matemática} + 2{,}433 * \text{Raz. Mat} + 2{,}670 * \text{Química}$$

La función discriminante lineal de Fisher para la variable Créditos Aprobados del ciclo 2017 II que está dada por:

$$Z_i = -5{,}014 + 0{,}114 * \text{Matemática} + 0{,}172 * \text{Raz. Mat} + 0{,}156 * \text{Química}$$

Criterio de Decisión:

Si $Z_i < 0$, se clasifica al estudiante en el grupo aprobó menos de 12 créditos matriculados.

Si $Z_i > 0$, se clasifica el estudiante en el grupo aprobó al menos 12 créditos matriculados.

Con la función discriminante hallada se procedió a la evaluación de cada uno de los casos y se encontró el número de casos asignados correctamente a ambos grupos de Créditos Aprobados, es decir aprobó al menos 12 o menos de 12 créditos matriculados en el primer ciclo 2017 II de la UNALM.

Tabla 75

Resultados de clasificación discriminante para Créditos Aprobados. Ciclo 2017 II

			Créditos Aprobados Pronosticado		
			0	1	Total
Créditos Aprobados observado	N	0	27	41	68
		1	20	89	109
	%	0	39,7	60,3	100,0
		1	18,3	81,7	100,0

De la Tabla 75 se observa que la función discriminante lineal de Fisher permitió la clasificación correcta del 65,5% de estudiantes de primer ciclo, de los cuales se clasificó correctamente al 81,7% a los estudiantes que aprobaron al menos 12 créditos y de al 39,7% para los estudiantes que aprobaron menos de 12 créditos matriculados en el primer ciclo 2017 II en la UNALM.

4.3 Modelo de Redes Neuronales tipo Perceptrón Multicapa

4.3.1 Modelo de Redes Neuronales para Resultados 2017 I

Para la variable dependiente Resultado de estudiantes en el primer curso de matemática del ciclo 2017 I de la UNALM, se construyó una red neuronal tipo perceptrón multicapa sobre una muestra de entrenamiento conformada por 211 estudiantes (70,3%) y la muestra de validación de 89 estudiantes, es decir, el 29,7% de estudiantes.

Arquitectura de la red: Se utilizaron 3 capas, una capa de entrada con 24 neuronas, la capa oculta con 3 neuronas y la capa de salida con dos neuronas que corresponde a la variable de estudio que consta de 2 valores: 1 si es aprobado y 0 si desaprueba. La función de activación en la capa oculta fue Tangente hiperbólica y en la capa de salida fue la función Softmax. La función de error empleada fue Entropía cruzada.

Con el entrenamiento de la red, se clasificó el Resultado en el primer curso de matemática en el primer ciclo 2017 I de la UNALM.

Tabla 76

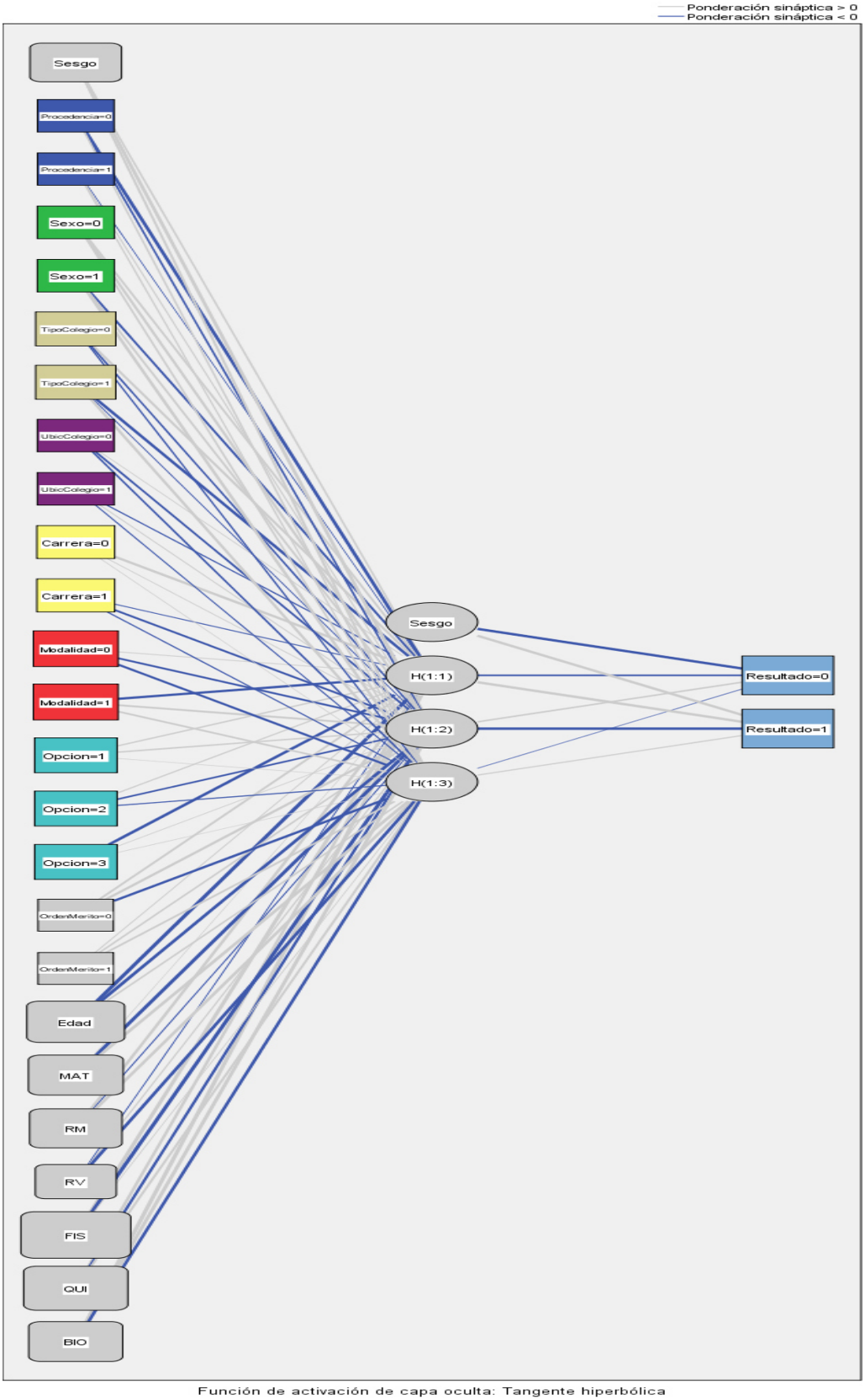
Modelo de redes neuronales para Resultado. Ciclo 2017 I

	Descripción	Valores
Entrenamiento	Error de entropía cruzada	101,549
	Porcentaje de pronósticos incorrectos	22,3%
	Regla de parada utilizada	1 paso consecutivo sin disminución del error[a]
	Tiempo de preparación	0:00:00.16
Validación	Error de entropía cruzada	37,916
	Porcentaje de pronósticos incorrectos	20,2%

[a] Los cálculos de error se basan en la muestra de comprobación.

Figura 13

Estructura de la Red Neuronal tipo Perceptrón Multicapa para Resultado. Ciclo 2017 I



En la Tabla 77 se observa que el error de clasificación fue de 22,3% y 20,2% en la muestra de entrenamiento y validación, respectivamente.

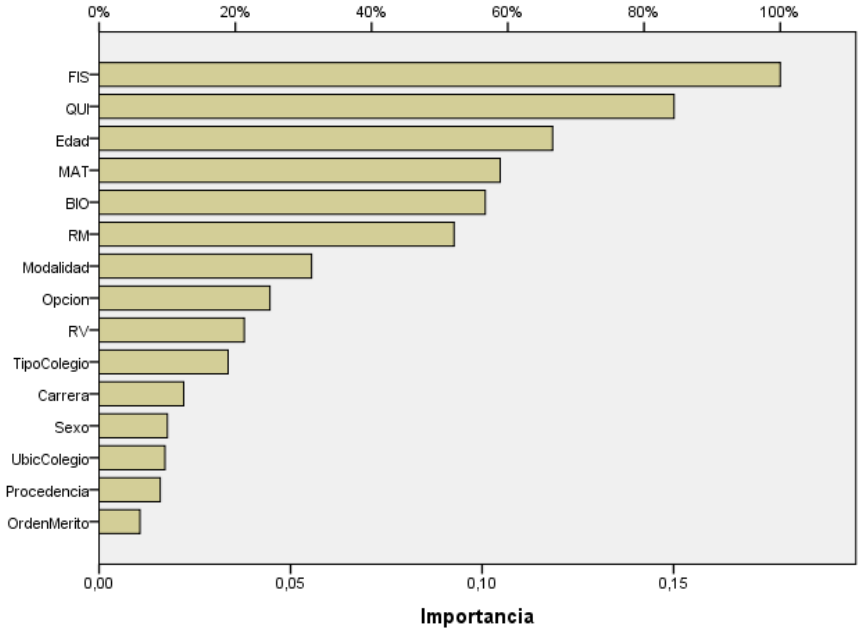
Tabla 77

Clasificación con redes neuronales para Resultado. Ciclo 2017 I

	Resultado observado	Resultado Pronosticado 0	1	Total
Muestra de Entrenamiento	0	42	27	60,9%
	1	20	122	85,9%
	% global	29,4%	70,6%	77,7%
Muestra de validación	0	16	10	61,5%
	1	8	55	87,3%
	% global	27,0%	73,0%	79,8%

Figura 14

Importancia Normalizada de las variables para Resultado. Ciclo 2017I



En la Figura 14 se observa que las cinco variables más importantes para el modelo de redes neuronales son: Física (100%), Química (84,4%), Edad (66,6%), Matemática (58,9%) y Biología (56,7%).

4.3.2 Modelo de Redes Neuronales para Resultados 2017 II

Para la variable dependiente Resultado de estudiantes en el primer curso de matemática del ciclo 2017 II de la UNALM, se construyó una red neuronal tipo perceptrón multicapa sobre una muestra de entrenamiento conformada por 182 estudiantes (71,9%) y la muestra de validación de 71 estudiantes, es decir, el 28,1% de estudiantes.

Arquitectura de la red: Se utilizaron 3 capas, una capa de entrada con 24 neuronas, la capa oculta con 5 neuronas y la capa de salida con dos neuronas que corresponde a la variable de estudio que consta de 2 valores: 1 si es aprobado y 0 si desaprueba. La función de activación en la capa oculta fue Tangente hiperbólica y en la capa de salida fue la función Softmax. La función de error empleada fue Entropía cruzada.

Con el entrenamiento de la red, se clasificó el Resultado en el primer curso de matemática en el primer ciclo 2017 II de la UNALM.

