Métodos de clasificación aplicados al perfil del estudiante por categorías de rendimiento global en la prueba Ser Bachiller 2019

Chávez Vinueza, Delfín Leonardo

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Maestría en Enseñanza de la Matemática

Trabajo de titulación, previo a la obtención del título de Magister en Enseñanza de la Matemática

Ing. Armijos Toro, Livino Manuel Mgs.

30 de septiembre del 2020

Curiginal

Document Information

Analyzed document	Tesis versión 19 feb 2021 - Leonardo Chavez.pdf (D110525889)
Submitted	7/16/2021 7:35:00 PM
Submitted by	ARMIJOS TORO LIVINO MANUEL
Submitter email	lmarmijos2@espe.edu.ec
Similarity	1%
Analysis address	lmarmijos2.espe@analysis.urkund.com

Sources included in the report

W	URL: https://www.researchgate.net/profile/Carlos_Bouza/publication/313364703_EL_PROBLEMA_DE_LA_POBREZA_Y_VIOLENCIA_FAMILIAR_UNA_MIRADA_PERSPECTIVA_DESDE_LA_INVESTIGACION_DE_OPERACIONES/links/5897eac292851c8bb67f09d4/EL-PROBLEMA-DE-LA-POBREZA-Y-VIOLENCIA-FAMILIAR-UNA-MIRADA-PERSPECTIVA-DESDE-LA-INVESTIGACION-DE-OPERACIONES.pdf Fetched: 12/13/2019 8:14:01 PM	1
SA	**INFORME FINAL FEDU 2019- LUIS ALCÁNTARA ZÁRATE.pdf** Document INFORME FINAL FEDU 2019- LUIS ALCÁNTARA ZÁRATE.pdf (D104337057)	1
SA	**olivera_c.pdf** Document olivera_c.pdf (D29718725)	1
W	URL: https://www.docsity.com/es/analisis-discriminante-3/5905046/ Fetched: 6/5/2021 3:00:53 PM	2
W	URL: https://core.ac.uk/download/pdf/61918191.pdf Fetched: 5/26/2020 5:54:25 AM	2
W	URL: https://www.fuenterrebollo.com/Master-Econometria/Analisis_Discriminante.pdf Fetched: 12/1/2020 11:15:38 PM	3

Firma:

Firmado electrónicamente por:
LIVINO MANUEL
ARMIJOS TORO

Ing. Armijos Toro, Livino Manuel Mgs.

Director

CC: 0916543747

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Certificación

Certifico que el trabajo de titulación: **Métodos de clasificación aplicados al perfil del estudiante por categorías de rendimiento global en la prueba Ser Bachiller 2019** fue realizado por el señor **Chávez Vinueza, Delfín Leonardo**; el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además ha sido revisado y analizado en su totalidad por la herramienta de verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Sangolquí, 20 de julio de 2021

Firma:

LIVINO MANUEL ARMIJOS TORO

Ing. Armijos Toro, Livino Manuel Mgs.

Director

C.C.: 0916543747

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Responsabilidad de Autoría

Yo, **Chávez Vinueza, Delfín Leonardo** con cédula de ciudadanía 1708175706, declaro que el contenido, ideas y criterios del trabajo de titulación: **Métodos de clasificación aplicados al perfil del estudiante por categorías de rendimiento global en la prueba Ser Bachiller 2019** es de mi autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 30 de septiembre de 2020

Firma:

Chávez Vinueza, Delfín Leonardo

CC: 1708175706

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrados

Autorización de Publicación

Yo, **Chávez Vinueza, Delfín Leonardo**, con cédula de ciudadanía 1708175706 autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **Métodos de clasificación aplicados al perfil del estudiante por categorías de rendimiento global en la prueba Ser Bachiller 2019** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 30 de septiembre de 2020

Firma:

Chávez Vinueza, Delfín Leonardo

CC: 1708175706

Dedicatoria

A la juventud que busca en los datos del presente, un espejo que refleje el futuro.

Leonardo

Agradecimiento

Al INEVAL, cuya información permite un análisis global del último año de bachillerato con todas sus consecuentes ventajas a las futuras generaciones.

Índice

Índice de Tablas

Índice de Figuras

RESUMEN

El proceso educativo está condicionado por varios factores, sin duda, los más preponderantes están en el ámbito social, económico y familiar del entorno estudiantil; la preocupación de los estudiantes y sus familiares, una vez que terminan la educación media, es el acceso a la educación superior. El presente estudio de investigación, empleó técnicas de Big Data. El mismo que permitió que, a partir del examen de ingreso Ser Bachiller 2019 y de la encuesta de Factores Asociados 2019, del INEVAL, pronosticar el acceso a la educación superior de bachilleres en todo el país. De las datas proporcionadas se construyó una sola matriz, sobre ésta se depuró la información; eliminando así variables similares y otras que no presentan información en por lo menos el 50 % de observaciones.

Así mismo, se eliminaron observaciones que no presentaron resultados de las evaluaciones parciales. Posteriormente, se redujo la matriz, seleccionando las variables utilizadas en estudios similares realizados en Colombia, Argentina y Costa Rica. Además, se revisaron las variables que emplean los proyectos Pisa y Terce, y así tomar los indicadores más recurrentes como las variables significativas para este estudio. Se logró una matriz de 265 915 observaciones por 48 variables; de estas 48 variables, 45 correrspondieron a factores asociados. Se aplicó el Análisis de Componentes Principales para representar los 45 factores asociados en un conjunto condensado de menor dimensión; sobre esta base se trabajaron los modelos predictivos que determinan el ingreso, o no, a la educación superior de un determinado individuo. Utilizando las técnicas de Regresión Logística, Análisis Discriminante y Máquina de Soporte Vectorial. Se estableció que de estos modelos, el más adecuado es el de Regresión Logística el mismo que produce sus pronósticos con un 70 % de eficiencia. Adicionalmente, se determinaron los factores asociados que más influyen en forma positiva y en forma negativa, para que el estudiante ingrese a la universidad.

Palabras clave:

- **BIG DATA**
- **ANÁLISIS DE COMPONENTES PRINCIPALES**
- **REGRESIÓN LOGISTICA**
- **ANÁLISIS DISCRIMINANTE**
- **MÁQUINA DE SOPORTE VECTORIAL**

ABSTRACT

To access to higher education is a concern for students who are finish high school. The educational process is conditioned by a great number of factors, the most preponderant is the social, economic, and family environment in which the student develops. This research study uses Big Data techniques that allow, to forecast the access to higher education of high school graduates throughout the country, with information based on: Ser Bachiller 2019 exams results, and the survey of Associated Factors 2019 INEVAL. With the data provided, single matrix was constructed. The matrix was depurated through the elimination of similar variables: those who does not have information in at least 50 % of the observations, observations without partial evaluations results were eliminated. The matrix experienced a new depuration whe we match used in similar studies carried out by countries in the region such as Colombia, Argentina, and Costa Rica. In addition, the variables used by Pisa and Terce are reviewed so that the most recurrent indicators are taken as the significant variables for this study, resulting in a matrix of 265 915 observations for 48 variables; of these 48 variables, 45 correspond to the Associated Factors. Applying Principal Component Analysis, the 45 Associated Factors are represented in a condensed set of smaller dimensions. This is the base on which the predictive model that determined the entrance or not of a certain individual to higher education is worked, using the techniques of Logistic Regression, Discriminant Analysis and Support Vector Machine, establishing that Logistic Regression produce a forecast with 70 % efficiency, after we chose this model to work, the study also determined which associated factors are the most influential in a positive or negative way for the University access of the student.

Key words:

- **BIG DATA**
- **PRINCIPAL COMPONENT ANALYSIS**
- **LOGISTIC REGRESSION**
- **DISCRIMINANT ANALYSIS**
- **SUPPORT VECTOR MACHINE**

Capítulo 1

Introducción

El Instituto Nacional de Evaluación Educativa (INEVAL) es una institución pública y autónoma que aparece de la promulgación de la Constitución de Montecristi en su artículo 346, con el fin de promover la educación de excelencia. Tiene como función establecer los indicadores de calidad de la educación a través de la evaluación continua del aprendizaje, del desempeño de los profesionales de la educación y de la gestión de los establecimientos educativos. El INEVAL lleva a cabo el respectivo proceso de elaboración, recepción , calificación y publicación de los resultados a nivel nacional de las evaluaciones a los diversos niveles educativos [12, INEVAL, 2020].

Desde la creación del INEVAL en noviembre de 2012, se realizó en julio de 2014 por primera vez la prueba Ser Bachiller que reemplaza a los exámenes de grado y equivale al 10 % de la nota final para obtener el título de bachiller. Esta evaluación es para medir el desempeño de los estudiantes en relación a los estándares establecidos por el Ministerio de Educación, para con este diagnóstico tomar las medidas necesarias tendientes a mejorar la calidad de la educación [11, INEVAL, 2020].

El proceso Ser Bachiller está compuesto por cuatro pruebas que miden el grado de dominio en Matemáticas, Lengua y Literatura, Ciencias Naturales y Ciencias Sociales. Para identificar los factores que contribuyen en una educación de calidad y considerando los diversos contextos en los que se desenvuelven los estudiantes, se aplicó una encuesta denominada de Factores Asociados a los futuros bachilleres evaluados [11, INEVAL, 2020].

Esta prueba se aplicó de manera censal, es decir que se evaluó a todos los estudiantes de Tercero de Bachillerato de Instuticiones Educativas fiscales, fiscomisionales, municipales, particulares laicos y particulares religiosos de todo el país.

En el proceso de evaluación Ser Bachiller 2019 el INEVAL, facilita a la Universidad de las Fuerzas Armadas ESPE, la información para su estudio en los posgrados que dicta esta universidad. La información Ser Bachiller fue entregada en un formato de hoja de cálculo, y se la puede obtener en la dirección electrónica: http://www.evaluacion.gob.ec/evaluaciones/ descarga-de-datos/

Esta información en muchos casos queda sin utilizar y esto suele ser por la dificultad de su proce-

samiento al emplear herramientas y métodos convencionales, esta dificultad repercute en la inadecuada toma de decisiones con el consecuente bajo rendimiento escolar.

La Minería de Datos en particular se refiere al procesamiento, análisis y visualización de grandes y complejos volúmenes de datos, para la toma de decisiones, por lo tanto, abre una nueva era para predecir, prevenir y personalizar dificultades de aprendizaje en todos los campos de la educación, pero en particular el seguimiento de procesos académicos para su permanente mejora [14, Jimenez y Alvarez, 2010].

Existe gran preocupación por parte de padres de familia y jóvenes por la aprobación o no de la prueba Ser Bachiller en el país, ya que esta prueba define la posibilidad de acceder a los estudios de tercer nivel.

El problema a estudiarse es el acceso a la universidad de los estudiantes que aprueban el Tercer año de Bachillerato, el mismo depende en gran parte de los Factores Asociados. Para este estudio se usarán los datos de la prueba Ser Bachiller aplicada por el INEVAL en el año 2019, para definir la relación entre estos factores y la nota que obtuvo el estudiante.

El presente trabajo se realiza con los resultados de la prueba Ser Bachiller sustentada por 514852 estudiantes y la encuesta de Factores Asociados con 327 variables tomada por el INEVAL en el año 2019. El estudio determinará el impacto de los Factores Asociados en la accesibilidad a la universidad pública de los estudiantes de Tercero de Bachillerato.

El objetivo general es aplicar técnicas de Minería de Datos para establecer un modelo matemático que a partir del perfil del estudiante dado por los Factores Asociados sea capaz de predecir su accesibilidad a la universidad pública.

Los objetivos específicos serán:

- Depurar la información proporcionada por el INEVAL a datos significativos.
- Formar grupos de variables.
- Plantear un sistema que permita jerarquizar las variables considerando las más concurrentes utilizadas en estudios similares realizados por otros países de la región.
- Establecer un modelo matemático que relacione el perfil socio económico y familiar del estudiante expresado en la encuesta de Factores Asociados y su acceso a la universidad.

Este trabajo se ha dividido en las siguientes partes: en el Capítulo 1, se presenta el alcance del trabajo, se establecen las interrogantes que conducen esta investigación y los enfoques considerados para responderla; en el Capítulo 2, se habla sobre el Big Data y las técnicas de análisis multivariante empleadas para cumplir los objetivos; el Capítulo 3, describe la metodología utilizada en el tratamiento de la data; en el Capítulo 4, se hace un análisis y discusión de los resultados obtenidos de acuerdo a los objetivos y preguntas de investigación planteadas y; finalmente, en el Capítulo 5 se presentan las conclusiones y recomendaciones.

Capítulo 2

Estado del arte

En este Capítulo se analiza los siguientes temas: Situación de la educación en el Ecuador, el Big Data y la información, Análisis Multivariante y culmina con una revisión de trabajos similares. En el análisis de la educación en el Ecuador se repasa sobre los informes de diversos procesos evaluativos en diferentes niveles educativos, En el tema del Big Data y la información se analiza su importancia en el campo de la educación, finalmente se menciona los diferentes programas computacionales utilizados en este estudio y que permiten procesar y obtener información a través del Big Data. En el análisis multivariante se recopila información de los métodos que se aplican en el presente estudio, estos son, Análisis de Componentes Principales, Regresión Logística, Análisis Discriminante y Máquina de Vectores de Soporte. Finalmente se realiza una mirada a diversos trabajos sobre el tema realizados con anterioridad.

Situación de la educación en el Ecuador

La educación en sus diversos niveles es un derecho humano fundamental y un bien público porque gracias a ella nos desarrollamos como personas, como ciudadanos responsables de nuestro país y contribuimos al desarrollo de la sociedad, por tanto la educación tiene un valor en sí misma y no sólo como herramienta para el crecimiento económico.

El informe de OREALC de la UNESCO del 2007 manifiesta que las políticas educativas que asumen una mirada más comprensiva sobre la calidad de la educación, incorporando en ella las dimensiones de equidad, relevancia, pertinencia, eficacia y eficiencia, y que se diseñan a partir del enfoque de derechos, pueden ayudar a responder a los desafíos que en el campo de la educación enfrentan los países de la región. Son los Estados quienes tienen la responsabilidad de asegurar que todos sus habitantes puedan hacer exigible su derecho a una educación de calidad. Es con más y mejor educación que las personas pueden ampliar el ejercicio de su libertad; y los países aumentar su productividad, abatir la pobreza, combatir las desigualdades y consolidar comunidades más cohesionadas, transparentes y democráticas [9, Gautier, 2007].

Para entender la accesibilidad de los bachilleres a las Instituciones de Educación Superior (IES), es necesario hacer una retrospectiva de como se ha dado la educación en los años pasados, tanto a lo referente a la Educación Básica como al Bachillerato, es así que:

El Ministerio de Educación del Ecuador para 1993 diagnóstica problemas serios en el bachillerato cuando el 60 % de alumnos no alcanza el nivel que se requiere para continuar con la educación superior.

APRENDA 2007 señala como una de las causas principales del bajo rendimiento académico a un factor en especial que es el tipo de plantel educativo, evidenciando una gran diferencia entre un plantel urbano particular versus un plantel rural fiscal.

En 2008 las pruebas Ser Ecuador enmarcaron al estudiante entre insuficiente y regular con un decrecimiento sostenido, mientras que la minoría bueno, muy bueno y excelente se reduce tendencialmente.

SITEAL en el año 2012, al evaluar la comprensión de los estudiantes de 3° a 6° grado a nivel de la región, se demuestra que a medida que pasan los años, la deficiencia cognitiva se acentúa.

En el Tercer Estudio Regional Comparativo y Explicativo (TERCE de la UNESCO) en el año 2013, el Ecuador se ubica dentro de los países con menos del 50 % de respuestas correctas

En 2013 La prueba Ser Ecuador concluye que los problemas educativos crecen en todas las áreas destacándose el nivel insuficiente y elemental.

Entre 2013 y 2016 las tendencias crecientes de los problemas de apropiación del conocimiento persisten entre 4° y 10° de Educación General Básica (EGB) y decrecen en 3° de Bachillerato General Unificado (BGU). Lo que se puede explicar por la presencia del examen Ser Bachiller y no por la solución al problema educativo.

La OREALC, indica que Ecuador presenta puntajes que no difieren con los de la región en las diferentes áreas del conocimiento a excepción de lectura en 6° grado donde los puntajes son significativamente menores [25, Torres, 2016].

PISA en el año 2017 [19, OCDE, 2015], muestra para jóvenes ecuatorianos de 15 años los siguientes datos:

- 29 % tiene un nivel mínimo en matemáticas.

- 43 % tiene un nivel mínimo en ciencias.
- 49 % tiene un nivel mínimo en lectura.

Calificando al estudiante ecuatoriano bueno para memorizar. La memorización es buena para tareas simples, a medida que la terea se complica y requiere de plantear planes estratégicos de resolución, la memoria hace daño en lugar de ayudar [25, Torres, 2016].

Los datos recopilados indican que es necesario mejorar los indicadores de la educación en los diferentes niveles en el país, por tanto cualquier esfuerzo por detectar las causas del bajo rendimiento escolar ayudará a crear políticas que permitan contrarrestar esas causas y lograr un mejor desempeño académico que redundará en beneficio de toda la sociedad.

El Big Data y la información

El Big Data está formado por conjuntos de datos de gran tamaño y más complejos, especialmente procedentes de nuevas fuentes de datos. Estos conjuntos de datos son tan voluminosos que el software de procesamiento de datos convencional sencillamente no puede gestionarlos. Sin embargo, estos volúmenes masivos de datos pueden utilizarse para enfrentar diversos tipos de problemas que antes no hubiera sido posible solucionar [10, Hurwitz, 2013].

En los actuales momentos existe una gran cantidad de información, que las nuevas tecnologías los computadores y programas crean, procesan y transmiten, estos datos se pueden guardar en diversas formas, independientemente de la fuente, se dividen en 2 grupos: estructurados y no estructurados [8, Garside, 2013].

Los datos estructurados es información altamente organizada dentro de una tabla, donde los procesadores de información pueden fácilmente manipularla y organizarla considerando diversos criterios; este tipo de información no se puede apreciar a simple vista, es el caso de un código de barras irreconocible a simple vista pero fácil de leer por una computadora [8, Garside, 2013].

La información no estructurada no tiene un modelo predefinido o no encaja en una hoja de cálculo, su contenido se expresa en textos con datos característicos como fechas y valores, también pertenecen a esta categoría los audios y videos; su carencia de estructura hace que su copilación tome tiempo en un sistema informático, esta información es fácil de entender por el investigador [8, Garside, 2013].

Un tipo de dato estructurado son las tablas que generalmente están conformadas por n observaciones o filas y por p variables representadas por las columnas de la tabla, las variables se pueden clasificar en:

- **Variables numéricas**

  Estas variables describen una cantidad medible, responden a la pregunta cuanto o cuantos, se conocen también como variables cuantitativas, por el número que la representa pueden ser:

  - Continuas, como edad o temperatura.
  - Discretas, como número de hijos o número de bachilleres.

- **Variables categóricas**

  Este tipo de variables señalan una característica propia de un individuo u objeto, algunas de estas variables son: color, sexo, estado civil, tipo de cuenta bancaria,etc.; sus valores son cualitativos no numéricos.

  En estas variables categóricas el nombre de la categoria puede ser nominado mediante un número, por ejemplo en la variable sexo sus posibles valores son hombre o mujer; estos valores puede ser declarado y almacenado como un número, donde: hombre=1; mujer=0. Con estas categorías de forma numérica no se puede hacer aritmética, pero si estadística.

Del Big Data a la inteligencia artificial

Para entrar al campo de la Inteligencia Artificial, se necesita de un mapa que defina y diferencie los siguientes conceptos:

Inteligencia Artificial, es una disciplina de la informática que busca la creación de máquinas que puedan imitar los comportamientos inteligentes diversos como la búsqueda de patrones; en este aspecto el rendimiento que logra un ordenador es superior al humano, pero carece de la diversidad que el cerebro humano permite [10, Hurwitz, 2013].

Máquina de Aprender es parte de la inteligencia artificial con la capacidad de lograr un aprendizaje automático, induce a la máquina la capacidad de aprender, esto es generar conocimiento a partir de un conjunto de experiencias, aquí se presentan aplicaciones como árboles de decisiones, modelos de regresión, modelos de clasificación, técnicas de caracterización y otros como Redes Neuronales

que plantea un aprendizaje en un número ilimitado de capas, cada una de las cuales guarda una característica de las múltiples que puede tener el fenómeno a estudiar [10, Hurwitz, 2013].

Todo esto se hace posible por la tendencia que se ha presentado en los últimos años de acumular más y más datos; acción que se ha hecho posible por la gran capacidad de los ordenadores para guardar información, a esto se conoce como Big Data.

Minería de datos

La minería de datos emplea técnicas y algoritmos para obtener de forma automática la información sintetizada que permita caracterizar las relaciones escondidas en una gran cantidad de datos, también pretende que la información tenga capacidad predictiva (facilita el análisis de datos en forma eficiente), esta secuencia se enfoca principalmente para visualizar, analizar y modelar la información [5, Beltrán, 2008].

Es el análisis y descubrimiento de un conocimiento tomando como punto de partida grades volúmenes de datos, utilizando algoritmos que permiten:

- Agrupar información por similitudes
- Clasificación de información.
- Transformar el volumen de información.
- Predicción de valores [5, Beltrán, 2008].

Las técnicas de minería de datos, se utilizan desde hace varios años y buscan encontrar patrones y tendencias en la información, su aplicación se da en diversos campos del conocimiento, una de estas áreas es la educativa.

La minería de datos consiste en la búsqueda de patrones mediante los siguientes procesos, cada uno de los cuales tiene características y requerimientos específicos:

- Análisis de componentes principales.
- Regresión Logística.
- Análisis discriminante.
- Máquina de soporte vectorial.

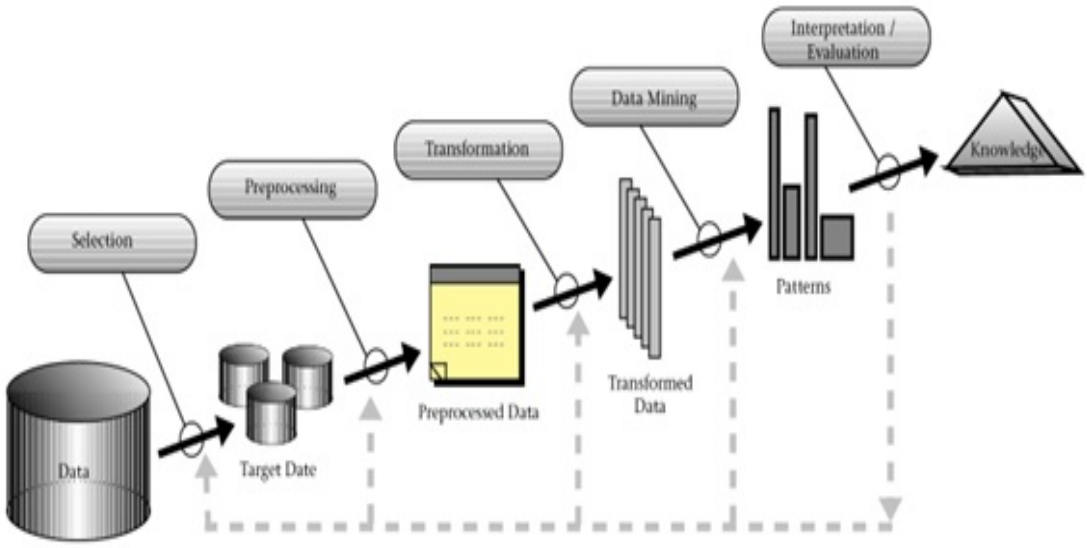
- Árbol de decisiones.
- Análisis Factorial.
- Modelamiento.

La aplicación del Big Data como también se conoce a la minería de datos, es enfrentar el elevado número de posibles relaciones que existen entre las observaciones o entre las variables,por lo que resulta imposible conocer cada una de ella. Para salvar este problema es necesario emplear estrategias de búsqueda definiendo objetivos que reflejen una visión clara de lo que se busca. El esfuerzo empleado en este proceso debe ser proporcional con el resultado buscado [4, Balcázar, 2012].

La información muchas veces cuenta con deficiencias, repeticiones, incongruencias; por este motivo debe ser previamente sometida a un proceso de depuración que permita observar como a través de una radiografía la información, para poder cumplir con el propósito de búsqueda que se tenga, Balcazar en su libro sobre Minería de Datos establece el esquema presentado en la Figura 1, donde se observa todo el proceso partiendo de una data hasta obtener el conocimiento útil y los posibles procesos de retroalimentación [4, Balcázar, 2012].

Figura 1

Proceso de Minería de Datos



Fuente: Minería de Datos Balcázar 2012 [4].

Programas

Las herramientas diseñadas para extraer conocimientos partiendo de Big Data siempre esta en continuo desarrollo, existen varios programas que permiten procesar y extraer información del Big Data, entre los que podemos mencionar los programas comerciales como Matlab, S Plus, SPSS, Excel, o los programas de software libre como Octave y R; en el presente estudio donde partimos de una base de datos que contiene gran cantidad de información y nos enfocamos en determinar la accesibilidad a la educación superior a partir de un conjunto extenso de factores asociados, se utilizan los programas Matlab y R

- Matlab

  Es la abreviatura de MATrix LABoratory, es un sotfware comercial, es un sistema de cómputo numérico con vectores y matrices, y por tanto se puede trabajar también con números escalares que pueden ser reales o complejos, con cadenas de caracteres y con otras estructuras de información más complejas; posee un lenguaje de programación propio de alto nivel y un entorno interactivo para el desarrollo de algoritmos, visualización de datos, análisis de datos y cálculo numérico. Matlab cuenta con una amplia gama de aplicaciones que incluyen procesamiento de señales e imágenes, comunicaciones, diseño de sistemas de control, sistemas de prueba y medición, modelado y análisis estadísticos y biología computacional. Para el análisis estadístico, Matlab posee el Toolbox Statistics que proporciona un conjunto completo de herramientas para evaluar e interpretar Big Data [20, Pateiro, 2016].