Tabla 78

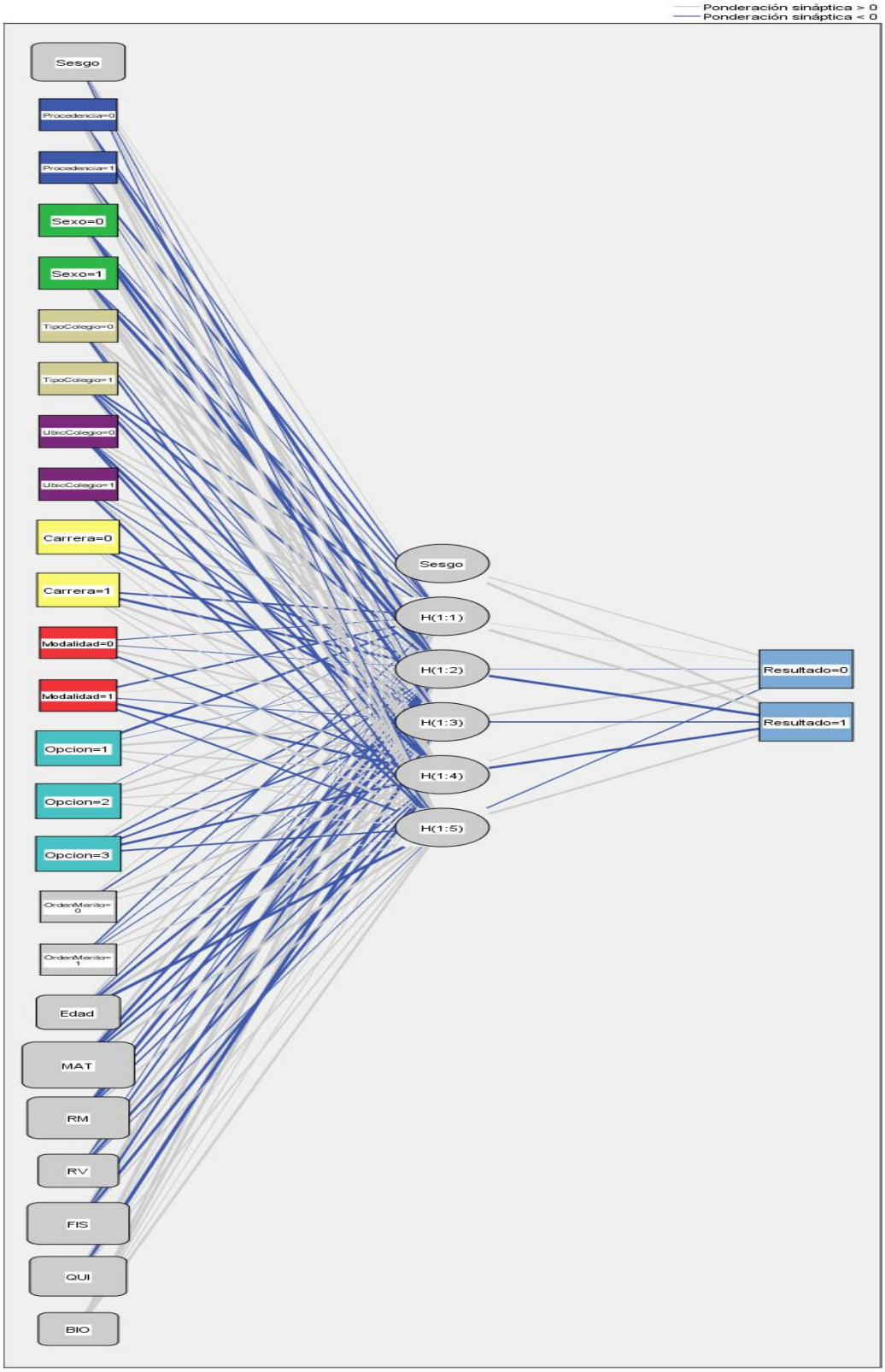
Modelo de redes neuronales para Resultado. Ciclo 2017 II

	Descripción	Valores
Entrenamiento	Error de entropía cruzada	111,179
	Porcentaje de pronósticos incorrectos	36,8%
	Regla de parada utilizada	1 paso consecutivo sin disminución del error[a]
	Tiempo de preparación	0:00:00.13
Validación	Error de entropía cruzada	39,227
	Porcentaje de pronósticos incorrectos	23,9%

[a] Los cálculos de error se basan en la muestra de comprobación.

Figura 15

Estructura de la Red Neuronal tipo Perceptrón Multicapa para Resultado. Ciclo 2017 II



En la Tabla 79 se observa que el error de clasificación fue de 36,8% y 23,9% en la muestra de entrenamiento y validación, respectivamente.

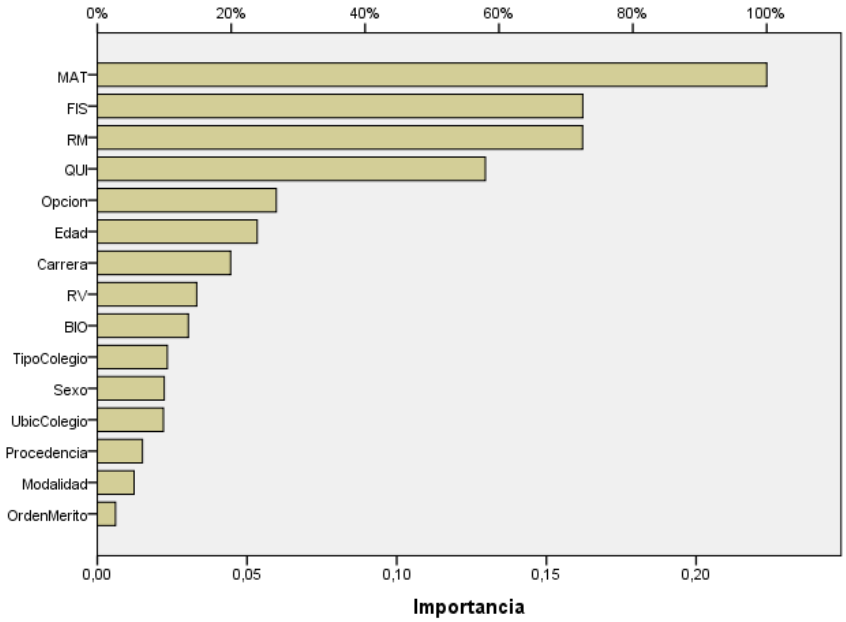
Tabla 79

Clasificación con redes neuronales para Resultado. Ciclo 2017 II

	Resultado observado	Resultado Pronosticado 0	1	Total
Muestra de Entrenamiento	0	57	27	67,9%
	1	40	58	59,2%
	% global	53,3%	46,7%	63,2%
Muestra de validación	0	23	8	74,2%
	1	9	31	77,5%
	% global	45,1%	54,9%	76,1%

Figura 16

Importancia Normalizada de las variables para Resultado. Ciclo 2017 II



En la Figura 16 se observa que las cinco variables más importantes para el modelo de redes neuronales son: Matemática (100%), Física y Razonamiento Matemático (72,5%), Química (58%) y Opción de ingreso (26,7%)

4.3.3 Modelo de Redes Neuronales para Créditos Aprobados 2017 I

Para la variable dependiente Créditos Aprobados, se construyó una red neuronal tipo perceptrón multicapa sobre una muestra de entrenamiento conformada por 201 estudiantes (67%) y la muestra de validación de 99 estudiantes, es decir, el 33% de estudiantes.

Arquitectura de la red: Se utilizaron 3 capas, una capa de entrada con 24 neuronas, la capa oculta con 2 neuronas y la capa de salida con dos neuronas que corresponde a las dos categorías de la variable: aprueba 12 o más créditos y menos de 12 créditos. La función de activación en la capa oculta fue Tangente hiperbólica y en la capa de salida fue la función Softmax. La función de error empleada fue Entropía cruzada.

Con el entrenamiento de la red, se clasificó a los estudiantes ingresantes de acuerdo a los Créditos Aprobados en el primer ciclo de estudios en la UNALM.

Tabla 80

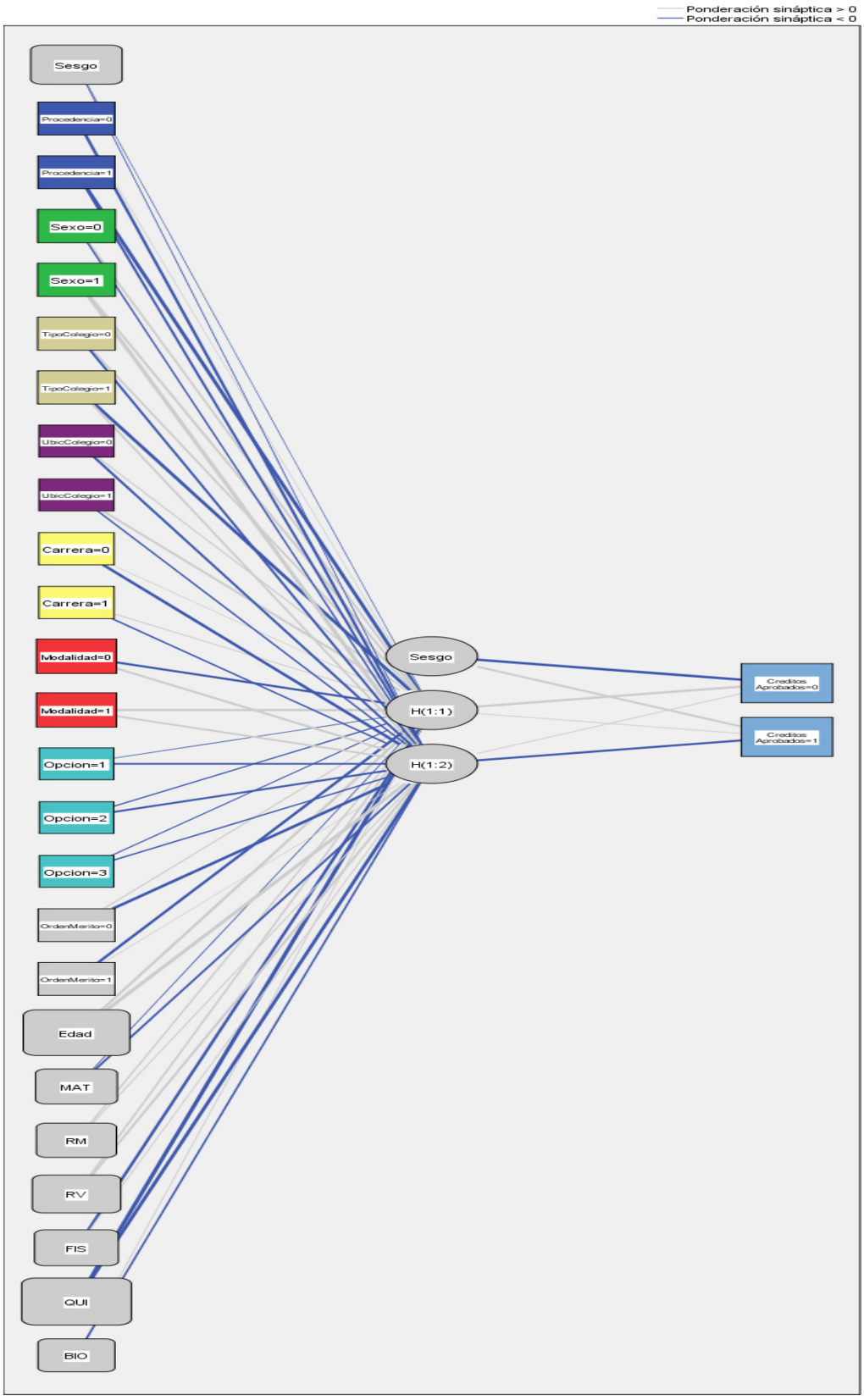
Modelo de redes neuronales para Créditos Aprobados. Ciclo 2017 I

	Descripción	Valores
Entrenamiento	Error de entropía cruzada	84,613
	Porcentaje de pronósticos incorrectos	17,9%
	Regla de parada utilizada	1 pasos consecutivos sin disminución del error[a]
	Tiempo de preparación	0:00:00.25
Validación	Error de entropía cruzada	45,557
	Porcentaje de pronósticos incorrectos	22,2%

[a] Los cálculos de error se basan en la muestra de comprobación.