- R

  Es un entorno de sotfware libre que se utiliza para el procesamiento y análisis estadístico de datos, implementado en el lenguaje S de GNU. El lenguaje R es ampliamente utilizado entre estadísticos y mineros de datos para el desarrollo de software estadístico y análisis de datos, su popularidad ha aumentado en los últimos años por el aporte de la comunidad en diversas áreas. Es una herramienta excelente para el análisis de datos representados en forma de vectores o matrices; es útil para transformar y sintetizar la información. Se compila y se ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOS. R posibilita cargar librerías o paquetes con finalidades específicas de cálculo [20, Pateiro, 2016].

Análisis multivariante

El crecimiento de la aplicación de técnicas estadísticas multivariante, se ha dado en todos los campos de la investigación científica, por la necesidad del manejo de gran cantidad de información para verificar o no una hipótesis, además de la gran capacidad que en la actualidad tienen los ordenadores [7, Cuadras, 2014].

El análisis multivariante permite investigar en forma simultánea dos o más variables con la premisa de establecer las relaciones simultáneas que se den entre las variables, diferenciándose de otros análisis porque no centran su atención en establecer la media o la varianza de una variable sino al análisis de covarianzas o correlaciones que reflejen la relación entre una cantidad de variables superior a tres.

El análisis de datos multivariantes permite realizar diversas tareas entre las que podemos mencionar:

- Métodos de reducción de datos, tratan de obtener representaciones de los datos en la forma más simple posible, sin sacrificar la información que contiene la data
- Métodos de Ordenamiento y agrupación, tratan de crear grupos de objetos o de variables que poseen las mismas características. por otro lado tratan de generar reglas para clasificar los objetos u observaciones en grupos bien definidos
- Métodos para determinar las relaciones de dependencia entre las variables, pues estas relaciones entre las variables son de interés.
- Métodos de predicción, una vez que se ha determinado las relaciones entre las variables, se trata de predecir los valores de una o más variables sobre la base de los datos observados en las demás variables [7, Cuadras, 2014].

En el presente trabajo se utiliza el método de Análisis de Componentes Principales (ACP) con el propósito de representar la data en forma más condensada, los métodos de Regresión Logística (RL), Análisis Discriminante (AD) y Máquina de Vectores de Soporte (MVS), que nos permiten determinar relaciones entre las variables y a la vez para las predicciones sobre las observaciones.

Análisis de Componentes Principales

Cuando se desarrolla una investigación con muchas variables que se encuentran correlacionadas

entre sí, la información impide conocer el papel que cumple cada variable en la interpretación del fenómeno, el ACP permite reducir la cantidad de variables para quedar con un número menor de ellas que reflejan el ámbito de estudio con la menor perdida de información posible, este nuevo grupo de variables se denomina componentes principales y son combinaciones lineales de las variables originales que aparecen para sintetizar la problemática analizada y facilitar su interpretación [21, Peña, 2008].

Aunque se necesita el mismo número de componentes principales que las p variables originales, para reproducir toda la variabilidad del sistema, generalmente la mayor parte de esa variabilidad es explicada por un número pequeño de k componentes principales. Las k primeras componentes principales reemplazan las p variables originales, logrando una reducción de la dimensión del sistema [21, Peña, 2008].

Algebraicamente, las componentes principales son combinaciones lineales de las p variables aleatorias; geométricamente, estas combinacioes lineales representan la selección de un nuevo sistema de coordenadas que se obtiene al rotar el sistema original donde X_1, $X_2 \cdots X_p$ son los ejes de coordenadas. Los nuevos ejes representan las direcciones ortogonales con variavilidad máxima y proporcionas una descripción más simple de la data [21, Peña, 2008].

Suponga una data de n observaciones y p variables, la que tiene una matriz de covarianzas Σ, la misma que posee pares de valores - vectores propios (λ_1, e_1), (λ_2, e_2), $\cdots (\lambda_p, e_p)$ donde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Entonces la i-ésima componente principal esta dada por la combinación lineal

$$Y_i = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \qquad i = 1, 2, \cdots, p$$

donde:

$$Var(Y_i) = e_i^t \Sigma e_i = \lambda_i \qquad i = 1, 2, \cdots, p$$

$$Cov(Y_i, Y_k) = e_i^t \Sigma e_k = 0 \qquad i \neq k$$

$$\sum_1^p Var(X_i) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_1^p Var(Y_i)$$

donde σ_{ii} es la varianza de la variable X_i

De lo expuesto anteriormente se puede concluir:

- Varianza total = $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \lambda_1 + \lambda_2 + \cdots + \lambda_p$,

- La proporción de la varianza total explicada por la k-ésima componente principal esta dada por:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} = \frac{\lambda_k}{Varianza\,total}$$

- Se puede seleccionar las primeras componentes principales hasta cubrir una proporción determinada de la varianza total.

- La k-ésima componente del vector propio i-ésimo $e_i = [e_{i1}, \cdots, e_{ik}, \cdots, e_{ip}]$ mide la importancia de la k-ésima variable sobre la i-ésima componente principal, independientemente de las demás variables.

- El coeficiente de correlación entre la i-ésima componente principal Y_i y la variable X_k esta dado por:

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i,k = 1,2,\cdot,p$$

- En la interpretación de las componentes principales se debe considerar los coeficientes e_{ik} de las componentes y las correlaciones ρ_{Y_i,X_k}, la que nos permite analizar la importancia de las variables [21, Peña, 2008].

La deterinación del número de componentes principales que se debe escoger no tiene una respuesta definitiva, se presenta diversos criterios que nos ayudaran en esta tarea.

- Un criterio es considerar un porcentaje de la variabilidad total y escoger el número de componentes principales que satisfacen este requerimiento.

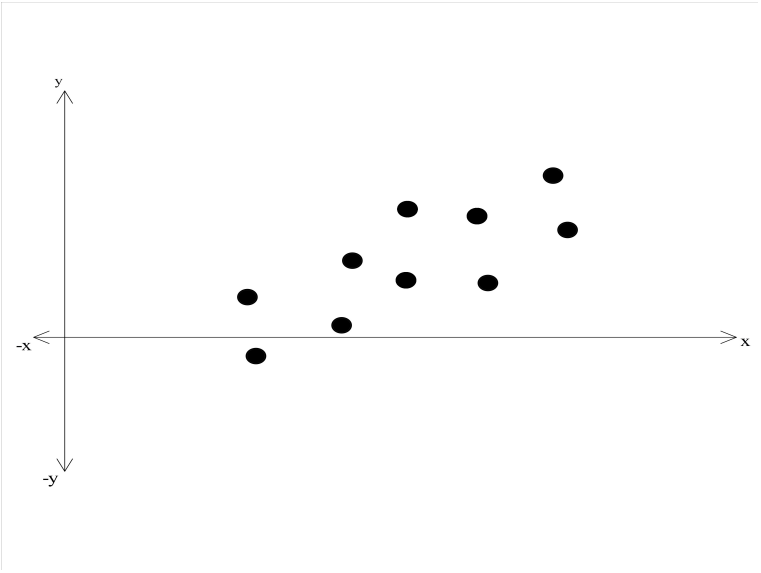
- Realizar un gráfico de λ_i contra i, seleccionamos las componentes principales hasta un punto donde los valores propios son aproximadamente iguales, es decir buscamos un codo en el gráfico, el que nos indica el número de componentes que se debe considerar. El resto de valores propios son relativamente pequeños y aproximadamente del mismo tamaño.

- Desechar aquellas componentes asociadas a valores menores a una cota prefijada, que suele considerarse como la varianza media. Cuando se trabaja con la matriz de correlaciones el valor medio es 1; en este caso el criterio es considerar las componentes principales asociadas con valores propios mayores que 1 [3, Baíllo, 2008].

Considerando la transformación más básica mediante componentes principales, que consiste en reducir de R^2 a R, se tiene el siguiente ejemplo: Se desea transformar los datos representados en la Figura 2 en dos dimensiones, a una dimensión [24, Serrano, 2018].

Figura 2

Conjunto de datos en dos dimensiones



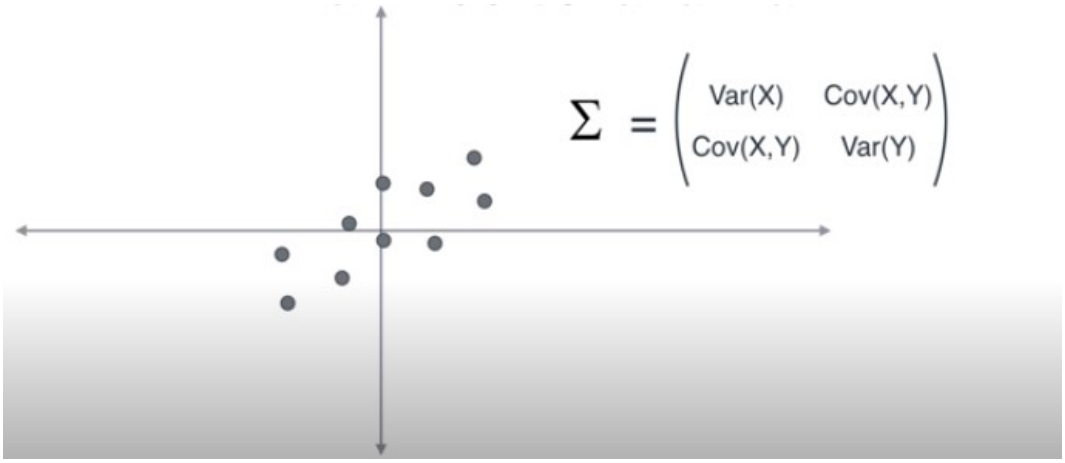
Fuente: Analisis de Componentes Principales. Luis Serrano [24].

Se determina el punto promedio, para ubicarle en el origen de coordenadas, como lo indica la Figura 3. Se considera que este conjunto de datos tiene una matriz de covarianzas Σ

$$\Sigma = \begin{pmatrix} varx & cov(x,y) \\ cov(x,y) & vary \end{pmatrix}$$

Figura 3

Conjunto de datos en dos dimensiones centrados en el origen

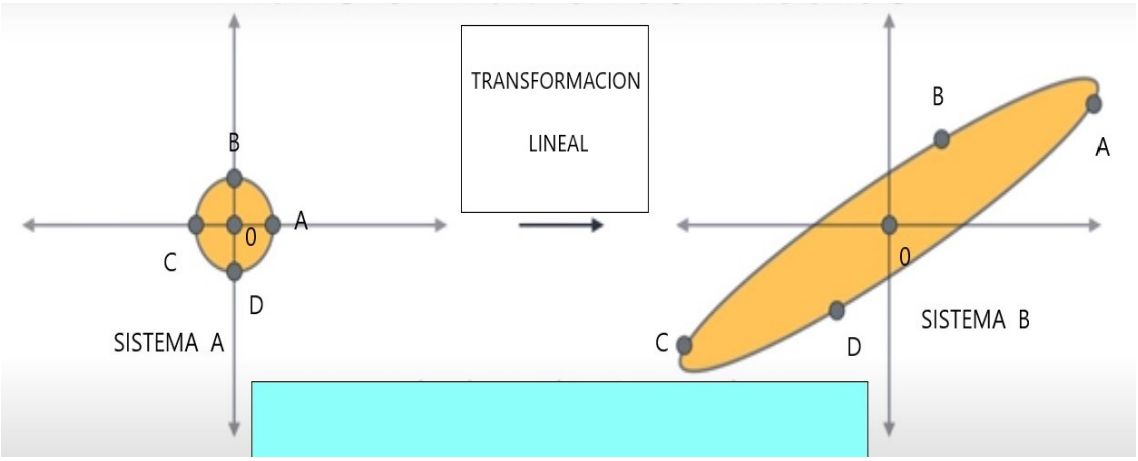


Fuente: Analisis de Componentes Principales. Luis Serrano [24].

Considerando a la matriz de covarianzas Σ como una transformación lineal, los puntos de la Figura 4 del Sistema A se transforman en los puntos del Sistema B.

Figura 4

Aplicando Σ como transformación lineal

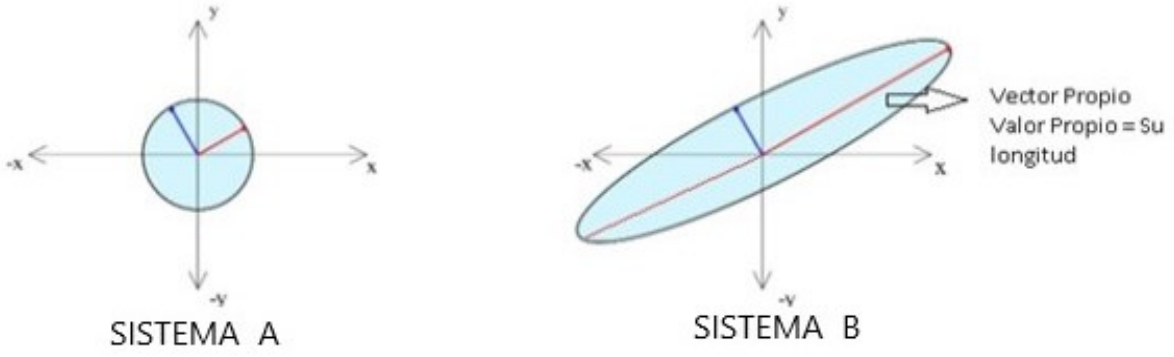


Fuente: Analisis de Componentes Principales. Luis Serrano [24].

Todos los vectores del sistema A al aplicarles la transformación lineal sufren una modificación de su magnitud y una rotación, a excepcion de los vectores propios que solo sufren una variación de su magnitud pero no rotación como lo indica la Figura 5. se busca al mayor vector propio cuya dirección indicará la primera componente principal .

Figura 5

Vectores propios y valores propios

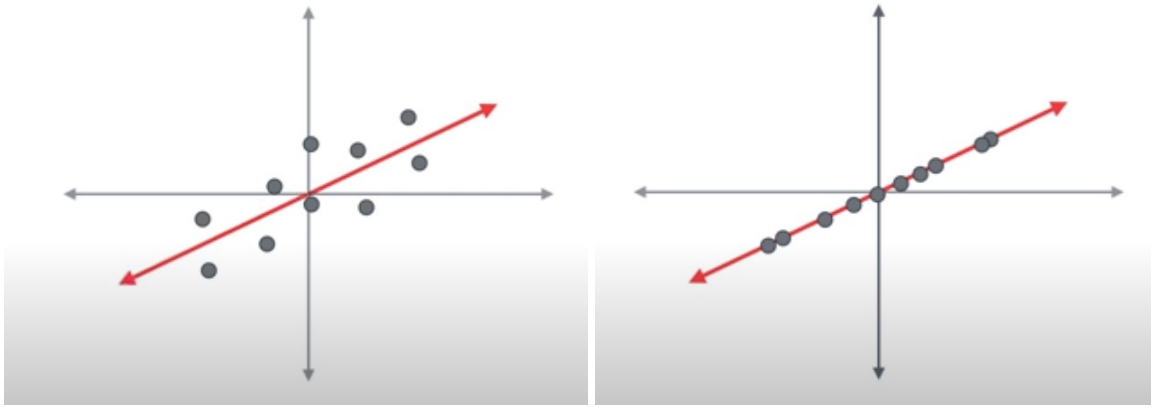


Fuente: Analisis de Componentes Principales. Luis Serrano [24].

Sobre un eje, en la dirección del mayor vector propio, se proyectan todos los puntos, obteniéndose una representación de toda la data en una sola dirección, como lo indica la Figura 6, manteniendo la máxima variabilidad de los datos.

Figura 6

Representación de los datos en una dimensión



(a) Componente principal (b) Proyeccion de los datosl

Fuente: Análisis de Componentes Principales. Luis Serrano [24].

Regresión Logística

El modelo de Regresión Logística analiza el comportamiento de una variable categórica llamada dependiente, que es binomial, es decir que solo puede tener dos valores, generalmente 0 y 1, como función de varias variables explicativas, que pueden ser cuantitativas o cualitativas.

Como la variable dependiente solo tiene dos valores posibles 0 y 1, no es posible enfrentar este modelo como una regresión lineal simple o múltiple, ya que la variable dependiente no satisface las condiciones de normalidad y de varianza constante, si a pesar de estas consideraciones se aplica la regresión lineal, esta podrá predecir valores negativos o superiores a 1 claramente fuera del intervalo [0,1], y el error es muy superior al obtenido si se ajusta mediante una función logística. La solución debe considerar por linealizar de alguna forma lo que es una relación no lineal, siendo esta la base del modelo de Regresión Logística

Plantemiento del modelo [1, Aldas, 2017]:

Se requiere analizar la relación de una variable dependiente dicotómica Y que toma valores 0 y 1, en función de una variable cuantitativa X.

En un modelo de regresión lineal la relación entre la variable dependiente Y y la explicativa X será:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

Si en lugar de una variable explicativa se tiene n de ellas, el modelo será:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{i2} + \cdots + \beta_n X_{ni} + \epsilon_i$$

A diferencia de la regresión lineal, la regresión logística no estudia las características de la recta como su intercepto con la ordenada ni tampoco su pendiente, es decir no predice el valor de Y_i dados los valores de X_i, sino que estima la probabilidad de que ocurra Y_i, es decir de que $Y_i = 1$ dados los valores de X_i. Por supuesto, considerando la relación lineal entre la variable dependiente y las independientes para calcular la probabilidad de ocurrencia de Y_i en lugar de su predigción, utilizando la siguiente función:

$$Pr(Y_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i})}} = \frac{1}{1 + e^{-Y_i}}$$

y para el caso de n variables independientes la función será

$$Pr(Y_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{i2} + \cdots + \beta_n X_{ni})}} = \frac{1}{1 + e^{-Y_i}}$$

En la Regresión Logística, el modelo se estima mediante la minimización de la función de máxima verosimilitud, que es un planteamiento similar al evaluar cuanta información queda por explicar

después de que el modelo se ha estimado

Para encontrar la máxima verosimilitud se utiliza la función:

$$LL = \sum_{i=1}^{N} [Y_i \ln(Pr(Y_i)) + (1 - Y_i) \ln(1 - Pr(Y_i))],$$

esta función toma valores cercanos a cero cuando la probabilidad predicha acierta en clasificar el caso como 0 o 1; pero en caso de desacierto crece haciéndose muy grande. La función de máxima verosimilitud no es sino la suma de valores que miden aciertos (valores pequeños) y desaciertos (valores grandes). Mayor valor implicará menor capacidad de la estimación en predecir los valores reales; es decir los parámetros β estimados serán aquellos que minimicen la función de máxima verosimilitud [1, Aldaz, 2017]

Contraste de hipótesis [1, Aldas, 2017]:

La estrategia es calcular la máxima verosimilitud de un modelo en que la función solo es formada por β_0, es decir un modelo donde las variables independientes no tienen influencia y luego se calcula la función de máxima verosimilitud para el modelo que estamos estimando. Si este segundo modelo es significativamente más pequeño que el primero se puede concluir que alguna variable debe estar ejerciendo una influencia significativa en la predicción de la variable dependiente, su β debe ser diferente de cero.

Para determinar si la diferencia es significativa, se ha determinado que $-2LL$ sigue una distribución Ji-cuadrado con grados de libertad $(K_m - K_0)$, donde:
K_m = número de parámetros a estimar en el modelo
K_0 = número de parámetros a estimar en el modelo base.

$$\chi^2 = 2LL(0) - 2LL(m) \qquad gl = K_m - K_0$$

A esta diferencia se le conoce como razón de máxima verosimilitud. En el programa estadístico R a $-2LL$ se denomina deviance.

Para el cálculo del modelo logístico en R, se usa la función glm del paquete stats, indicando que se trata de una Regresión Logística mediante el modificador family = binomial.

Contraste de los coeficientes individuales [1, Aldas, 2017]:

Una vez descartada la hipótesis nula, de que todos los coeficientes son nulos, se necesita saber cual es su contribución individual a la explicación de la variable dependiente, esto se realiza mediante el estadístico denominado test de Wald,

$$W_j = \frac{\beta_j}{SE}$$

que tiene una distribución normal. Las estimaciones de estos parámetros presenta R en la solución del modelo de Regresión Logística.

Interpretación de los coeficientes de regresión [1, Aldas, 2017]:

La influencia de los coeficientes estandarizados β en la Regresión Logística se determina mediante los odds ratio.

Se define odd de un acontecimiento como la razón entre su probabilidad de ocurrencia y la de no ocurrencia

$$odd = \frac{Pr(Y=1)}{Pr(Y=0)} = \frac{\frac{1}{1+e^{-Y}}}{1-\frac{1}{1+e^{-Y}}} = \frac{\frac{1}{1+e^{-Y}}}{\frac{1+e^{-Y}-1}{1+e^{-Y}}} = \frac{1}{e^{-Y}} = e^{Y}$$

como:

$$e^{Y} = e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{i2}+\cdots+\beta_n X_{ni}} = e^{\beta_0} e^{\beta_1 X_{1i}} e^{\beta_2 X_{i2}} \cdots e^{\beta_n X_{ni}}$$

el término e^{β_i} se denomina odd ratio y su interpretación es: es el factor en el que se incrementa el odd cuando la variable independiente i-ésima se incrementa en una unidad y el resto permanece constante.

En R la función glm del paquete stats no ofrece los odds ratio, pero su cálculo no tiene mayor dificultad: $oddratio = e^{\beta}$. La función toOR del paquete logit nos proporciona estos valores.

Si una variable se incrementa en una unidad, no significa que la probabilidad de ocurrencia se multiplica por el odd ratio, sino que el odd lo hace.

$$nuevo\,odd = viejo\,odd * odd\,ratio * cambio\,en\,la\,variable$$

y como

$$odd = \frac{Pr(Y=1)}{1-Pr(Y=1)}$$

despejando $Pr(Y=1)$ se tiene la nueva probabilidad

$$Pr(Y=1) = \frac{odd}{1+odd}$$

Evaluación del ajuste del modelo [1, Aldas, 2017]:

Para medir lo bien que el modelo predice la variable dependiente se han construidos algunos coeficientes de determinación, entre los que se puede mencionar:

- Seudo R^2 de McFadden

$$R^2_{MF} = \frac{-2LL(0) - (-2LL(M))}{-2LL(0)}$$

- R^2 de Cox y Snel

$$R^2_{CS} = 1 - e^{\frac{1}{N}(2LL(M) - 2LL(0))}$$

  donde N es el tamaño de la muestra

- R^2 de Nogelberne

$$R^2_N = \frac{R^2_{CS}}{1 - e^{\frac{2LL(0)}{N}}}$$

- Tablas cruzadas

  Una manera óptima para evaluar la precisión de las estimaciones del modelo es comparar los valores de la variable dependiente con los valores predichos. La expresión $Pr(Y) = \frac{1}{1+e^{-Y}}$ nos permite calcular las probabilidades de todas las observaciones. Si adoptamos un criterio: que cuando esta probabilidad es mayor que α, con $0 \leq \alpha \leq 1$, se asigna al grupo 1 y si es inferior a α, se le asigna al cero, se puede predecir el grupo al que de acuerdo a las variables independientes es más probable que pertenezca.

  Dado que se conoce su pertenencia real se puede generar una tabla con los valores reales y con los pronosticados, cuantos más elementos se encuentren en la diagonal principal de esta tabla mejor será la precisión del modelo.

  Sea fit el modelo de Regresión Logística resuelto mediante R con la función glm, entonces el objeto fit guardará las probabilidades de todas las observaciones en fit$fitted.values; si guardamos estas probabilidades en la variable predict.fit, podemos reasignar a esta variable los valores de 1 y 0, según el valor de la probabilidad sea mayor o menor de un valor preestablecido α. Con los valores reales de la variable dependiente y con los valores predichos se crea una tabla cruzada con el comando CrossTable del paquete gmodels.

  El código en R que realiza lo indicado es

```
predict.fit<-fit$fiftted.values
predict.fit[predict.fit>=.50]<-1
predict.fit[predict.fit<.50]<-0
CrossTable(datod$y,predict.fit,format=SPSS,chisq=FALSE,
prop.c=FALSEprop.r=FALSE)
```

En el análisis de la precisión de un modelo se debe tener en cuenta los siguientes parámetros:

- Sensibilidad: % de positivos que son clasificados como positivos.
- Especificidad: % de negativos que son clasificados como negativos.
- Falsos positivos: % de negativos clasificados como positivos
- Falsos negativos: % de positivos clasificados como negativos

El modelo puede calibrarse en función de estos parámetros, modificando el valor de la probabilidad de ocurrencia, a partir de la cual se predice la pertenencia al grupo 1 o al grupo 0.

Análisis Discriminante

El análisis discriminante es un modelo que se utiliza para explicar la pertenencia de los individuos a diversos grupos, a partir de los valores de un grupo de variables que definen el perfil del individuo que se desea clasificar, con la restricción de que cada individuo debe pertenecer a un solo grupo. En este análisis se construye una variable categórica con tantas categorías como grupos existan y esta constituye la variable dependiente.