Figura 17

Estructura de la Red Neuronal tipo Perceptrón Multicapa para Créditos Aprobados. Ciclo 2017 I



En la Tabla 81 se observa que el error de clasificación fue de 17,4% y 22,2% en la muestra de entrenamiento y validación, respectivamente.

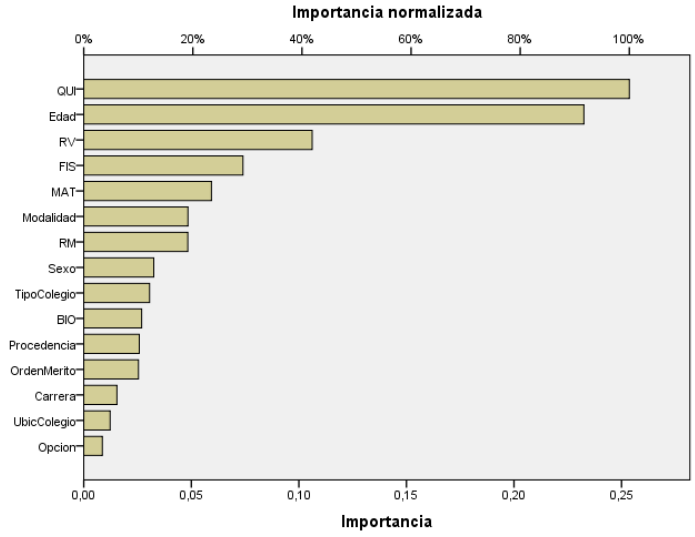
Tabla 81

Clasificación con redes neuronales para Créditos Aprobados. Ciclo 2017 I

	Créditos Aprobados observado	Créditos Aprobados Pronosticado		
		0	1	Total
Muestra de Entrenamiento	0	2	34	05,6%
	1	1	164	99,4%
	% global	01,5%	98,5%	82,6%
Muestra de validación	0	1	21	04,5%
	1	1	76	98,7%
	% global	02,0%	98,0%	77,8%

Figura 18

Importancia Normalizada de las variables para Créditos Aprobados. Ciclo 2017 I



En la Figura 18 se observa que las cinco variables más importantes para el modelo de redes neuronales son: Química (100%), Edad (91,7%), Razonamiento Verbal (41,9%), Física (29,2%) y Matemática (23,4%).

4.3.4 Modelo de Redes Neuronales para Créditos Aprobados 2017 II

Para la variable dependiente Créditos Aprobados, se construyó una red neuronal tipo perceptrón multicapa sobre una muestra de entrenamiento conformada por 178 estudiantes (70,4%) y la muestra de validación de 75 estudiantes, es decir, el 29,6% de estudiantes.

Arquitectura de la red: Se utilizaron 3 capas, una capa de entrada con 24 neuronas, la capa oculta con 3 neuronas y la capa de salida con dos neuronas que corresponde a las dos categorías de la variable: aprueba 12 o más créditos y menos de 12 créditos. La función de activación en la capa oculta fue Tangente hiperbólica y en la capa de salida fue la función Softmax. La función de error empleada fue Entropía cruzada.

Con el entrenamiento de la red, se clasificó a los estudiantes ingresantes de acuerdo a los Créditos Aprobados en el primer ciclo de estudios en la UNALM.

Tabla 82

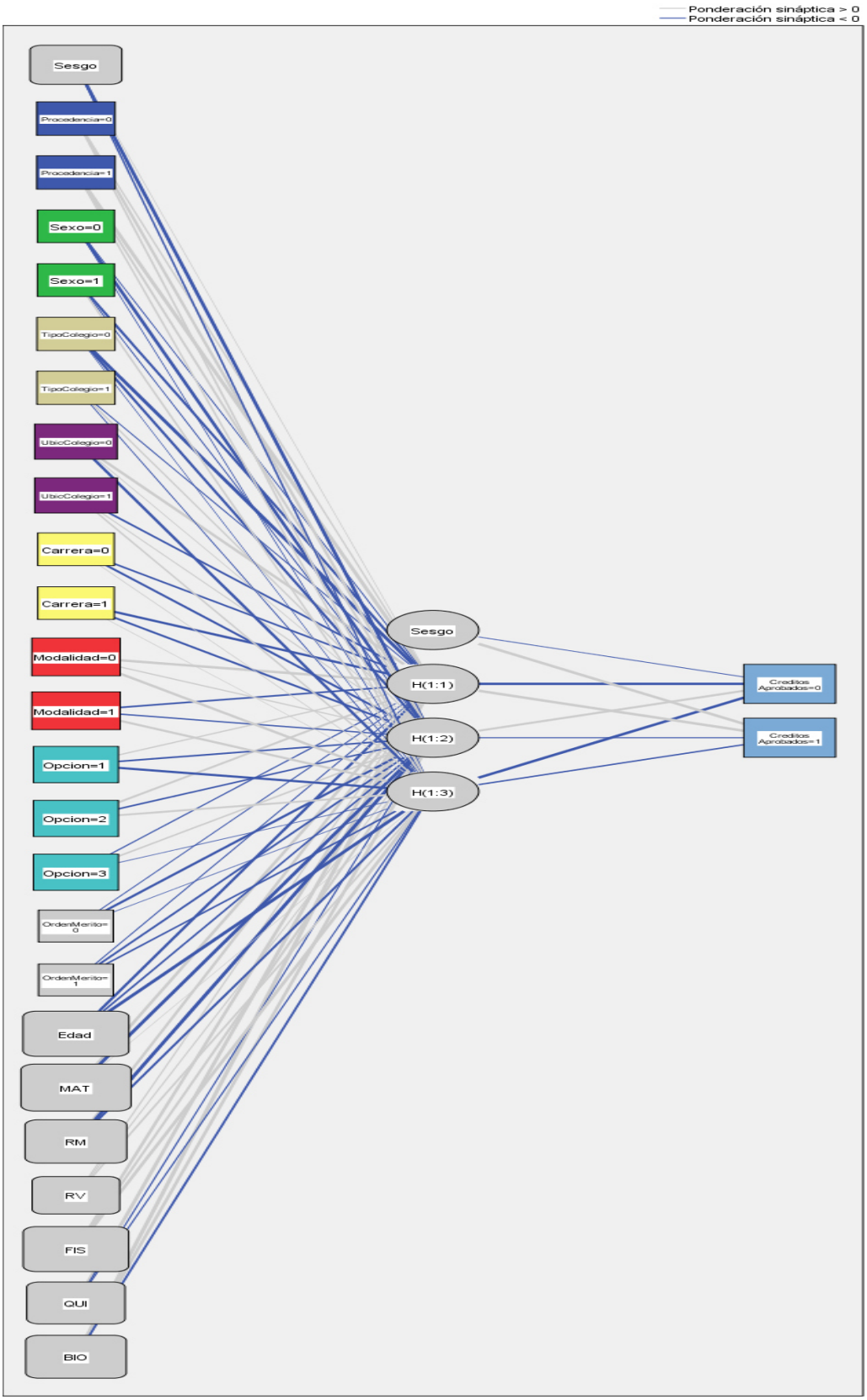
Modelo de redes neuronales para Créditos Aprobados. Ciclo 2017 II

	Descripción	Valores
Entrenamiento	Error de entropía cruzada	100,866
	Porcentaje de pronósticos incorrectos	30,3%
	Regla de parada utilizada	1 pasos consecutivos sin disminución del error[a]
	Tiempo de preparación	0:00:00.11
Validación	Error de entropía cruzada	44,909
	Porcentaje de pronósticos incorrectos	32,0%

[a] Los cálculos de error se basan en la muestra de comprobación

Figura 19

Estructura de la Red Neuronal tipo Perceptrón Multicapa para Créditos Aprobados. Ciclo 2017 II



En la Tabla 83 se observa que el error de clasificación fue de 30,3% y 32,0% en la muestra de entrenamiento y validación, respectivamente.

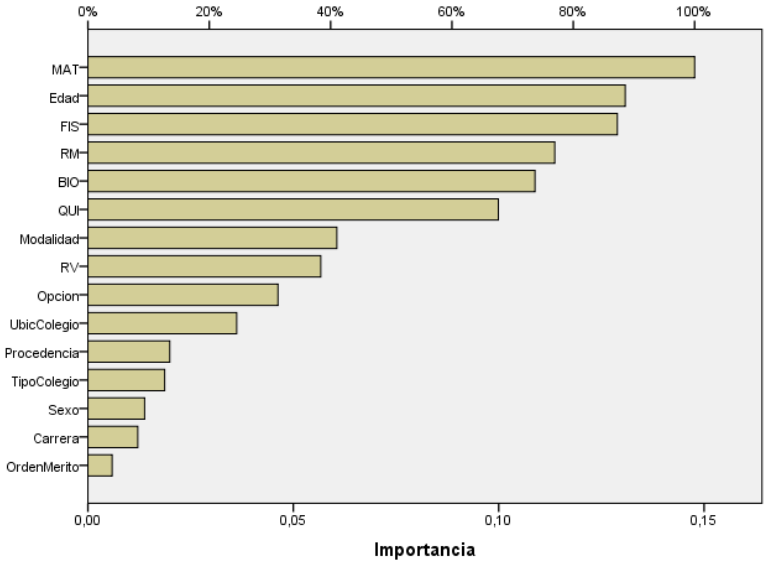
Tabla 83

Clasificación con redes neuronales para Créditos Aprobados. Ciclo 2017 II

	Créditos Aprobados observado	Créditos Aprobados Pronosticado		
		0	1	Total
Muestra de Entrenamiento	0	29	35	45,3%
	1	19	95	83,3%
	% global	27,0%	73,0%	69,7%
Muestra de validación	0	20	13	60,6%
	1	11	31	73,8%
	% global	41,3%	58,7%	68,0%

Figura 20

Importancia Normalizada de las variables para Créditos Aprobados. Ciclo 2017 II



En la Figura 20 se observa que las cinco variables más importantes para el modelo de redes neuronales son: Matemática (100%), Edad (88,6%), Física (87,3%), Razonamiento Matemático (77%) y Biología (73,7%).

4.4 Técnicas de evaluación de clasificadores

Se realizó el análisis comparativo de los tres modelos propuestos para el presente estudio. La comparación se aplicó sobre la muestra de validación estuvo conformada por 166 estudiantes que representa el 30% del total de estudiantes para los modelos discriminante y logístico. En cambio, para el modelo de redes neuronales se empleó una muestra de 153 estudiantes que representan el 28% del total de estudiantes. Los indicadores fueron: Tasa de clasificación correcta, Sensibilidad, Especificidad y Área bajo la curva ROC. A continuación, se muestran los indicadores de validez de la capacidad discriminatoria de los modelos obtenidos para el desempeño académico:

Tabla 84

Comparativo para Resultado en el primer curso de matemática. Ciclo 2017 I

Modelo	Tasa de clasificación correcta	Sensibili-dad	Especifi-cidad	Área Curva ROC
Discriminante	0,789	0,500	0,906	0,807
Logístico	0,822	0,538	0,938	0,849
Redes neuronales	0,798	0,873	0,615	0,830

En la Tabla 84 se puede observó que el modelo que discriminó mejor el Resultado de los estudiantes de primer ciclo académico en el curso de Matemática en el ciclo 2017 I de la UNALM, es el modelo de regresión logística con una tasa de clasificación correcta del 82,2%. Los estudiantes que aprobaron el primer curso se estimaron con una precisión moderada del 53,8% (Sensibilidad), además, los estudiantes que desaprobaron el primer curso de matemática se estimaron con una alta precisión del 93,8% (Especificidad) y finalmente, se obtuvo casi el 85% del área bajo la curva ROC. Sin embargo, se debe señalar que el modelo de redes neuronales permitió una alta precisión (87,3%) en la clasificación de estudiantes que aprobaron el primer curso.

Tabla 85

Comparativo para Resultado en el primer curso de matemática. Ciclo 2017 II.

Modelo	Tasa de clasificación correcta	Sensibili-dad	Especifi-cidad	Área Curva ROC
Discriminante	0,605	0,438	0,727	0,640
Logístico	0,622	0,511	0,701	0,652
Redes neuronales	0,761	0,775	0,742	0,745

En la Tabla 85 se observó que el modelo que discriminó mejor el Resultado de los estudiantes de primer ciclo 2017 II en el curso de Matemática de la UNALM es el de Redes Neuronales tipo Perceptrón Multicapa con una tasa de clasificación correcta del 76,1%. Los estudiantes que aprobaron el primer curso se estimaron con una precisión moderada del 77,5% (Sensibilidad) y los estudiantes que desaprobaron el primer curso de matemática se estimaron con una precisión del 74,2% (Especificidad). Además, con el modelo Redes Neuronales tipo Perceptrón Multicapa se obtuvo el 74,5% de área bajo la curva ROC. Los indicadores del modelo de redes neuronales fueron muy superiores a los de los modelos discriminante y logístico.

Tabla 86

Comparativo para Créditos Aprobados. Ciclo 2017 I.

Modelo	Tasa de clasificación correcta	Sensibili-dad	Especifi-cidad	Área Curva ROC
Discriminante	0,789	0,250	0,943	0,810
Logístico	0,744	0,300	0,871	0,840
Redes neuronales	0,778	0,987	0,045	0,747

Se observó en la Tabla 86 que el modelo que discrimina mejor los dos grupos de estudiantes con menos de 12 créditos y 12 o más créditos aprobados en el primer ciclo académico 2017 I en la UNALM es el modelo discriminante con una tasa de clasificación correcta del 78,9%.

Los estudiantes que aprueban 12 o más créditos matriculados se estimaron con una baja precisión del 25% (Sensibilidad) y los estudiantes que aprueban menos de 12 créditos matriculados se estimaron con una alta precisión del 94,3% (Especificidad). Cabe señalar, que con el modelo logístico se alcanzan indicadores muy cercanos al obtenido con el modelo discriminante, sin embargo, se debe señalar que el modelo de redes neuronales permite una alta precisión (98,7%) para clasificar a los estudiantes que culminaron el primer ciclo de estudios con menos de 12 créditos y 12 o más créditos.

Tabla 87

Comparativo para Créditos Aprobados. Ciclo 2017 II.

Modelo	Tasa de clasificación correcta	Sensibili-dad	Especifi-cidad	Área Curva ROC
Discriminante	0,658	0,286	0,875	0,696
Logístico	0,697	0,500	0,813	0,766
Redes neuronales	0,680	0,738	0,606	0,733

En la Tabla 87 se observó que los modelos logístico y de redes neuronales tipo perceptrón multicapa presentan indicadores muy similares para clasificar mejor a los estudiantes que culminaron el primer ciclo de estudios con menos de 12 créditos y 12 o más créditos aprobados. Siendo importante identificar a los estudiantes que aprueben 12 o más créditos, se seleccionó al modelo de redes neuronales tipo perceptrón multicapa con una tasa de clasificación correcta del 68,0%. Los estudiantes que aprueban 12 o más créditos matriculados se estimaron con una alta precisión del 73,8% (Sensibilidad) y los estudiantes que aprueban menos de 12 créditos matriculados se estimaron con una alta precisión del 60,6% (Especificidad), cabe señalar que, en este indicador, el modelo discriminante presentó un valor más alto de 87,5%. Además, con el modelo de redes neuronales tipo perceptrón multicapa se obtuvo el 73,3% de área bajo la curva ROC.

CONCLUSIONES

1. Se compararon tres métodos multivariantes: regresión logística binaria, discriminante y de redes neuronales tipo perceptrón multicapa a través del análisis del desempeño académico de los estudiantes de primer ciclo de la UNALM, medido por el Resultado y los Créditos Aprobados. El Resultado mide el promedio final obtenido por el estudiante en el curso de matemática en el primer ciclo de estudios universitarios con dos categorías: Aprobado y Desaprobado. Los Créditos Aprobados mide el número de créditos aprobados al finalizar el primer ciclo de estudios, también, con dos categorías: Doce o más créditos y Menos de doce créditos aprobados.

2. El modelo de regresión logística binaria clasificó mejor el desempeño académico para el Resultado en el ciclo 2017 I, con una tasa de clasificación correcta de 82,2% considerado como alta precisión. Con este modelo se encontró la mejor clasificación correcta (93,8%) para el grupo de estudiantes que desaprobaron el curso de matemática. Así también, con este modelo se encontró la mejor capacidad discriminante entre los dos grupos de estudiantes que aprobaron o no el curso de matemática, con un 84,9% de área bajo la curva ROC.

   Con el modelo logístico para los Créditos Aprobados en el ciclo 2017 II, se encontró la mejor capacidad discriminante (76,6%) entre los dos grupos de estudiantes.

3. El modelo del análisis discriminante clasificó mejor el desempeño académico para los Créditos Aprobados en el ciclo 2017 I, con una tasa de clasificación correcta de 78,9% que fue considerado como alta precisión, sin embargo, esta tasa fue muy cercana a los obtenidos por los otros dos modelos. De igual forma clasificó con un acierto de 94,3% a los estudiantes que aprobaron menos de doce créditos, seguido del modelo de regresión logística que obtuvo el 87,1% de acierto. Así también, con el modelo discriminante para el ciclo 2017 II, la clasificación de estudiantes que aprobaron menos de doce créditos tuvo un acierto máximo de 87,5%.

4. El modelo de redes neuronales tipo perceptrón multicapa clasificó con mayor precisión el desempeño académico de los estudiantes según el Resultado en el ciclo 2017 II. La tasa de clasificación correcta fue 76,1%. Con respecto a la clasificación correcta del grupo de estudiantes que aprobaron el curso de matemática, se halló un 77,5% de acierto mediante este modelo y la mejor clasificación correcta del grupo de estudiantes que desaprobaron el curso de matemática fue de 74,2%. De igual forma, este modelo ofreció la mejor capacidad discriminante entre los dos grupos de estudiantes que aprobaron o no el curso de matemática, con un 74,5% de área bajo la curva ROC. Así también, con el modelo de redes neuronales se encontró una clasificación correcta de los estudiantes que aprobaron doce o más créditos en el ciclo 2017 II, con una alta precisión de 73,8%.

5. Con respecto a las variables intervinientes para el ciclo 2017 I, las que más influyeron en el promedio final del curso de matemática fueron las notas obtenidas en el examen de admisión en las áreas de Matemática, Física y Química, y para clasificar los créditos aprobados por los estudiantes al culminar el primer ciclo de estudios fueron las notas de Física, Química, Razonamiento Verbal y Matemática que también corresponden al factor de admisión. Por lo que concluye los factores sociodemográficos y escolares no estarían asociados al desempeño académico de los estudiantes de primer ciclo.

6. Con respecto a las variables intervinientes para el ciclo 2017 II, las que más influyen en el promedio final del curso de matemática fueron las notas obtenidas en el examen de admisión en las áreas de Matemática, Física y Razonamiento Matemático. Con respecto a la clasificación de los estudiantes según los créditos aprobados al finalizar el primer ciclo, las variables más importantes fueron Matemática, Edad, Física y Razonamiento Matemático, que corresponden a los factores de admisión y sociodemográfico. Por lo tanto, el factor escolar no estaría incidiendo en el desempeño académico de los estudiantes en el primer ciclo de estudios universitarios.

7. De los resultados obtenidos, se concluye que las mejores precisiones en la clasificación de los estudiantes de primer ciclo de acuerdo a su desempeño académico, se obtuvieron con los modelos de regresión logística y de redes neuronales. El factor de admisión resultó el de mayor importancia en el proceso de clasificación y la nota promedio en el área de matemática del examen de admisión fue la variable incluida en los modelos obtenidos bajo los tres métodos multivariantes estudiados. Estos resultados son importantes debido a que presentan un nuevo escenario a tomar en cuenta en la gestión académica de la universidad.

8. Este estudio nos revela que sí es posible encontrar perfiles o características en los estudiantes que permitan clasificarlos a priori en función de su desempeño académico en especial en las áreas de matemática que son materia de numerosos trabajos de investigación para identificar las características más importantes en los estudiantes con o sin éxito en el primer ciclo de la universidad. Para ello el empleo de los métodos multivariantes juegan un rol decisivo en la búsqueda de nuevos indicadores que permitirán definir las acciones de control necesarias para mitigar los resultados adversos.

RECOMENDACIONES

1. Resulta conveniente incluir en el análisis a modelos multivariantes como el de ecuaciones estructurales, análisis cluster, de correspondencias, entre otros. Además, para futuros estudios de investigación sobre la clasificación del desempeño de estudiantes universitarios, se recomienda la inclusión de factores alternativos que también permitan diferenciar el desempeño académico de los estudiantes, tales como el psicológico, motivacional, estrategias de aprendizaje, hábitos de estudios, tipo de preparación preuniversitaria, entre otros.

2. Con la finalidad de realizar un estudio sobre el desempeño de estudiantes universitarios a lo largo de todo el periodo de estudios desde la fecha de ingreso hasta la graduación, se recomienda realizar un estudio de investigación con un diseño de tipo longitudinal a fin de realizar el seguimiento al desenvolvimiento académico de los estudiantes.

3. También se recomienda realizar el análisis comparativo de los modelos de clasificación del desempeño académico de los estudiantes de primer ciclo de estudios universitarios que se encuentren matriculados no solo en el curso de matemática sino también en las demás asignaturas básicas del programa de Estudios Generales.

REFERENCIAS BIBLIOGRÁFICAS

Aguiar, M., Gutiérrez, H., Lara, A. & Villalpando, J. (2011). El rendimiento académico de las mujeres en matemáticas: análisis bibliográfico y un estudio de caso en educación superior en México. Revista electrónica "Actualidades investigativas en Educación" 11(2). Costa Rica. Recuperado de https://www.redalyc.org/pdf/447/44720020016.pdf

Aguirre, E. (2013). *Comparación de las redes neuronales perceptrón multicapa y la regresión logística en el desarrollo de modelos de credit Scoring de clientes de un banco.* (tesis de pregrado). Universidad Nacional Agraria La Molina. Lima, Perú.

Anatomía topográfica (2019) Neurona multipolar [Gráfico]. Recuperado de https://www.anatomiatopografica.com/celulas/neuronas-multipolares/

Bazán, J. & Sotero H. (1997). Una aplicación al estudio de actitudes hacia la matemática en la UNALM. *Anales Científicos UNALM*. Recuperado de http://argos.pucp.edu.pe/~jlbazan/download/1998_62.pdf

Basogain, X. (2009). *Redes Neuronales Artificiales y sus Aplicaciones*. Escuela superior de Ingeniería de Bilbao, Departamento de Ingeniería de Sistemas y Automática. Recuperado de http://www.ciberesquina.una.edu.ve:8080/2014_2/350_E.pdf

Castro E. & Ávila J. (2013). *Acta Latinoamericana de Matemática Educativa*, Vol. 26. Flores R. Colegio Mexicano de Matemática Educativa A. C. y Comité Latinoamericano de Matemática Educativa A. C. D. F. México. Recuperado de https://core.ac.uk/download/pdf/20482748.pdf

Castro, J. (2006). *Fundamentos para la implementación de red neuronal perceptrón multicapa mediante software*. (tesis de pregrado). Universidad de San Carlos de Guatemala. Guatemala.