Las variables que intervienen en la clasificación de los individuos toman el nombre de variables clasificadoras o explicativas, cuya información se sintetiza en unas funciones llamada funciones discriminantes, que son las utilizadas en el proceso de clasificación [1, Aldas, 2017].

El análisis discriminante se utiliza con dos finalidades:

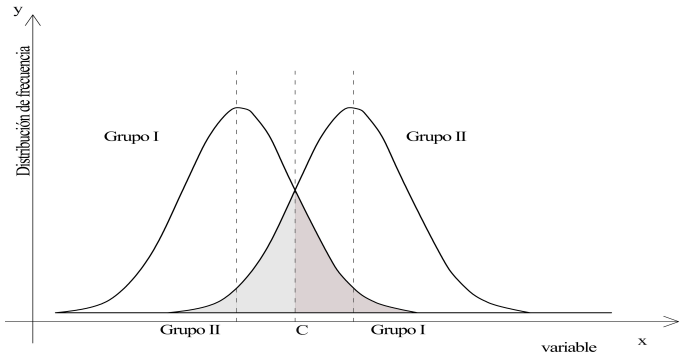
- Explicativa: Determina la contribución de cada variable a la clasificación correcta de los individuos.
- Predictiva: Se busca establecer el grupo al que más probablemente pertenece el individuo, conociendo los valores que toman las variables explicativas.

Clasificación en dos grupos mediante una variable clasificatoria [1, Aldas, 2017]:

Para una mejor comprensión del modelo consideremos el caso más simple que consiste en clasificar a los individuos en dos grupos en función de una sola variable clasificatoria. Considerando que se tiene para los dos grupos las mismas distribuciones de frecuencias y la misma varianza; es decir, los dos grupos coinciden en todo excepto en su media como se observa en la Figura 7, las distribuciones de frecuencia se solapan, debido a este solapamiento pueden cometerse errores de clasificación.

Figura 7

Funciones de distribución de 2 grupos similares con diferentes medias



Fuente: Análisis multivariante aplicado con R Aldas [1].

Sea $\bar{X}_I$ y $\bar{X}_{II}$ las medias de los grupos I y II respectivamente, el punto de intersección de las dos funciones corresponde al valor medio de $\bar{X}_I$ y $\bar{X}_{II}$ al que llamaremos C, es igual a

$$C = \frac{\bar{X}_I + \bar{X}_{II}}{2}$$

El valor de C permite definir el criterio de clasificación y se denomina punto de corte:

- Si $X_i < C$, se clasifica al individuo i en el grupo I.
- Si $X_i > C$, se clasifica al individuo i en el grupo II

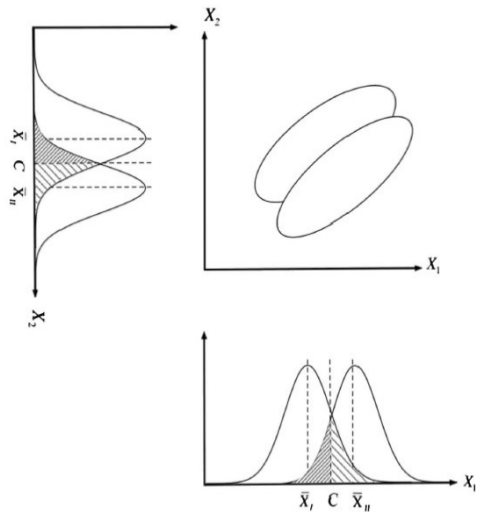
Con este criterio se cometen errores de clasificación, como se observa en la Figura 7, el área sombreada a la derecha de C representa a los individuos del grupo I incorrectamente clasificados en el grupo II, de igual forma el área sombreada a la izquierda del punto C representa a los individuos del grupo II incorrectamente clasificados en el grupo I

Clasificación en dos grupos mediante dos variables clasificatorias [1, Aldas, 2017]:

En este análisis se considera la presencia de dos variables clasificatorias, la Figura 8 muestra las elipses de concentración de los datos de dos distribuciones de frecuencias bivariantes donde las variables X_1 y X_2 estan correlacionadas positivamente. Las dos elipses tienen el mismo tamaño pero su centro es diferente

Figura 8

Distribuciones de frecuencias bivariantes y sus proyecciones sobre los ejes X_1 y X_2



Fuente: Análisis multivariante aplicado con R. Aldas [1].

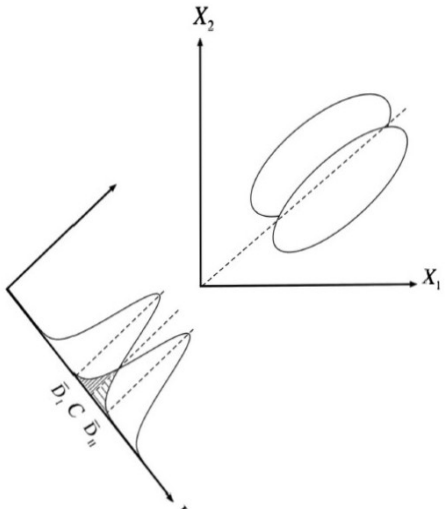
.

Debajo del eje X_1 se presenta la proyección de las distribuciones de frecuencias bivariante sobre este eje, la misma que presenta las distribuciones univariante de la variable X_1, las distribuciones de los dos grupos se encuentran solapadas. Realizando el mismo proceso sobre el eje X_2, se tiene las distribuciones de frecuencia marginales de la variable X_2, las que también se encuentran solapadas. Mientras mayor sea el solapamiento mayor sera el número de individuos clasificados equivocadamente.

Se puede obtener una mejor función discriminante utilizando las dos variables conjuntamente. Proyectando las dos distribuciones de frecuencia sobre un eje oblicuo como lo indica la Figura 9 obteniéndose distribuciones de frecuencia que están menos solapadas que las distribuciones marginales. Variando la inclinación se obtienen distribuciones con distinto grado de solapamiento. al momento de obtener el mínimo solapamiento se tendra el eje óptimo, llamado eje discriminante.

Figura 9

Distribuciones de frecuencias bivariantes y sus proyecciones sobre el eje discriminante



Fuente: Análisis multivariante aplicado con R. Aldas [1].

La variable obtenida en la proyección sobre el eje discriminate a la que nominaremos como D es la función discriminante, la figura 10 muestra lo indicado. Se observa que el solapamiento de la función discriminante es menor al que presentan las distribuciones de frecuencias de la variable sobre los ejes X_1 y X_2, lo que implica que con la distribución de frecuencias de la función discriminante se consigue que el número de individuos clasificados erróneamente sea menor.

Clasificación en dos grupos mediante k variables clasificatorias [1, Aldas, 2017]:

Generalizando el proceso realizado para dos variables explicativas a k variables, con la finalidad de que la mayor cantidad de individuos clasificados sea correcta, se utiliza la función discriminante de Fisher, quién resolvió analíticamente este problema. La función discriminante de Fisher D, se plantea como una combinación lineal de las k variables explicativas:

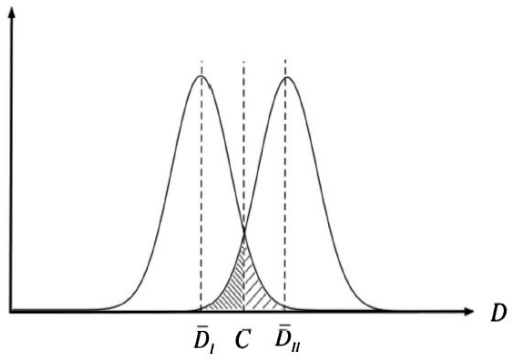
$$D = u_1X_1 + u_2X_2 + ... + u_kX_k$$

El problema consisde en determinar los coeficientes u_j de la combinación lineal. Como existen n observaciones, se puede expresar la función discriminante para cada una de las n observaciones

$$D_i = u_1X_{i1} + u_2X_{i2} + ... + u_kX_{ik} \qquad i = 1, 2, \cdots, n$$

Figura 10

Funciones de distribución de frecuencias sobre el eje discriminante



Fuente: Análisis multivariante aplicado con R. Aldas [1].

donde: D_i es la puntuación discriminante de la observación i-ésima.

Considerando que las variables explicativas están expresadas en desviaciones respecto a la media, D_i también lo estara.

Expresando la anterior ecuación en forma matricial para las n observaciones se tiene:

$$\begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{K1} \\ X_{12} & X_{22} & \cdots & X_{K2} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{Kn} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_K \end{bmatrix}$$

En forma simplificada se puede expresar como

$$\mathbf{d=Xu}$$

La variabilidad de la función discriminante se expresa como:

$$\mathbf{d'd=u'X'Xu} \tag{1}$$

La matriz $\mathbf{X'X}$ representa la matriz de suma de cuadrados y productos cruzados (SCPC) total de las variables explicativas X, que pueden descomponerse en las matrices SCPC entre grupos y SCPC

intragrupos; es decir,

$$\mathbf{X}'\mathbf{X} = \mathbf{T} = \mathbf{F} + \mathbf{W} \tag{2}$$

donde:

- **T**: matriz SCPC total
- **F**: matriz SCPC entre grupos
- **W**: matriz SCPC intagrupos

Reemplazando la ecuación 2 en la ecuación 1, se tiene:

$$\mathbf{d}'\mathbf{d}=\mathbf{u}'\mathbf{T}\mathbf{u}=\mathbf{u}'\mathbf{F}\mathbf{u} + \mathbf{u}'\mathbf{W}\mathbf{u}$$

Las matrices **T**, **F** y**W** se pueden calcular con los datos muestrales, pero los coeficientes u_i se deben determinar. Para su determinación Fisher utilizó el criterio: la función discriminante separará mejor los grupos cuando los grupos resultantes sean lo más iguales posibles dentro de ellos y lo más diferentes posible entre grupos, lo que se traduce en maximizar la razón entre la variabilidad entre grupos y la variabilidad intragrupos. Lo que se puede expresar como

$$MAXIMIZAR \quad \lambda = \frac{\mathbf{u}'\mathbf{F}\mathbf{u}}{\mathbf{u}'\mathbf{W}\mathbf{u}}$$

Esta expresión permite determinar los coeficientes normalizados u_j con $j = 1, 2, \cdots, k$, donde normalizados significa que la suma de sus cuadrados es uno.

Para determinar el punto de corte discriminante, que permitirá separar de mejor manera los grupos, se calcula el valor de la función discriminante en los centroides de cada grupo, para luego promediar estos valores, es decir

$$C = \frac{\bar{D}_I + \bar{D}_{II}}{2}$$

donde

- $\bar{D}_I$ valor de la función discriminante evaluado en el centroide del grupo I
- $\bar{D}_{II}$ valor de la función discriminante evaluado en el centroide del grupo II

Con este valor de C, clasificamos a los individuos i con el siguiente criterio:

- Si $D_i <$ C $\longrightarrow$ el individuo i pertenece al grupo I
- Si $D_i >$ C $\longrightarrow$ el individuo i pertenece al grupo II

Esta clasificación definirá 2 grupos con las siguientes características

- Ser lo más homogéneos posible dentro del grupo.
- Ser lo más diferentes posible entre grupos.

Calidad de la clasificación [1, Aldas, 2017]:

El cálculo de la función discriminante mediante el programa estadístico R se realiza utilizando la función **lda** del paquete **MASS**, que proporciona los coeficientes de la función discriminante. Este paquete estadístico posee además la función **predict** que permite ya sea clasificar nuevos casos o clasificar los casos que se utilizaron para generar la función discriminante. Al aplicar la función predict sobre los datos originales, se obtiene para todos los individuos la clasificación que realiza el modelo, lo que permite conjuntamente con los datos reales realizar una tabla cruzada para determinar la calidad de la clasificación

Supuestos del análisis discriminante [1, Aldas, 2017]:

- Todos los grupos tienen la misma matriz de covarianzas (hipótesis de homocedasticidad)
- Cada grupo tiene una distribución normal multivariante (hipótesis de normalidad)
- Las muestras de cada grupo son aleatorias e independientes.

Para verificar la hipótesis de homocedasticidad mediante R, se puede utilizar la función boxM del paquete biotools. Para el análisis de la normalidad multivariante se puede usar la función mshaphiro.test del paquete mvnormtest, este test necesita que las variables esten en filas por lo que es necesario modificar la data.

Predicción de nuevos casos [1, Aldas, 2017]:

La predicción de nuevos casos puede hacerce automáticamente utilizando la función predict del paquete MASS en el programa R, aplicando a una nueva base de datos que contiene los individuos que se desea clasificar

Máquina de Soporte Vectorial [2, Amat, 2017]

Es un modelo de clasificación supervisado, es decir tiene una variable dependiente, que indica a que grupo pertenece la observación, tomará el valor de 1 si la observación pertenece al grupo I y de -1 si pertenece al grupo II. Este algoritmo se fundamenta en buscar una recta, superficie o hiperplano que separe lo mejor posible las observaciones en dos grupos, Dada una situación en la que se pueda tener diversas soluciones que separen las clases, la solución ideal será la que maximice las distancias entre las observaciones de las diferentes clases; por tanto se debe buscar estos puntos a los que se llamará vectores de soporte que maximicen esa distancia.