Catena, A., Ramos, M. & Trujillo, H. (2003). *Análisis Multivariado.Un manual para investigadores.* Madrid, España: Biblioteca Nueva.

Cea, M. (2002). *Análisis Multivariable. Teoría y Práctica en la investigación social*. Madrid, España: Síntesis.

Closas, A., Hisgen, C. & Sanz, M. (2017). Estrategias de aprendizaje y su relación con el rendimiento académico mediante regresión logística. *Cuaderno de Pedagogía Universitaria, 13* (25), 8-20. Recuperado de https://cuaderno.pucmm.edu.do/index.php/cuadernodepedagogia

Cuadras, C. (2018). *Nuevos Métodos de Análisis Multivariante.* Barcelona, España: CMC Editions.

De Diego, S. (2015). *Análisis discriminante: aplicación a la búsqueda de variables relacionadas con la capacidad de exportar.* (trabajo para obtención de grado). Universidade da Coruña. España.

De Guzmán, M. (2007). Enseñanza de las ciencias y la matemática. *Revista Iberoamericana de Educación* (043). Organización de Estados Iberoamericanos para la Educación, la Ciencias y la Cultura. Madrid, España.

De la Fuente, S. (2011). *Análisis Discriminante*. Universidad Autónoma de Madrid. España.

De la Garza, J., Morales, B. & Gonzáles, B. (2013). *Análisis Estadístico Multivariante: un enfoque teórico y práctico*. México. Editorial McGraw-Hill.

Del Valle, A. (2017). *Curva ROC y sus aplicaciones*. (tesis de pregrado) Universidad de Sevilla. Sevilla, España.

Edel, R. (2003). El rendimiento académico: concepto, investigación y desarrollo. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*. Recuperado de http://www.redalyc.org/ articulo.oa?id =5511 0208

Fernández, R. (2016). Neuropsicología aplicada a la educación: implicación de las funciones ejecutivas en el desarrollo lecto-escritor. Programa de intervención. Área de innovación y Desarrollo, S.L. Recuperado de https: //issuu.com/3ciencias/docs/neuropsicolog__a_aplicada_a_la_educ.

Ferreyra, M. (2007). Determinantes del desempeño universitario: Efectos heterogéneos de un modelo censurado. (tesis de maestría). Universidad Nacional de la Plata. Argentina. Recuperado de http://www.depeco.econ o.unlp.edu.ar/maestria/tesis/048-tesisferreyra . pdf

Figuera, P., Dorio, I., & Forner, Àngel. (2003). Las competencias académicas previas y el apoyo familiar en la transición a la universidad. *Revista de Investigación Educativa*, 21(2), 349-369. Recuperado a partir de https://revistas.um.es/rie/article/view/99251.

Flores, R. (2015). *El rendimiento académico en matemática del estudiante de la Escuela Normal Mixta del Litoral Atlántico determina la aprobación de Matemática General en la Universidad Pedagógica Nacional Francisco Morazán*. (tesis de maestría). Universidad Pedagógica Nacional Francisco Morazán. San Pedro Sula, Honduras.

Fuentes, R. (2006). Factores que afectan al rendimiento o académico del estudiante de enfermería de la Universidad Mariano Gálvez. (tesis de pregrado). Universidad Mariano Gálvez. Guatemala, Guatemala.

Gonzáles, D. (2015). *Relación entre el rendimiento académico en matemáticas y variables afectivas y cognitivas en estudiantes preuniversitarios de la Universidad Católica Santo Toribio de Mogrovejo.* (tesis doctoral). Universidad de Málaga. Málaga, España.

Gonzáles, M. (2017). Relación entre el puntaje de ingreso y el rendimiento académico de los estudiantes de la Facultad de Ciencias Matemáticas de la UNMSM que ingresaron mediante los exámenes de admisión 2010-2013. *PESQUIMAT*, 20(2): 41-52.

Hair, J., Anderson, R., Tatham, R. & Black, W. (1999). *Análisis multivariante.* Madrid, España: Prentice Hall.

Hernández, F. & Correa, J. (2009). Comparación entre tres técnicas de clasificación. *Revista Colombiana de Estadística*, 32 (2), 247-265.

Hernández, R., Fernandez, C. & Baptista, M. (2006). *Metodología de la investigación*. México D.F., México: McGraw-Hill Interamericana.

Iglesias, T. (2013). Métodos de Bondad de Ajuste en Regresión Logística. Trabajo Máster en Estadística Aplicada. Universidad de Granada. Recuperado de https://masteres.ugr.es/moea/pages/tfm-1213/tfm_iglesiascabo_tania/!

Jiménez, R. (1998). *Metodología de la Investigación. Elementos básicos para la investigación clínica.* La Habana, Cuba: Ciencias Médicas.

Johnson, D. (2004), *Métodos multivariados aplicados al análisis de datos*, México DF, México: International Thomson Editores.

Johnson, R. & Wichern, D. (2007), *Applied Multivariate Statistical Analysis*, New Jersey, Estados Unidos: Pearson Prentice Hall.

Kovalevski, L. & Macat, P. (2012). *Decimoséptimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística*, Recuperado de https://www.fcecon.unr.edu.ar/web-nueva/sites/default /files/u16/Decimocuartas/kovalevski_macat_alternativas_no_parametricas.pdf.

Larrañaga, P., Inza, I. & Moujahid, A. (1997) Tema 8. Redes Neuronales. *Redes Neuronales*. Universidad del País Vasco. España. Recuperado de http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t8neuronales.pdf

Lévy, J. & Valera, J. (2008). *Análisis Multivariable para las ciencias sociales*. Madrid, España: Pearson Educación.

Lizares, M. (2017). *Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico.* (tesina de pregrado). Universidad Nacional Mayor de San Marcos. Lima, Perú.

Llerena, G. (2017). *La innovación educativa y su relación con el rendimiento académico de estudiantes del curso de economía general de la UNALM. Lima, Perú: Fondo editorial UNALM.*

López, M., Arballo, M., Porcel, E., Mata, L. & Barreto, S. (2012). *Redes Neuronales para predecir el rendimiento académico de los alumnos ingresantes a la carrera de Bioquímica de la FACENA-UNNE en función de sus conocimientos matemáticos previos*. Semantic Scholar. Recuperado de https://pdfs.semanticscholar.org/80be/76e7b3c256b3e765603ab077ddfb14264883.pdf

López-Roldán, P. & Fachelli, S. (2016). *Análisis de regresión logística en Metodología de la Investigación Social Cuantitativa.* Barcelona, España. Universitat Autònoma de Barcelona. (1a ed.). Recuperado de http://ddd.uab.cat/record/163570.

Madrid, J. (2017). *Propuesta de un modelo estadístico para caracterizar y predecir la deserción estudiantil Universitaria.* (tesis de maestría). Universidad Nacional de Colombia. Medellín, Colombia.

Martín, Q. & De Paz, Y. (2007). *Aplicación de las redes neuronales artificiales a la regresión.* Madrid, España: La Muralla.

Marin, J. (2012). *Introducción a las redes neuronales aplicadas*. Universidad Carlos III de Madrid. España. Recuperado de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Expertos/PresNN12.pdf

Marin, J. (s.f.). *Modelos avanzados con Regresión Logística*. Universidad Carlos III de Madrid. España. Recuperado de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Categor/Tema4Cate.pdf

Ministerio de Educación. (2017). *¿Cuánto aprenden nuestros estudiantes? Resultados de la Evaluación Censal de Estudiantes ECE 2016*. Recuperado de http://umc.minedu.gob.pe/wp-content/uploads/2017/04/Resultados-Nacionales-2016.pdf

Ministerio de Educación. (2017). *El Perú en PISA 2015 Informe nacional de resultados*. Lima, Perú. Recuperado de http://umc.minedu.gob.pe/wp-content/uploads/2017/04/Libro_PISA.pdf

Ministerio de Educación. (2016). *La competencia matemática en estudiantes peruanos de 15 años. Predisposiciones de los estudiantes y sus oportunidades para aprender en el marco de PISA 2012.* Publicaciones 2012 Serie de Evaluaciones y Factores Asociados. Lima, Perú. Recuperado de http://umc.minedu.gob.pe/wp-content/uploads/2016/05/Estudio_Pisa_web_VF.pdf

Miranda, C. (2009). *Estudio comparativo del clasificador Naive-Bayes con el análisis discriminante lineal y la regresión logística.* (tesis de maestría). Universidad Nacional Agraria La Molina. Lima, Perú.

Montaño, J. (2002). *Redes Neuronales Artificiales aplicadas al Análisis de Datos*. (tesis doctoral). Universitat de Les Illes Balears. España.

Muñoz, E. (2015). *Aplicación de redes neuronales y regresión logística para predecir el éxito de la compra de deuda de una entidad financiera*. (tesis de maestría). Universidad Nacional Agraria La Molina. Lima, Perú.

OECD. (2016). *PISA 2015 Results (I): Excellence and Equity in Education*. Paris, Francia. Recuperado de http://dx.doi.org/10.1787/9789264266490

OECD. (2019). *OECD*. Recuperado de: oecd.org

OECD. *El programa PISA de la OCDE, qué es y para qué sirve*. Organización para la Cooperación y el Desarrollo Económico OCDE. Paris, Francia.

Pano, A. (2015). *Realización FPGA de una red neuronal artificial para reproducir comportamiento caótico.* (tesis de maestría). Instituto Nacional de Astrofísica óptica y electrónica. Puebla, México.

Pérez, C. (2013). *Análisis multivariante de datos*. Madrid, España: Ibergaceta Grupo Editorial.

Pérez, C. (2004). *Técnicas de Análisis Multivariante de Datos*. Madrid, España: Pearson Educación.

Pérez, F. Ocaris, & Fernández, H. (2007). Las Redes Neuronales y la Evaluación del Riesgo de Crédito. *Revista Ingenierías Universidad de Medellín,* 6(10), 77-91. Recuperado de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1692-33242007000100007&lng=en&tlng=es.

Picón, C. (2011). ¿Son más corruptos los países menos abiertos a los mercados internacionales? Aplicación de un modelo predictivo de clasificación basado en redes neuronales. *Revista de Economía del Caribe* (8).

Porcel, E., Dapozo, G., López, M. (2009). *Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios*. XI Workshop de Investigadores en Ciencias de la Computación. Red de Universidades con Carreras en Informática (RedUNCI), Universidad de La Plata, 635-639. Recuperado de http://sedici.unlp.edu.ar/handle/10915/19846

Rado, J. (2014). *Predicción del rendimiento en el examen de admisión a la UNALM utilizando técnicas de análisis discriminante lineal y análisis discriminante con algoritmos genéticos*. (tesis de pregrado). Universidad Nacional Agraria La Molina. Lima, Perú.

Real Academia Española. (s.f.). *Diccionario de la lengua española* (23.3a ed.). Recuperado de https://dle.rae.es/

Ramón, P. & Plasencia, S. (2010). *Factores relacionados con el rendimiento académico en matemática en los estudiantes de la Universidad Nacional de Educación "Enrique Guzmán y Valle" en el año 2010*. Recuperado de http://www.une.edu.pe/investigacion/CIE%20CIENCIAS%202010/CIE-2010-88%20RAMON%20PEDRO.pdf

Rivera, E. (2005). Introducción a las Redes Neuronales. *Revista Científica, (6), pp. 59-70* . Recuperado de http://hdl.handle.net/11715/238

Rodríguez, M. & Mora, R. (2001). Análisis discriminante en Estadística Informática: casos y ejemplos con el SPSS. Publicaciones de la Universidad de Alicante. España. Recuperado de http://hdl.handle.net/ 10045/12080

Saavedra, L., Ramos, J., Mita, M. & Del Águila, V. (2017). Detección temprana del rendimiento académico de estudiantes universitarios de primer ciclo mediante el análisis discriminante. Ingeniería Industrial, (35), 77-98. ISSN: 1025-9929. Recuperado de https://www.redalyc.org /articulo.oa?id=3374/337453922004

Salazar, C. (2005). *Análisis discriminante con la metodología CART y comparación con el método de regresión logística*. (tesis de pregrado). Universidad Nacional Agraria La Molina. Lima, Perú

Sánchez, E. & Alanis, A. (2006). *Redes neuronales: conceptos fundamentales y aplicaciones a control automático*. Madrid, España: Pearson Educación.

Sauñe, J. & Gamarra, M. (2017). *Desempeño docente y el rendimiento académico en el área de matemática en los aprendices del I ciclo del SENATI sede San Martín de Porres. Lima.* (tesis de maestría). Universidad César Vallejo.

Serrano, A., Soria, E. & Martín, J. (2009). Redes Neuronales Artificiales. Universitat de Valencia. España.

Stanghellini, E. (2009). *Introduzione ai metodi statistici per il credit scoring.* Perugia, Italia: Springer.

Sucari, R. (2018). *Comparación del análisis discriminante no métrico, árboles de clasificación Chaid y la regresión logística multinomial.* (tesis de maestría). Universidad Nacional Agraria La Molina. Lima, Perú

Tomassini, M. (2018). *Factores educativos y familiares que influyen en el desempeño académico de estudiantes de agronomía.* (tesis de maestría). Universidad Nacional Agraria La Molina. Lima, Perú.

Torrado, M. & Berlanga, V. (2013). Análisis discriminante mediante SPSS. REIRE *Revista d'Innovació i Recerca en Educació, 6 (2), 150-166*. Recuperado de https://www.academia.edu/11986195/An%C3%A1lisis_Discriminante_mediante_SPSS.

Uriel, E. & Aldás, J. (2005). *Análisis multivariante aplicado.* Madrid, España: Thomson.

Úsuga, O. & Patiño, C. (2008). Análisis discriminante no métrico y regresión logística en el problema de clasificación. *Revista Tecnológicas Nº 21*. Recuperado de https://www.researchgate.net/publication/279441761_Analisis_discriminante_no_metrico_y_regresion_logistica_en_el_problema_de_clasificacion/fulltext/55a777d108ae345d61db3b05/Analisis-discriminante-no-metrico-y-regresion-logistica-en-el-problema-de-clasificacion.pdf

Vargas, M. & Montero, E. (2016). Factores que determinan el rendimiento académico en Matemáticas en el contexto de una universidad tecnológica: aplicación de un modelo de ecuaciones estructurales. *Universitas Psychologica*. Vol. 15 No. 4. Colombia. Recuperado de https://www.researchgate.net/publication/315349649_Factores_que_determinan_el_rendimiento_academico_en_Matematicas_en_la_Universidad_Nacional_de_Ingenieria_UNI_Nicaragua_un_modelo_de_ecuaciones_estructurales/fulltext/5c0d68184585157ac1b6b32b/Factores-que-determinan-el-rendimiento-academico-en-Matematicas-en-la-Universidad-Nacional-de-Ingenieria-UNI-Nicaragua-un-modelo-de-ecuaciones-estructurales.pdf

Veliz, C. (2016). Análisis multivariante: métodos estadísticos para la investigación. Buenos Aires, Argentina: Cengage Learning.

Vergara G. & Peredo H. (2017). Relación del desempeño académico de estudiantes de primer año de universidad en Chile y los instrumentos de selección para su ingreso. *Revista Educación, 41(2).* Universidad de Costa Rica. Costa Rica. Recuperado de https://www.redalyc.org/jatsRepo/440/44051357006/html/index.html

Villanueva, I. (2017). *Comparación de árboles de clasificación CART y redes neuronales perceptrón multicapa para la clasificación de estudiantes de secundaria que consumen o no alcohol.* (tesina de pregrado). Universidad Nacional Mayor de San Marcos. Lima, Perú.

Vivanco, F. (2018). *Clasificación del rendimiento en la prueba de admisión de la UNALM utilizando análisis discriminante lineal y máquina de soporte vectorial*. (tesis de pregrado). Universidad Nacional Agraria La Molina. Lima, Perú.

Vivo, J., Sánchez, M., Franco, M. (2004). Estudio del rendimiento académico universitario basado en curvas ROC. *Revista de Investigación Educativa*, 22(2), 327-340. Recuperado de https://rehttps://revistas.um.es/rie/article/view/98581vistas.um.es/rie/article/view/98581.

Yacolca, M. (2001). *Rendimiento de los estudiantes ingresantes de la UNALM por el sistema de ingreso por canales.* (tesis de pregrado). Universidad Nacional Agraria La Molina. Lima, Perú.

Zambrano, C., Rojas, D., Carvajal, K., Acuna, G. (2011). Análisis de rendimiento académico estudiantil usando data warehouse y redes neuronales. *Revista Ingeniare,* 19(3) 369-381. Recuperado de http://dx.doi.org/10.4067/S0718-33052011000300007.

Apéndice A. Tablas de Contingencia

Tabla A1

Créditos Aprobados y Procedencia

Procedencia	Créditos Aprobados		
	50% o más	Menos del 50%	Total
Lima y Callao	381	59	440
	87%	13%	100%
	79%	81%	80%
Otras provincias	99	14	113
	88%	12%	100%
	21%	19%	20%
Total	480	73	553
	89%	11%	100%
	100%	100%	100%

Tabla A2

Cursos Desaprobados y Procedencia

Procedencia	Cursos Desaprobados		
	2 cursos o menos	Tres o más cursos	Total
Lima y Callao	361	79	440
	82%	18%	100%
	80%	80%	80%
Otras provincias	93	20	113
	82%	18%	100%
	20%	20%	20%
Total	454	99	553
	82%	18%	100%
	100%	100%	100%

Tabla A3

Resultado y Carreras Profesionales

Carreras Profesionales	Resultado		
	Aprobados	Desaprobados	Total
Biología, Ingeniería Ambiental e Industrias Alimentarias	93	45	138
	67%	33%	100%
	27%	21%	25%
Agronomía, Meteorología, Ingeniería Forestal, Economía, Estadística Informática, Ingeniería en Gestión Empresarial, Ingeniería Agrícola, Zootecnia y Pesquería	250	165	415
	60%	40%	100%
	73%	79%	75%
Total	343	210	553
	62%	38%	100%
	100%	100%	100%

Tabla A4

Créditos Aprobados y Carreras Profesionales

Carreras Profesionales	Créditos Aprobados		
	50% o más	Menos del 50%	Total
Biología, Ingeniería Ambiental e Industrias Alimentarias	120	18	138
	87%	13%	100%
	25%	25%	25%
Agronomía, Meteorología, Ingeniería Forestal, Economía, Estadística Informática, Ingeniería en Gestión Empresarial, Ingeniería Agrícola, Zootecnia y Pesquería	360	55	415
	87%	13%	100%
	75%	75%	75%
Total	480	73	553
	87%	13%	100%
	100%	100%	100%

Tabla A5

Cursos Desaprobados y Carreras Profesionales

Carreras Profesionales	Cursos Desaprobados		
	2 cursos o menos	Tres o más cursos	Total
Biología, Ingeniería Ambiental e Industrias Alimentarias	120	18	138
	87%	13%	100%
	26%	18%	25%
Agronomía, Meteorología, Ingeniería Forestal, Economía, Estadística Informática, Ingeniería en Gestión Empresarial, Ingeniería Agrícola, Zootecnia y Pesquería	334	81	415
	80%	20%	100%
	74%	82%	75%
Total	454	99	553
	82%	18%	100%
	100%	100%	100%

Apéndice B. Análisis Exploratorio Univariado

Tabla B1

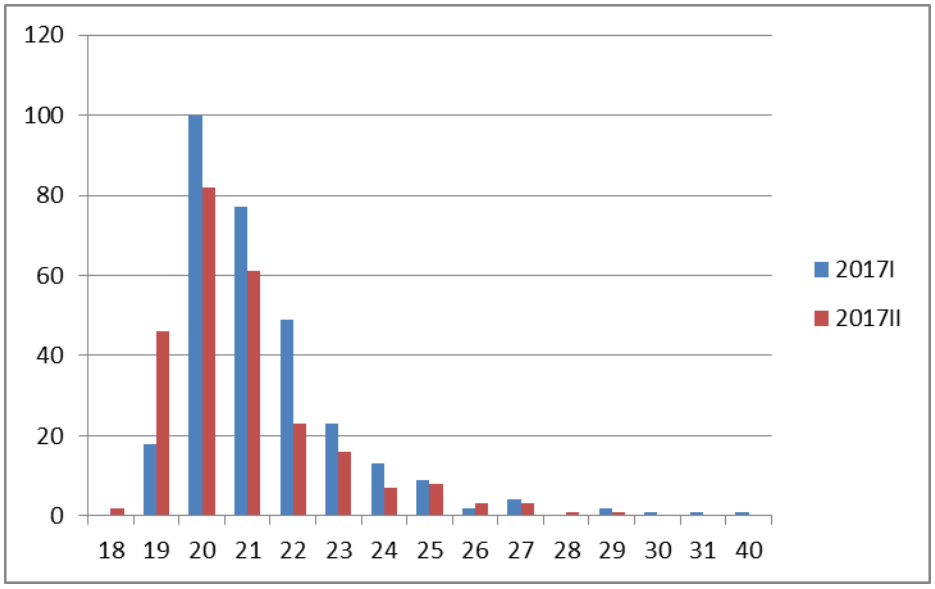
Estudiantes ingresantes matriculados en el curso de matemática según Edad.

Edad	2017 I	2017 II	Total
18	0	2	2
19	18	46	64
20	100	82	182
21	77	61	138
22	49	23	72
23	23	16	39
24	13	7	20
25	9	8	17
26	2	3	5
27	4	3	7
28	0	1	1
29	2	1	3
30	1	0	1
31	1	0	1
40	1	0	1
Total	300	253	553

En el año 2017, el 33% de los estudiantes matriculados en el curso de matemática, ingresó a la UNALM a la edad de 20 años. A nivel general, en el año 2017 I, el 81% de estudiantes tenían entre 18 y 22 años de edad cumplidos. En tanto que en el ciclo 2017 II, el 85% de estudiantes tenían entre 18 y 22 años cumplidos. Además, el promedio de edad de los estudiantes que ingresaron a la UNALM fue de 21 años en ambos ciclos académicos.

Figura B1

Estudiantes ingresantes matriculados en el curso de matemática según Edad.



La distribución de los estudiantes matriculados en el curso de matemática en ambos ciclos académicos, según la edad. Para el ciclo 2017 I, el 6% de estudiantes tenían 19 años en contraste con el ciclo 2017 II que fue superior al 18%. Así también, para el ciclo 2017 I más del 16% de estudiantes tenía 22 años en cambio esta cifra se redujo para el ciclo 2017 II en el que registró un 9%.

Tabla B2

Estudiantes ingresantes matriculados en el curso de matemática según Sexo.

Sexo	2017 I	2017 II	Total
Femenino	144	127	271
Masculino	156	126	282
Total	300	253	553

En el ciclo 2017 I, el 48% de los estudiantes ingresantes fueron mujeres, mientras que el 52% de estudiantes fueron varones. Para el ciclo 2017 II no se observaron mayores diferencias entre ambos grupos de estudiantes.