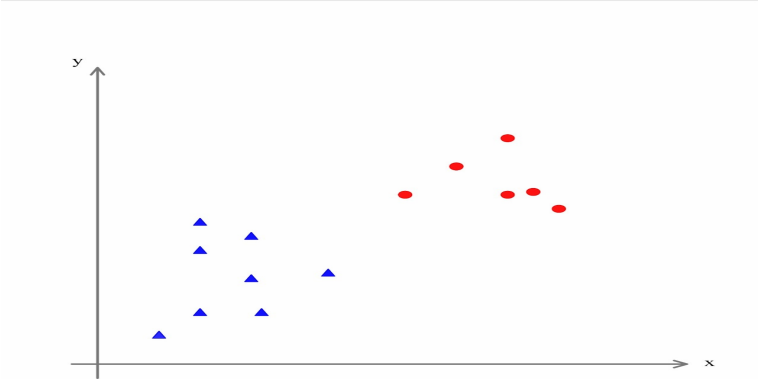
Las características principales de los vectores de soporte son:

- Son puntos de la data que estan muy cercanos de la recta, superficie o hiperplano de decisión.
- Son puntos muy dificiles de clasificar.
- Tienen mucha influencia en la posición de la superficie de decisión, es decir la modificación de los vectores modificará el clasificador.

En este estudio se utilizará exclusivamente para clasificar. Dados el conjunto de datos, el algoritmo genera un elemento separador (hiperplano óptimo) que define claramente los grupos formados y a la vez permite clasificar a nuevos elementos. Lo expuesto se representa en forma gráfica, la Figura 11, representa a un conjunto de datos pertenecientes a dos clases diferentes a los que se quiere clasificar utilizando este modelo.

Figura 11

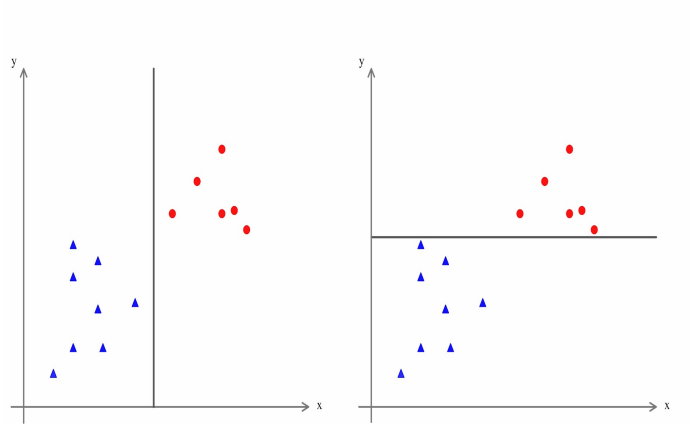
Conjunto de datos que pertenecen a dos grupos



Construir el hiperplano que separe las dos clases existentes en los datos, presenta varias posibili-

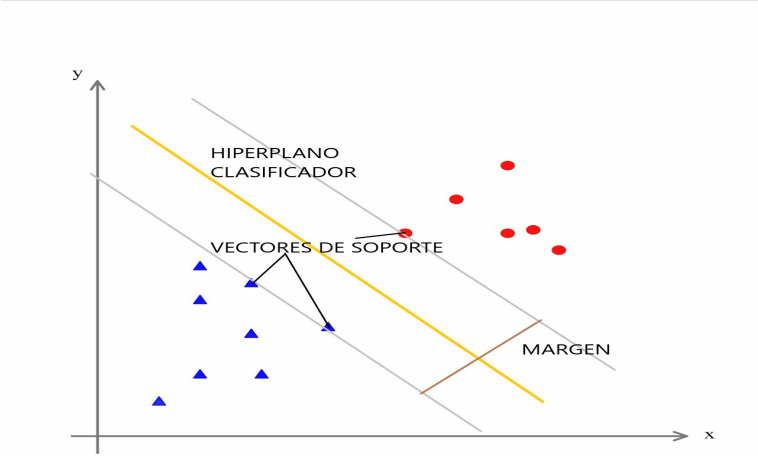
dades como se observa en la Figura 12.

Figura 12
Diferentes hiperplanos que separan dos grupos



En ambos casos el hiperplano separa perfectamente en dos clases los datos, pero lo ideal es buscar una posición que maximice la distancia entre los elementos de frontera. La Figura 13 indica el hiperplano óptimo donde se ha maximizado la distancia entre los elementos mas próximos de las dos clases. Se crea 2 líneas paralelas equidistantes, una hacia el lado positivo y otra hacia el lado negativo del hiperplano, definiendo los vectores de soport y la banda entre los datos toma el nombre de margen máxmo.

Figura 13
Hiperplano óptimo y elementos del modelo



El modelo cuando necesita clasificar un elemento nuevo, en función de los vectores de soporte

ubica en uno o en otro lado del hiperplano, clasificándole.

Si no es posible realizar la clasificación en el espacio dimensional dado, por existir una mezcla entre las observaciones de los diversos grupos, el modelo aumenta la dimensión para determinar si en este nuevo espacio vectorial es posible construir el hiperplano clasificador, proceso que puede continuar hasta lograr una separación adecuada.

El cálculo aplicando el modelo de Máquina de Vectores de Soporte, mediante el programa estadístico R se realiza utilizando la función svm del paquete e1071, este paquete estadístico posee además la función predict que permite ya sea clasificar nuevos casos o clasificar los casos que se utilizaron para generar el modelo. Al aplicar la función predict sobre los datos originales, se obtiene para todos los individuos la clasificación que realiza el modelo, lo que nos permitirá conjuntamente con los datos reales realizar una tabla cruzada para determinar la calidad de la clasificación.

Revisión de trabajos realizados

La Tesis de Maestría de Tania Pesantes de la Escuela Politécnica Nacional del año 2013, tiene por tema: Construcción de un índice a partir de datos SNIESE 2010 mediante Análisis de Componentes Principales. Este estudio busca construir un Indice de Pertenencia de 71 universidades del país, es decir conocer la vinculación con la sociedad y el entorno productivo que estas universidades tienen en sus localidades. La técnica multivariante que se utiliza es el Análisis de Componentes Principales no lineales, a partir de variables que son de caracter cuantitativo y cualitativo. De este grupo de variables, se busca un conjunto menor de variables abstractas representativas, para repetir el proceso hasta la obtención de un solo valor que es una proporción del índice buscado. El índice esta en un rango [0,100] puntos; por lo que la universidad mejor puntada es asignada con 100 puntos y a las otras universidades les corresponde un valor proporcional [23, Pesantes, 2013].

La Tesis de Maestría de Juan Carlos López de la Escuela Politécnica Nacional del año 2017, con el tema: Satisfacción laboral del maestro y rendimiento académico en Matemáticas a partir de los factores asociados de las pruebas Ser Maestro y Ser Bachiller 2014. A partir de la información de los maestros y utilizando modelos de Ecuaciones Extructuradas, se construye un índice de satisfacción del maestro, con los datos de los estudiantes se construyen modelos multinivel para establecer las variables que más aportan al aprendizaje en el aula. La clasificación binaria de satisfacción del maestro a partir del índice construido se introduce en el modelo multinivel óptimo para identificar la inferencia de la satisfacción del maestro en el rendimiento académico en matemáticas del bachiller

[18, López, 2015].

El informe: Resultados de Pisa Para el Desarrollo primera edición 2918, La aplicación de Pisa Para el Desarrollo, en Ecuador esta a cargo del INEVAL y plantea un sistema de evaluación del rendimiento académico de 325 297 adolescentes de 15 años en 2 018, para comparar y posicionar con otros jóvenes de 90 países. Los Factores Asociados al rendimiento académico que plantea Pisa, se agrupan en:

- Recursos económicos,
- apoyo familiar,
- calidad de la institución,
- ambiente escolar,
- salud y bienestar,
- entorno social.

La evaluación consta de alrededor de 400 items que se relacionan al desempeño del estudiante en: comprensión lectora, actitud matemática, conocimiento literario y destrezas en ciencias [19, OCDE, 2015].

TERCE de UNESCO: Es una evaluación educativa a gran escala que involucra a 16 países de la región, con la participación de más de 3 000 escuelas, en el tercero y sexto grado de primaria; desarrollada en las áreas de matemáticas, lenguaje y ciencias. Para esta evaluación se utiliza 2 elementos:

- Prueba para medir conocimientos,
- Cuestionarios al director, profesores, alumnos y padres de familia.

La prueba y las encuestas son realizadas por el Laboratorio Latinoamericano de Evaluación de la Calidad Educativa (LLECE), que forma parte de la UNESCO. Con la prueba y la encuesta se realiza el análisis de los Factores Asociados que tienen mayor incidencia en el resultado académico. El propósito es dar cuenta del estado de la educación en la región y guiar la toma de decisiones en políticas públicas educativas [17, LLECE, 2015].

Capítulo 3

Marco Metodológico

En el presente capítulo se abordará la metodología para predecir la accesibilidad de los bachilleres a la universidad pública en el año 2019, a partir del examen Ser Bachiller y la encuesta de Factores Asociados, partiendo de la preparación de los datos, etapa que da paso al Análisis de Componentes Principales (PCA), para continuar con los modelos predictivos:

- Regresión Logística
- Análisis Discriminante
- Máquina de soporte vectorial

De la comparación de los resultados de cada uno de estos modelos, se determinará la predicción con mayor precisión.

Preparación de la información

En la limpieza de la información, se reducirá la data mediante la utilización de argumentos como: evitar la duplicidad de información, corregir incoherencias, solucionar la falta de información entre otras causas; este proceso de filtrado de la información, afectará tanto a observaciones como a variables, hasta obtener una matriz de datos donde todas las observaciones tengan información significativa en todas las variables escogidas.

Esta matriz debe contener datos valederos y completos que simplifique los procesos de regresiones y transformaciones lineales, llegando a un punto de equilibrio en el cual se evidencie que la información en su totalidad no pierda la esencia del estudio.
Se empleará dos mecanismos de reducción de información, el primero es la Depuración de la información en el que se elimina observaciones y variables respaldados por un argumento y el segundo el Perfilamiento que jerarquiza las variables que han sido utilizadas en estudios similares a nivel internacional.

Depuración de la data

Los datos iniciales que entregó el Instituto Nacional de Evaluación (INEVAL), consisten en 2 matrices, la primera llamada de Factores Asociados, contiene información social, económica, familiar y de desenvolvimiento en el ambiente escolar; la segunda llamada de Notas contiene los resultados de la evaluación Ser Bachiller 2019, en las asignaturas Lenguaje, Matemáticas, Ciencias Naturales, Ciencias Sociales, el promedio de estas, la nota modificada por el índice socio ecnómico del INEVAL, información del establecimiento escolar, de la ubicación geográfica y sobre su situación personal.

$$FactoresAsociados = \begin{pmatrix} f & \cdots & f \\ \vdots & \ddots & \vdots \\ f & \cdots & f \end{pmatrix} = 513806 \; x \; 324$$

$$Notas = \begin{pmatrix} n & \cdots & n \\ \vdots & \ddots & \vdots \\ n & \cdots & n \end{pmatrix} = 514852 \; x \; 34$$

Selección de datos válidos:

Se realizará un proceso mediante el cual se eliminará variables (columnas de la matriz) u observaciones (filas de la matriz), por las siguientes causas:

- Eliminación de filas o columnas de la data por no contener información
- Eliminación de columnas cuando la característica conceptual de dos o más factores asociados sea similar
- Eliminación de columnas cuando todas las observaciones tienen igual respuesta o esta no es significativa

Eliminación de filas o columnas por no contener de información.

La Tabla 1, representa un extracto de la data, sobre la cual ejemplificamos los criterios antes indicados, así:

- Las variables 2 y 7 se eliminan por no tener información.
- La variable 3 se elimina porque todos sus valores son iguales.

- La observación 2 se elimina por no contener información en la mayoría de sus celdas

Tabla 1
Eliminación de filas y columnas en la data

	Var.1	Var.2	Var.3	Var.4	Var.5	Var.6	Var.7	.....	Var.n
Observa. 1	1	99999	0	1	0	1	99999		0
Observa. 2	0	99999	0	99999	99999	99999	99999		99999
Observa. 3	0	99999	0	1	0	1	99999		1
.....									
Observa. n	1	99999	0	0	0	1	9999		0

Nota: Los valores 99999 representan datos faltantes

La matriz de Notas tiene un alto porcentaje (alrededor del 40 %) de observaciones que no presentan los valores de las evaluaciones parciales ni del valor promedio, las cuales deben ser eliminadas.