Tabla B3

Estudiantes ingresantes matriculados en el curso de matemática según Lugar de Procedencia.

Procedencia	2017 I	2017 II	Total
Lima y Callao	237	203	440
Otras provincias	63	50	113
Total	300	253	553

La mayoría de los estudiantes de la UNALM provenían de la provincia de Lima y Callao, siendo el 79% en el ciclo 2017 I y 80% en el ciclo 2017 II. A nivel general, en el año 2017, el 20% de estudiantes ingresantes a la UNALM provenían de otras provincias o regiones del Perú.

Tabla B4

Estudiantes ingresantes matriculados en el curso de matemática según Tipo de Colegio.

Tipo de Colegio	2017 I	2017 II	Total
Público	127	88	215
Privado	173	165	338
Total	300	253	553

En los ciclos 2017 I y 2017 II, el 58% y 65% de estudiantes ingresantes de la UNALM culminaron sus estudios de nivel secundario, en un colegio de gestión privada, respectivamente. A nivel general, en el año 2017, se observa que el 39% de estudiantes ingresantes y matriculados en la UNALM culminaron los estudios secundarios en un colegio de gestión pública.

Tabla B5

Estudiantes ingresantes matriculados en el curso de matemática según Ubicación del Colegio.

Ubicación del Colegio	2017 I	2017 II	Total
Lima y Callao	275	234	509
Otras provincias	25	19	44
Total	300	253	553

El 92% de estudiantes ingresantes en el ciclo 2017 I, culminaron sus estudios secundarios en Lima y Callao, en tanto que en el ciclo 2017 II fue el 93%. A nivel general, el 8% de estudiantes ingresantes en el año 2017 culminaron los estudios secundarios en otras provincias o departamentos del Perú.

Tabla B6

Estudiantes ingresantes matriculados en el curso de matemática según Carrera Profesional.

Carrera profesional	2017 I	2017 II	Total
Biología, Ingeniería Ambiental e Industrias Alimentarias	67	71	138
Agronomía, Meteorología, Ingeniería Forestal, Economía, Estadística Informática, Ingeniería en Gestión Empresarial, Ingeniería Agrícola, Zootecnia y Pesquería	233	182	415
Total	300	253	553

El 22% y 28% del total de estudiantes ingresantes a la UNALM pertenecieron a las carreras profesionales de Biología, Ingeniería Ambiental e Industrias Alimentarias, en los ciclos 2017 I y 2017 II, respectivamente. Las carreras profesionales se agruparon de acuerdo a los desempeños observados en los ciclos anteriores.

Tabla B7

Estudiantes ingresantes matriculados en el curso de matemática según Modalidad de Ingreso.

Modalidad de ingreso	2017 I	2017 II	Total
Concurso ordinario	288	244	532
Otras modalidades	12	9	21
Total	300	253	553

El 96% de estudiantes de ambos ciclos académicos ingresó a la universidad bajo la modalidad de Concurso ordinario. Se debe precisar que dentro de la categoría Otras modalidades se han considerado aquellas que obligatoriamente rinden la prueba de admisión que consta de 100 preguntas y tiene una duración de 3 horas.

Tabla B8

Estudiantes ingresantes matriculados en el curso de matemática según Opción de Ingreso.

Opción de ingreso	2017 I	2017 II	Total
Primera	137	140	277
Segunda	113	96	209
Tercera	50	17	67
Total	300	253	553

En el ciclo 2017 I, el 46% de estudiantes ingresaron a la carrera profesional elegida como primera opción. El 38% de los estudiantes ingresó a la carrera profesional elegida como segunda opción y el 17% ingresó a la tercera opción.

Tabla B9

Puntajes del Examen de Admisión. Ciclos 2017 I y 2017 II

Áreas evaluadas	2017 I	2017 II	Promedio Ponderado
Matemática	14,3	11,0	12,8
Razonamiento Matemático	14,0	12,1	13,1
Razonamiento Verbal	10,0	11,8	10,8
Física	14,9	12,9	14,0
Química	13,2	14,0	13,5
Biología	11,4	09,1	10,4
Puntaje General	12,9	11,8	12,4

En el examen de admisión 2017 I, el promedio más alto fue de 14,9 en el área de Física, seguido de 14,3 en Matemática y 14,0 en Razonamiento Matemático. Los promedios más bajos fueron de 10,0 en Razonamiento verbal y 11,4 en Biología. Para el examen de admisión 2017 II, el promedio más alto fue de 14,0 en el área de Química, seguido de 12,9 en Física y 12,1 en Razonamiento Matemático. En tanto, los promedios más bajos fueron de 09,1 en Biología y 11,0 en Matemática. Considerando ambos exámenes de admisión, el promedio ponderado más alto fue de 14,0 en el área de Física, seguido de 13,5 en Química y 13,1 en Razonamiento Matemática. Los promedios ponderados más bajos fueron de 10,4 en Biología y 10,8 en Razonamiento verbal. Los estudiantes de primer ciclo ingresaron a la universidad con un puntaje general promedio de 12,9 y 11,8 en los ciclos 2017 I y 2017 II, respectivamente.

Tabla B10

Estudiantes ingresantes matriculados en el curso de matemática según Orden de Mérito. Ciclos 2017 I y 2017 II

Orden de mérito	2017 I	2017 II	Total
Mitad superior	218	212	430
Mitad inferior	82	41	123
Total	300	253	553

El 73% y 84% de los estudiantes obtuvo un orden de mérito de ingreso correspondiente a la mitad superior de todos los ingresantes matriculados en los ciclos académicos 2017 I y 2017 II, respectivamente.

Tabla B11

Estudiantes ingresantes matriculados en el curso de matemática según Resultado.

Resultado	2017 I		2017 II		Total	
	n	%	n	%	n	%
Aprobado	205	68,3	138	54,5	343	62,0
Desaprobado	95	31,7	115	45,5	210	38,0
Total	300	100,0	253	100,0	553	100,0

El 68,3% y 54,5% de los estudiantes matriculados aprobaron el curso de matemática en los ciclos 2017 I y 2017 II, respectivamente. En general, en el año 2017, el 62,0% de estudiantes aprobó el curso de matemática.

Tabla B12

Estudiantes ingresantes matriculados en el curso de matemática según los Créditos Aprobados.

Créditos aprobados	2017 I		2017 II		Total	
	n	%	n	%	n	%
12 o más	267	89,0	213	84,2	480	86,8
Menos de 12	33	11,0	40	15,8	73	13,2
Total	300	100,0	253	100,0	553	100,0

El 89% y 84,2% de los estudiantes de primer ciclo en la UNALM aprobaron al menos 12 créditos del total de créditos matriculados en los ciclos 2017 I y 2017 II, respectivamente. En general, el 86,8% de estudiantes aprobaron al menos 12 de los créditos del total de créditos matriculados en el año 2017.

Apéndice C. Análisis Exploratorio Bivariado

Tabla C1

Resultado y Sexo de estudiantes ingresantes matriculados en el curso de matemática

Ciclo	Sexo	Resultado		
		Aprobados	Desaprobados	Total
2017 I	Hombre	45	99	144
		31%	69%	100%
		47%	48%	48%
	Mujer	50	106	156
		32%	68%	100%
		53%	52%	52%
	Total	95	205	300
		32%	68%	100%
		100%	100%	100%
2017 II	Hombre	62	65	127
		49%	51%	100%
		54%	47%	50%
	Mujer	53	73	126
		42%	58%	100%
		46%	53%	50%
	Total	115	138	253
		45%	55%	100%
		100%	100%	100%

Del total de estudiantes de primer ciclo en la UNALM, las mujeres aprobaron en mayor porcentaje (52%). En cambio, del total de estudiantes que desaprobaron el primer curso de matemática, el 51% y 49% correspondió a los hombres y mujeres, respectivamente.

Tabla C2

Créditos Aprobados y Sexo de estudiantes ingresantes matriculados en el curso de matemática

Ciclo	Sexo	Créditos Aprobados		
		12 o más	Menos de 12	Total
2017 I	Hombres	26	118	144
		18%	82%	100%
		45%	49%	48%
	Mujeres	32	124	156
		21%	79%	100%
		55%	51%	52%
	Total	58	242	300
		19%	81%	100%
		100%	100%	100%
2017 II	Hombres	50	77	127
		39%	61%	100%
		52%	49%	50%
	Mujeres	47	79	126
		37%	63%	100%
		48%	51%	50%
	Total	97	156	253
		38%	62%	100%
		100%	100%	100%

En el ciclo 2017 I las estudiantes mujeres aprobaron 12 o más créditos en mayor proporción (55%) respecto a los varones. Un escenario muy parecido ocurrió con los estudiantes que aprobaron menos de 12 créditos. En el ciclo 2017 II, los estudiantes varones aprobaron 12 o más créditos en mayor proporción que las mujeres y lo contrario ocurrió con los estudiantes que aprobaron menos de 12 créditos.

Tabla C3

Resultado y Edad de estudiantes ingresantes matriculados en el curso de matemática

Ciclo	Edad	Resultado		
		Aprobados	Desaprobados	Total
2017 I	Menor a 20 años	14	4	18
		78%	22%	100%
		7%	4%	6%
	Mayores o igual a 20 años	191	91	282
		68%	32%	100%
		93%	96%	94%
	Total	205	95	300
		68%	32%	100%
		100%	100%	100%
2017 II	Menor a 20 años	38	10	48
		79%	21%	100%
		28%	9%	19%
	Mayores o igual a 20 años	100	105	205
		49%	51%	100%
		72%	91%	81%
	Total	138	115	253
		55%	45%	100%
		100%	100%	100%

Los estudiantes ingresantes en ambos ciclos académicos y con edades menores a 20 años, tuvieron proporciones similares de aprobación del curso de matemática, siendo de 78% y 79%, respectivamente. Del total de estudiantes que desaprobaron el curso, más del 90% de estudiantes correspondieron a los estudiantes con 20 o más años, en ambos ciclos académicos.

.

Tabla C4

Créditos Aprobados y Edad de estudiantes ingresantes matriculados en el curso de matemática

Ciclo	Edad	Créditos Aprobados		
		12 o más	Menos de 12	Total
2017 I	Menor a 20 años	15	3	18
		83%	17%	100%
		6%	5%	45%
	Mayores o igual a 20 años	227	55	282
		80%	20%	100%
		94%	95%	55%
	Total	242	58	300
		81%	19%	100%
		100%	100%	100%
2017 II	Menor a 20 años	40	8	48
		83%	17%	100%
		26%	8%	19%
	Mayores o igual a 20 años	116	89	205
		57%	43%	100%
		74%	92%	81%
	Total	156	97	253
		62%	38%	100%
		100%	100%	100%

Existe una marcada diferencia en el desempeño académico de los estudiantes al finalizar su primer ciclo de estudios. Así tenemos que, el 83% de los estudiantes menores de 20 años aprobaron al menos 12 créditos en ambos ciclos. Con respecto a los estudiantes con 20 o más años, se observaron diferencias en la aprobación de al menos 12 créditos, en el ciclo 2017 I aprobaron el 80% de estudiantes, mientras que en el ciclo 2017 II solo aprobó el 57%. Con respecto a los estudiantes que aprobaron menos de 12 créditos, se evidenció que el 95% y 92% corresponden a estudiantes con 20 o más años para los ciclos 2017 I y 2017 II, respectivamente.