Cuando la característica conceptual de los factores asociados sea similar:

En este proceso se calificará el concepto de las variables para encontrar similaridades, que permitan eliminar respuestas parecidas. La Tabla 2, representa un ejemplo en el que se presentan dos variables similares, de las cuales se toma como única variable a la signada con el número 1

Tabla 2
Eliminación de variables similares

Variable	Concepto	Categorías
Variable 1	Tiene hijos	si
		no
Variable 2	¿Cuantos hijos tiene?	0
		1
		2

Cuando todas las observaciones tienen igual respuesta o esta no es significativa:

Tabla 3
Eliminación de variables que no aportan información

Variable 1	Variable 2	Variable 3
1	1	0
1	1	0
0	1	1
1	1	0
0	1	1
0	1	1
1	1	0

En la Tabla 3, la variable 2 todos tienen la misma respuesta para todas las observaciones, por

lo que esta debe ser eliminada ya que no aporta con información. Un ejemplo de este proceso es la variable 'Es honesto', la respuesta en el 100 % de observaciones fue si, esta homogeneidad no define perfiles en las observaciones, por lo que esta variable es eliminada.

Depuración de la data:

Con los criterios indicados la depuración de la data se realizará de la siguiente manera:

- En la matriz de Factores Asociados, se mantendrá exclusivamente a aquellas variables cuya información es mayor al 50 % de las observaciones.

$$FactoresAsociados = \begin{pmatrix} f & \cdots & f \\ \vdots & \ddots & \vdots \\ f & \cdots & f \end{pmatrix} = 511470\ x\ 139$$

- En la matriz Notas se eliminará:
  - A las observaciones que no tienen información ya que no dieron la evaluación.
  - A las observaciones que solo tienen el parámetro de asignación a la universidad y no tienen resultados de las evaluaciones.
  - A las variables duplicadas en la matriz de Factores Asociados.

$$Notas = \begin{pmatrix} n & \cdots & n \\ \vdots & \ddots & \vdots \\ n & \cdots & n \end{pmatrix} = 299015\ x\ 16$$

- Se unen estas dos matrices, la de Factores Asociados y la de Notas y sobre esta se eliminan las variables que posen la misma característica conceptual de otra, luego de este proceso se obtiene:

$$FactoresAsociadosNotas = \begin{pmatrix} fn & \cdots & fn \\ \vdots & \ddots & \vdots \\ fn & \cdots & fn \end{pmatrix} = 299015\ x\ 152$$

Perfilación

Esta etapa, busca reducir exclusivamente la cantidad de variables, para lo cual se considera estudios sobre acceso a la universidad realizados por universidades de la región, así como de insti-

tuciones que analizan la gestión educativa de la juventud mientras cursan la educación básica y el bachillerato, estos estudios fueron realizados en los últimos 3 años.

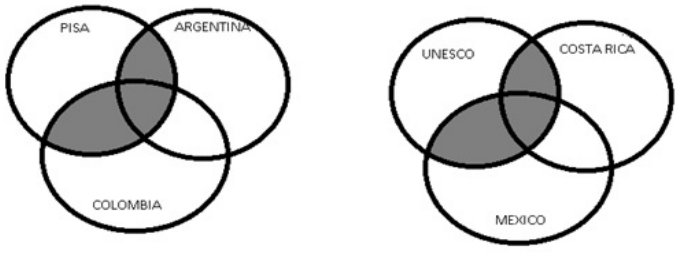
Los estudios indicados, utilizan Factores Asociados similares a los que contempla el Ineval en la encuesta realizada a los estudiantes que rindieron la prueba Ser Bachiller; las universidades y organizaciones dedicadas a observar la educación, no utilizan los mismos factores; con la finalidad de lograr similitud con estos estudios, se tomará como principales factores a aquellos que sean recurrentes.

Los estudios considerados pertenecen a las siguientes universidades e instituciones:

- Unesco
- Universidad de Costa Rica
- Universidad de México
- Pisa
- Universidad de Argentina
- Universidad de Colombia

A estas estudios se los dividió en dos grupos de tres, en cada grupo está presente una institución con cobertura mundial. Los Factores Asociados significativos serán los que utilicen en forma común las diferentes instituciones, dando privilegio a Pisa y Unesco de acuerdo al siguiente esquema:

Figura 14
Análisis de las variables utilizadas en estudios similares



Los Factores Asociados seleccionados son los que se encuentran en las regiones sombreada de la Figura 14, los mismos que constituyen 45 variables que se detallan a en la tabla 4

Para un mejor manipuleo de los factores asociados seleccionados, al código asignado por el INEVAL se le reduce a cuatro caracteres, generalmente eliminando los últimos tres, esta reducción permitirá realizar gráficos con mayor claridad en la información, este cambio es dirigido exclusivamente al código de la variable, manteniendo todas las demás características de las variables sin ninguna alteración. En la Tabla **??** de los anexos, se presenta la lista de variables escogidas con sus respectivas categorías y los valores asignados a cada categoría.

Los códigos originalmente asignados a estas variables y los nuevos códigos simplificados se listan a continuación manteniendo el mismo orden.

“nm_regi" “tp_sost" “aguaabe" “allcabe" “amocbbe" “atrsabe" “basuabe"
“cboxabe" “clapabe" “clavabe" “cocqbbe" “crefbbe" “cuarpbe" “desaabe"
“expeabe" “htardbe" “inasabe" “injhbbe" “inmebbe" “inteabe" “jhogabe"
“luzeabe" “nbanabe" “ncarabe" “ncelabe" “nhijabe" “niessbe" “nlibcbe"
“nmicabe" “nvcaabe" “ocjhabe" “ocstbbe" “pardabe" “piscabe" “prdscbe"
“rabupbe" “segebbe" “sexoabe" “tclabbe" “tesccbe" “tpviabe" “tvstabe"
“ucorabe" “veccabe" “vivfybe"

“regi", “sost", “agua", “allc", “amoc", “atrs", “basu",
“cbox", “clap", “clav", “cocq", “cref", “cuarp", “desa",
“expe", “htar", “inas", “injh", “inme",“inte", “jhog",
“luze", “nban", “ncar", “ncel", “nhij", “niess", “nlib",
“nmic", “nvca", “ocjh", “ocst", “pard", “pisc", “prds",
“rabu", “sege", “sexo", “tcla", “tesc", “tpvi", “tvst",
“ucor", “vecc", “vivy"

Tabla 4
Variables seleccionadas de la data.

Código	Nombre de la variable
regi	Región natural del Ecuador
sost	Tipo de sostenimiento
sexo	¿ Eres hombre o mujer?

Sigue en la siguiente página

Código	Nombre de la variable
agua	Tienes agua potable o entubada en tu casa
allc	Se llevan bien entre compañeros de aula
amoc	Te siento amenazado por algún compañero
atrs	En el último mes de clase que tuviste llegabas tarde al colegio
basu	Tienes servicio de recolección de basura en casa
clav	Tienes lavadora de ropa en casa
cocq	Tienes cocina con horno
cref	Tienes refrigeradora en casa
cuarp	Tienes cuarto propio
desa	Tienes desague o alcantarillado en tu casa
expe	Lo que aprendí en el colegio cubre mis expectativas de aprendizaje
htar	Dedico tiempo para estudiar y hacer tareas
inas	En el último mes de clase no asistía a clases en algunas materias
injh	Señale el nivel de instrucción más alto que ha llegado el jefe del hogar
inme	Los maestros están interesados en que los estudiantes estemos bien
inte	Tienes conexión a internet en tu casa
jhog	¿Quién es el jefe de tu hogar?
luze	Tienes luz eléctrica en casa
nban	Cuantos baños hay en funcionamiento en tu hogar
ncar	Tienes automóvil en tu casa
niess	Tienes familiares que viven en tu casa y estan afiliados al IESS
nlib	En tu casa tienen libros
nmic	Tienen microondas en tu hogar
nvca	Tienen cámara de video digital en tu casa
ocjh	Señale el trabajo del jefe del hogar
ocst	¿Cuál es tu principal ocupación?
pard	Tipo de material de las paredes de tu casa
pisc	Tipo de materiales de los pisos de tu casa
prds	Tus familiares o representantes te preguntan por tus deberes del colegio
rabu	En el último o presente período escolar faltaste más de 15 días al colegio
sege	Me siento seguro cuando estoy en el colegio

Sigue en la siguiente página

Código	Nombre de la variable
tcla	¿Los maestros se preocupaban de que aprovechemos el tiempo ?
tesc	¿Cuanto tiempo tardas en llegar a tu colegio?
tpvi	En qué tipo de viviendas habitas
tvst	Tienes televisión por cable o satelital en tu casa
ucor	Alguien de tu hogar utiliza correo electrónico que no sea del trabajo
vecc	Compran ropa en centros comerciales
vivy	Vives con cónyuge conviviente o pareja
cbox	Tienes PlayStation o Xbox en casa
clap	Tienes lapton
nhij	Tienes hijos
ncel	¿Cuantos celulares con conexión a internet hay en tu hogar?

Nota: Al código proporcionado por el INEVAL se eliminó los últimos 3 caracteres.

En las variables seleccionadas, a las categorías que poseen pocos individuos $(< al\, 5\,\%)$ del total se la elimina, incluyéndola en otra categoría similar, lo que permite tener categorías robustas para evitar distorciones en los análisis predictivos posteriores. El número de observaciones en cada una de las categorías de las variables seleccionadas se presenta en la Tabla 5

Considerando las variables de identificación de las observaciones (Código de la observación e ID), la variable dependiente (promedio de las evaluaciones Ser Bachiller) y los 45 factores asociados seleccionados, se tiene una matriz de 299015 observaciones y 48 variables, la misma que contiene algunas celdas sin información, lo que dificulta la aplicación de los modelos predictivos, por lo que se procede a un control fila por fila, eliminando la fila donde exista algún dato faltante logrando una matriz final completa; es decir, que todas las celdas contiene información válida, y sobre la cual plantearemos los modelos predictivos. Las dimensiones de esta matriz son:

$$FactoresAsociadosNotasFinal = \begin{pmatrix} fn & \cdots & fn \\ \vdots & \ddots & \vdots \\ fn & \cdots & fn \end{pmatrix} = 265915 \ * \ 48$$

Tabla 5

Número de observaciones en cada categoría

	regi		sost		agua		allc		amoc		atrs
0	151251	0	189897	0	33899	0	21830	0	259678	0	51432
1	114664	1	21843	1	232016	1	244085	1	6237	1	214483
		2	54175								

	basu		cbox		clap		clav		cocq		cref
0	37814	0	223106	0	161954	0	98121	0	75942	0	20184
1	228101	1	42809	1	103961	1	167794	1	189973	1	245731

	cuarp		desa		expe		htar		inas		injh
0	94224	0	75357	0	21301	0	39569	0	18806	0	89627
1	171691	1	190558	1	244614	1	226346	1	247109	1	121700
										2	54588

	inme		inte		jhog		luze		nban		ncar
0	24089	0	104369	0	12328	0	3442	0	5670	0	178662
1	241826	1	161546	1	36784	1	262473	1	171998	1	87253
				2	216803			2	88247		

	ncel		nhij		niess		nlib		nmic		nvca
0	18706	0	226912	0	115339	0	24574	0	171890	0	230416
1	40443	1	39003	1	150576	1	195070	1	94025	1	35499
2	206776					2	46271				

	ocjh		ocst		pard		pisc		prds		rabu
0	71669	0	185116	0	27309	0	149821	0	45401	0	110096
1	156663	1	80799	1	238606	1	116094	1	220514	1	155819
2	37583										

	sege		sexo		tcla		tesc		tpvi		tvst
0	27030	0	136057	0	34085	0	86193	0	242481	0	139588
1	238885	1	129858	1	231830	1	165563	1	23434	1	126327
						2	14159				

	ucor		vecc		vivy
0	107617	0	152689	0	231798
1	158298	1	113226	1	34117

Técnicas Multivariantes

Es un grupo de modelos estadísticos que analizan en forma simultánea un conjunto de variables y su interacción, de esta forma se obtiene una visión global del fenómeno analizado, logrando mayor perspectiva que con los métodos clásicos univariantes. La utilización de estas técnicas en este estudio permite:

- Revisar los datos representados en un conjunto menor de nuevas variables, transformadas de las originales con mínimas pérdidas de información, estas nuevas variables al ser combinaciones lineales de las originales, son de definición abstracta, que dependen de las variables pri-

marias; utilizando métodos objetivos que definen el número de factores, indicadores o nuevas variables necesarias para describir en forma adecuada una realidad se obtiene una reducción del número de variables y por tanto una simplificación de la data. El modelo estadístico que hace lo antes indicado se conoce como Análisis de Componentes Principales.

- Clasificar a los individuos que de antemano ya forman un grupo (bachilleres) a los cuales se les quiere separar en dos nuevos grupos.
  - Los que ingresan a la universidad pública ecuatoriana.
  - Los que no ingresan a la universidad pública ecuatoriana

Los modelos que nos permiten clasificar a los individuos en este estudio son:

- Regresión Logística
- Análisis Discriminante.
- Máquina de Soporte Vectorial

Reducción de la dimensión de la data

La matriz resultante de la depuración de la data contiene 45 variables explicativas, este número de variables resulta grande para una aplicación adecuada de los métodos clasificatorios, por lo que previamente aplicamos el modelo ACP, el mismo que nos permitirá reducir el número de variables con cierta pérdida de la variabilidad de la data.