Tabla C5

Resultado y Procedencia de estudiantes ingresantes matriculados en el curso de matemática

Ciclo	Procedencia	Resultado		
		Aprobados	Desaprobados	Total
2017 I	Lima y Callao	162	75	237
		68%	32%	100%
		79%	79%	79%
	Otras provincias	43	20	63
		68%	32%	100%
		21%	21%	21%
	Total	205	95	300
		68%	32%	100%
		100%	100%	100%
2017 II	Lima y Callao	111	92	203
		55%	45%	100%
		80%	80%	80%
	Otras provincias	27	23	50
		54%	46%	100%
		20%	20%	20%
	Total	138	115	253
		55%	45%	100%
		100%	100%	100%

En ambos ciclos académicos, alrededor del 80% de los estudiantes que aprueban el primer curso de matemática en la UNALM, provienen de Lima y Callao. La misma proporción se evidencia para el grupo de estudiantes que desaprobaron el curso. Con respecto al desempeño de los estudiantes que provienen de otras provincias o regiones, se observaron algunas diferencias en ambos ciclos académicos, es así que el 68% y 54% de ellos, aprobaron el curso en los ciclos 2017 I y 2017 II respectivamente.

Tabla C6

Estadísticas de variables según Resultado de estudiantes ingresantes matriculados en el curso de matemática. Ciclo 2017 I

Variables	Resultado			
	Aprobado		Desaprobado	
	Promedio	Desv. Est.	Promedio	Desv. Est.
Edad	21,2	2,2	21,9	2,0
Matemática	14,9	3,1	13,1	3,1
Razonamiento Matemático	14,3	2,3	13,3	2,4
Razonamiento Verbal	09,8	2,3	10,4	2,2
Física	15,6	2,9	13,3	3,2
Química	13,5	3,0	12,4	2,9
Biología	11,5	3,6	11,3	3,7
Puntaje general	13,1	1,3	12,2	0,9

En el ciclo 2017 I no existen mayores diferencias entre las edades de los estudiantes que aprobaron y desaprobaron el primer curso de matemática siendo en promedio alrededor de los 21 años. Así también, para los grupos de aprobados y desaprobados en el primer curso de matemática, existen diferencias mayores a 0,5 en el promedio de todas las áreas evaluadas en el examen de admisión con excepción de Biología que además es el área que presenta la mayor variabilidad en las notas, lo mismo sucede con Matemática y Química. Los estudiantes que aprobaron y desaprobaron el primer curso de Matemática, tuvieron el promedio más bajo y desaprobatorio en el área de Razonamiento Verbal en el concurso de admisión 2017 I. En ambos grupos de estudiantes, los promedios más altos obtenidos en el examen de admisión fueron Física, Matemática y Razonamiento Matemático.

Tabla C7

Estadísticas de variables según Resultado de estudiantes ingresantes matriculados en el curso de matemática. Ciclo 2017 II

Variables	Resultado			
	Aprobado		Desaprobado	
	Promedio	Desv. Est.	Promedio	Desv. Est.
Edad	20,7	2,0	21,2	1,6
Matemática	11,9	3,7	09,8	3,6
Razonamiento Matemático	12,5	2,3	11,7	2,7
Razonamiento Verbal	11,5	2,0	12,2	2,2
Física	13,6	2,9	12,1	2,8
Química	14,2	2,7	13,8	2,8
Biología	08,8	3,6	09,5	3,3
Puntaje general	12,1	1,2	11,5	0,8

En el ciclo 2017 II l la edad promedio de los estudiantes aprobados y desaprobados es cercano a los 21 años. Con respecto a los estudiantes aprobados en matemática, en promedio obtuvieron la nota más baja y desaprobatoria en Biología. Con relación al grupo de estudiantes que desaprobaron el primer curso de matemática, en promedio obtuvieron las notas más bajas y desaprobatorias en Biología y Matemática. Al comparar ambos grupos se observa que existen diferencias mayores a 0,5 en el promedio de todas las áreas evaluadas en el examen de admisión con excepción de Química. Entre los estudiantes que aprobaron Matemática destacan sus altos promedios en las áreas de Química, Física y RM del examen de admisión; mientras que en el grupo de estudiantes desaprobados destacan sus promedios más altos en las áreas de Química, RV y Física. Las notas en las áreas de Matemática y Biología son las que presentaron mayor variabilidad en ambos grupos de estudiantes.

Tabla C8

Estadísticas de variables según Créditos Aprobados de estudiantes ingresantes matriculados en el curso de matemática. Ciclo 2017 I

Variables	Créditos Aprobados			
	12 o más		Menos de 12	
	Promedio	Desv. Est.	Promedio	Desv. Est.
Edad	21,2	2,1	22,2	2,1
Matemática	14,6	3,3	13,3	2,9
Razonamiento Matemático	14,1	2,3	13,5	2,5
Razonamiento Verbal	09,8	2,3	10,5	2,2
Física	15,2	3,1	13,6	3,0
Química	13,5	3,0	11,9	3,0
Biología	11,6	3,5	10,9	3,9
Puntaje general	13,0	1,3	12,3	0,9

De la Tabla 30 se observa que en el ciclo 2017 I la edad promedio de los estudiantes que aprobaron 12 o más créditos y los que aprobaron menos de 12 créditos fue de 21 y 22 años respectivamente. Así también, se observa que, para ambos grupos de estudiantes, existen diferencias mayores a 0,6 en el promedio de todas las áreas evaluadas en el examen de admisión. Las notas en el área de Biología son las presentan mayor variabilidad en ambos grupos de estudiantes. De igual forma, se observa que los estudiantes que aprobaron 12 o más créditos y los que aprobaron menos de 12 créditos, tuvieron el promedio más bajo y desaprobatorio en el área de Razonamiento Verbal en el concurso de admisión 2017 I. Finalmente, se evidencia que en ambos grupos de estudiantes, los promedios más altos obtenidos en el examen de admisión fueron en las áreas de Física, Matemática y Razonamiento Matemático, estos resultados coinciden con los obtenidos en el grupo de estudiantes que aprobaron o no el curso de Matemática en el primer ciclo de estudios universitarios.

Tabla C9

Estadísticas de variables según Créditos Aprobados de estudiantes ingresantes matriculados en el curso de matemática. Ciclo 2017 II

Variables	Créditos Aprobados			
	12 o más		Menos de 12	
	Promedio	Desv. Est.	Promedio	Desv. Est.
Edad	20,7	1,9	21,3	1,6
Matemática	11,5	3,8	10,1	3,6
Razonamiento Matemático	12,4	2,4	11,7	2,6
Razonamiento Verbal	11,7	2,0	12,0	2,3
Física	13,3	3,0	12,2	2,8
Química	14,3	2,6	13,5	2,9
Biología	09,0	3,5	09,3	3,3
Puntaje general	12,0	1,2	11,5	0,7

En el ciclo 2017 II la edad promedio de los estudiantes que aprobaron 12 o más créditos y los que aprobaron menos de 12 créditos fue de 21 años. Para ambos grupos de estudiantes, existen diferencias mayores a 0,6 en el promedio de todas las áreas evaluadas en el examen de admisión con excepción de Biología y RV. Las notas en las áreas de Biología y Matemática presentan mayor variabilidad en ambos grupos de estudiantes. Los estudiantes que aprobaron 12 o más créditos y los que aprobaron menos de 12 créditos, tuvieron el promedio más bajo y desaprobatorio en el área de Biología. En el grupo de estudiantes con 12 o más créditos aprobados, los promedios más altos obtenidos en el examen de admisión fueron en Física, Matemática y RM. Mientras que para el grupo de estudiantes con menos de 12 créditos aprobados en el primer ciclo son Química, Física y RV.

Apéndice D. Matriz de Consistencia

<table>
<tr><th colspan="6">“USO DE LOS MÉTODOS MULTIVARIANTES PARA EL ANÁLISIS DEL DESEMPEÑO DE LOS ESTUDIANTES DE LA EDUCACIÓN SUPERIOR (CASO: ESTUDIANTES INGRESANTES EN EL PRIMER CURSO DE MATEMÁTICA DE LA UNALM)”</th></tr>
<tr><th>Problema General</th><th>Objetivo General</th><th>Hipótesis General</th><th>Variables</th><th>Metodología</th><th>Población</th></tr>
<tr><td>¿Cuál de los métodos multivariantes analiza mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM?</td><td>Comparar los métodos multivariantes que analizan mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM</td><td>El método multivariante que analiza mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM es la regresión logística binaria.</td><td rowspan="4"><u>Variables Independientes</u>
Los métodos multivariantes: análisis discriminante, regresión logística binaria y de redes neuronales tipo perceptrón multicapa

<u>Variable Dependiente</u>
El porcentaje de clasificación correcta, especificidad, sensibilidad y el área bajo la curva ROC

<u>Variable Interviniente</u>
El desempeño académico</td><td rowspan="4"><u>Tipo</u>
Descriptivo y correlacional

<u>Diseño</u>
No experimental

<u>Técnicas de Recolección</u>
Acopio de información a partir de los datos de admisión y de las actas de evaluación final del curso de matemática</td><td rowspan="4"><u>Población</u>
Estudiantes ingresantes y matriculados en el primer curso de matemática de la UNALM en los ciclos académicos 2017 I y 2017 II

<u>Unidad de análisis</u>
Estudiante ingresante y matriculado en el primer curso de matemática en la UNALM</td></tr>
<tr><th>Problemas Específicos</th><th>Objetivos Específicos</th><th>Hipótesis Específicas</th></tr>
<tr><td>¿La regresión logística binaria analiza mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM?</td><td>Analizar el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM mediante la regresión logística binaria</td><td></td></tr>
<tr><td>¿El análisis discriminante analiza mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM?</td><td>Analizar el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM mediante el análisis discriminante</td><td>El método multivariante de regresión logística binaria es mejor que el análisis discriminante para analizar el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM.</td></tr>
<tr><td>¿Las redes neuronales tipo perceptrón multicapa analizan mejor el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM?</td><td>Analizar el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM mediante las redes neuronales tipo perceptrón multicapa</td><td>El método multivariante de regresión logística binaria es mejor que el de redes neuronales tipo perceptrón multicapa para analizar el desempeño académico de los estudiantes ingresantes en el primer ciclo de la UNALM.</td></tr>
</table>

Apéndice E. Glosario de términos

Autovalor. El autovalor de una función representa la proporción de variabilidad total de las puntuaciones sobre el conjunto de todas las funciones atribuibles a la función. Cuanto más alto sea su valor, la función discriminará mucho. (De la Fuente, 2011)

Correlaciones canónicas. Miden las desviaciones de las puntuaciones discriminantes entre grupos respecto a las desviaciones totales sin distinguir grupos. Si su valor se acerca a la unidad la dispersión se deberá a las diferencias entre grupos, con lo cual, la función discriminará mucho entre los grupos. (De la Fuente, 2011)

$$\eta = \sqrt{\frac{\lambda}{1+\lambda}}$$

Tolerancia. Miden la multicolinealidad de una variable, se basa en la correlación múltiple de una variable respecto del resto de variables. Su valor es igual a $1 - R_i^2$, donde R_i^2 es la correlación múltiple al cuadrado de la variable respecto a las demás variables. Cuando más baja es la tolerancia, la correlación será más alta y por tanto habrá multicolinealidad.

VIF. Es el Factor de Influencia de la Varianza. Sirve para detectar la multicolinealidad de variables. Es el recíproco de la Tolerancia. Su valor es igual a:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Cuanto más cercano a 1 sea el valor de VIF entonces no existirá multicolinealidad, en cambio, cuanto más alto sea su valor, significará la existencia de multicolinealidad.