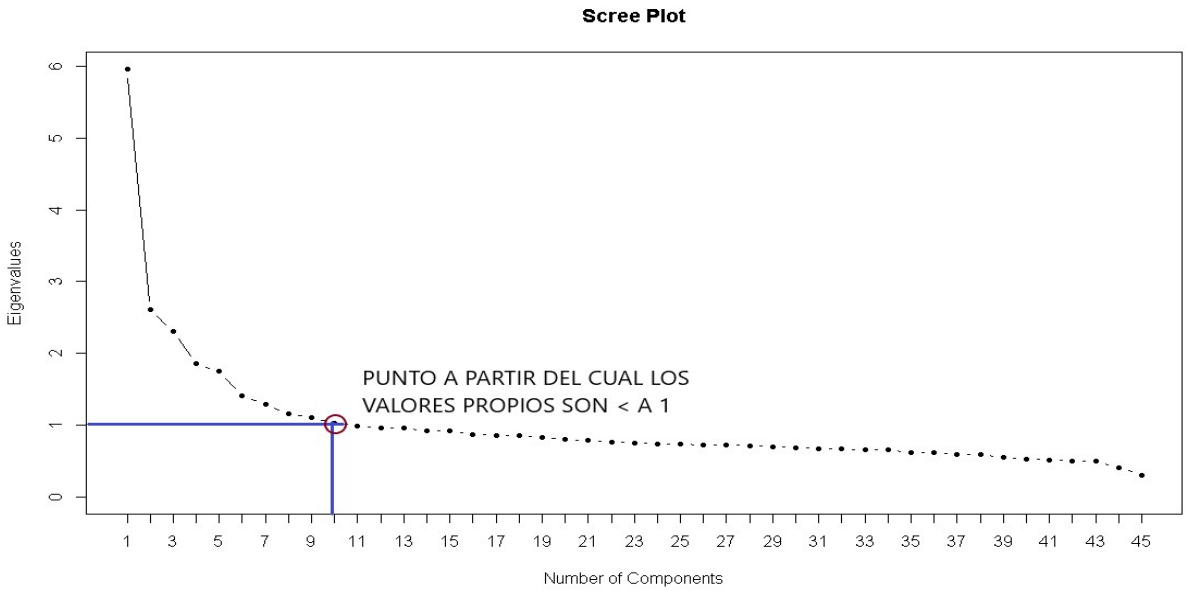
Para la seleccíon del número adecuado de componentes principales que se debe determinar, construimos un gráfico de la magnitud de los valores propios ordenados descendentemente versus su posición llamado Scree Plot, el mismo que lo observamos en la Figura 15

De acuerdo al criterio de que el número de componentes principales debe ser igual al número de valores propios mayores a 1, o de escoger el punto de cambio de pendiente de la recta que une los valores propios resulta que el número de componentes principales de acuerdo a estos criterios será 10.

El porcentaje de la variabilidad total explicada por estos 10 componentes es del 43.70 %, valor bajo que indica que con 10 componentes principales la cantidad de información que se pierde de la data es alta.

Figura 15

Scree Plot: Magnitud de los valores propios vs. su posición



Para mejorar el porcentaje de la variabilidad de la data, que es explicada por las componentes principales se aumenta el número de las mismas, así tememos que para 20 componentes principales el porcentaje de variabilidad explicadas por estas, sube al 63.84 %

Estos valores bajos del porcentaje de variabilidad del sistema indica una baja correlación entre las variables. Los resultados del ACP para 10 y 20 dimensiones se pueden observar en los anexos.

Clasificación de las observaciones de la data

Con los 2 resultados de la reducción de la dimensión de la data, obtenidos en el ACP, es decir con los modelos de 10 y de 20 dimensiones se aplicará los modelos clasificatorios descritos.

Previamente a realizar el proceso de clasificación de las observaciones en las nuevas datas, procedemos a dividir cada una de estas matrices en dos submatrices, la primera llamada de entrenamiento, la misma que contendrá el 90 % de los datos, escogidos aleatoriamente y será la matriz que permita desarrollar el módelo clasificatorio. La segunda matriz, llamada de prueba contiene el restante 10 % de los datos y permite verificar la eficiencia del modelo.

La eficiencia de los modelos clasificatorios se determinará mediante una tabla cruzada, comparando los valores pronosticados por el modelo clasificatorio con los valores reales proporcionados por la data, según lo indica la tabla 6:

Tabla 6

Tabla cruzada entre valores reales y valores pronosticados

TABLA CRUZADA			
	Valores pronosticados		
V. reales	0	1	
0	Especificidad	Falsos positivos	Total fracasos
	V-	F+	reales
1	Falsos negativos	Sensibilidad	Total éxitos
	F-	V+	reales
	Total fracasos p.	Total éxitos p.	TOTAL

Además de los totales parciales de exitos y fracasos tanto para valores reales como para los pronosticados se presentan ciertas características que relacionan los resultados obtenidos con los valores reales, lo que permite determinar la confiabilidad del modelo; estos parámetros se les conoce como:

- **Especificidad**: es el porcentaje las observaciones pronosticadas en el grupo 0 y que realmente pertenecen al grupo 0 respecto al total de observaciones del grupo 0, para este estudio indica el porcentaje los individuos calificados que no acceden a la universidad y en la realidad no lo hacen respecto al total de los que no ingresan.

- **Falsos positivos**: es el porcentaje de las observaciones pronosticadas en el grupo 1 y que realmente pertenecen al grupo 0, respecto al total de observaciones del grupo 0; en este estudio es el porcentaje los individuos calificados como que acceden a la universidad y en realidad no lo hacen, respecto al total de los que no ingresan.

- **Falsos negativos**: es el porcentaje de las observaciones pronosticadas en el grupo 0 cuando realmente pertenecen al grupo 1, respecto al total de observaciones del grupo 1; en este estudio es el porcentaje de los individuos calificados como que no acceden a la universidad y en realidad si lo hacen respecto al total de los que si ingresan.

- **Sensibilidad**: Es el porcentaje de las observaciones pronosticadas en el grupo 1 y que si pertenecen a este grupo, respecto al total de observaciones del grupo 1; en este estudio es el porcentaje de los individuos calificados como que acceden a la universidad y que en realidad lo hacen, respecto al total de los que si ingresan.

- **Aciertos** Un criterio que determina la eficiencia del modelo es el porcentaje de aciertos del mismo, que estará determinado por el porcentaje entre el total de observaciones pronosticadas

0 y que pertenecen al grupo 0, más las observaciones pronosticadas como 1 y que realmente pertenecen al grupo 1, respecto al total de observaciones.

Una vez determinados los resultados de los modelos clasificatorios, se puede determinar las componentes principales que inciden positiva y negativamente en el resultado obtenido, así como los Factores Asociados de mayor relevancia que permiten al bachiller acceder o no a la universidad pública.

Regresión Logística

Para la determinación del modelo de RL aplicamos sobre las matrices de entrenamiento de 10 y 20 componentes, la función glm del paquete MASS, del programa estadístico R. Los resultados se presentan en las tablas 7 y 8

Tabla 7
Resultados de la Regresión Logística para 10 componentes

Coeficientes del modelo de RL para matriz de 10 componentes					
Intercepto:1.4936	D1: 0.6396	D2: 0,4698	D3: 0.0228	D4: -0.5760	D5 0.2759
	D6: 0.0576	D7: 0.1058	D8: 0.1706	D9: -0.0444	D10: -0.1002

Tabla 8
Resultados de la Regresión Logística para 20 componentes

Coeficientes del modelo de RL para matriz de 20 componentes					
Intercepto:1.5035	D1: 0.6411	D2: 0,4763	D3: 0.0193	D4: -0.4771	D5 0.4311
	D6: 0.0434	D7: 0.1155	D8: 0.1605	D9: -0.0744	D10: -0.1114
	D11: 0.0031	D12: 0.0390	D13: 0.0251	D14: 0.0387	D15: 0.0632
	D16: 0.0159	D17: 0.0562	D18: 0.1130	D19: -0.0079	D20: -0.0262

Considerando como criterio de separación de las observaciones un valor para la probabilidad $Pr. = 0.5$ sobre los modelos para 10 y 20 componentes, se obtienen los resultados que se presentan en las tablas 9 y 10

Tabla 9
Resultados RL: Tabla cruzada para 10 componentes y Pr.=0.5

10 dimensiones y Pr.=0.5			
	Valores pronosticados		
V. reales	0	1	
0	9 836	44 348	54 184
	18.15 %	81.85 %	22.64 %
1	8 879	176 261	185 140
	4.80 %	95.20 %	77.36 %
TOTAL	18 715	220 609	239 324

RESUMEN	
Aciertos	77.76 %
Sensibilidad	95.20 %
Especificidad	18.15 %
Falsos positivos	81.85 %
Falsos negativos	4.80 %

Tabla 10

Resultados RL: Tabla cruzada para 20 componentes y Pr.=0.5

20 dimensiones y Pr.=0.5			
	Valores pronosticados		
V. reales	0	1	
0	10 278	43 906	54 184
	18.9 %	81.03 %	22.64 %
1	9 171	175 964	185 140
	4.95 %	95.03 %	77.36 %
TOTAL	19 449	219 875	239 324

RESUMEN	
Aciertos	77.82 %
Sensibilidad	95.03 %
Especificidad	18.90 %
Falsos positivos	81.03 %
Falsos negativos	4.95 %

La RL aplicada sobre los dos modelos, con 10 y 20 dimensiones, bajo las condiciones planteadas, define un elevado porcentaje de falsos positivos, además el porcentage de éxitos pronosticados en ambos casos es alrededor del 92 %, el modelo tiene la tendencia de pronosticar a todas las observaciones positivamente.

Con la finalidad de mejorar los resultados de la RL, se realiza diversos análisis modificando el valor de la probabilidad que clasifica las observaciones, tanto para el modelo con 10 dimensiones como para el de 20 dimensiones. El resumen de estos análisis se presentan en las tablas 11 y 12 Los resultados parciales de estos análisis se presentan en el anexo 3.

Tabla 11

Resumen de RL para el modelo de 10 dimensiones y diversos valores de Pr.

Resultados de la RL para el modelo de 10 dimensiones y diversas probabilidades						
	0.50	0.70	0.73	0.75	0.77	0.80
ACIERTOS	77.76	73.61	71.89	70.45	68.79	65.68
SENSIBILIDAD	95.20	79.39	75.16	71.92	68.39	62.39
ESPECIFICIDAD	18.15	53.80	60.74	65.40	70.17	76.96
FALSOS POSITIVOS	81.85	46.11	39.26	34.53	29.83	23.04
FALSOS NEGATIVOS	4.80	20.61	24.84	28.08	31.61	37.61

Tabla 12

Resumen de RL para el modelo de 20 dimensiones y diversos valores de Pr.

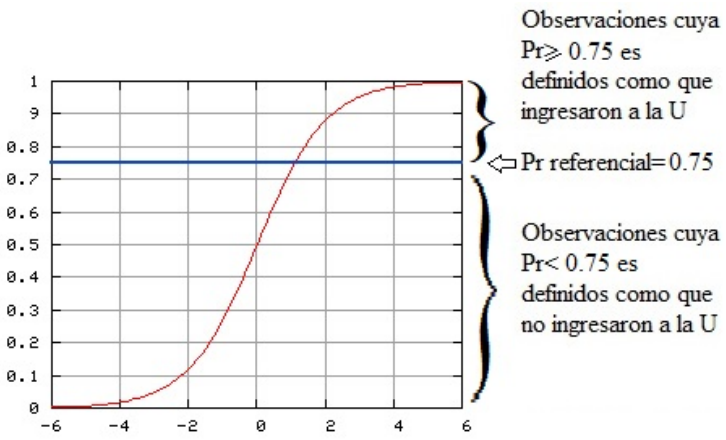
Resultados de la RL para el modelo de 20 dimensiones y diversas probabilidades						
	0.50	0.70	0.73	0.75	0.77	0.80
ACIERTOS	77.82	73.76	72.09	70.67	69.01	65.94
SENSIBILIDAD	95.03	79.37	75.21	72.08	68.60	62.66
ESPECIFICIDAD	18.90	54.60	61.44	65.88	70.45	77.16
FALSOS POSITIVOS	81.03	45.40	38.56	34.12	29.55	22.84
FALSOS NEGATIVOS	4.95	20.63	24.79	27.92	31.40	37.34

Al aplicar la Regresión Logística a las datas con 10 y 20 dimensiones y diferentes valores de

probabilidad referencial, lo primero que se observa es que en los dos casos los valores obtenidos para condiciones similares son muy parecidos. Si buscamos un equilibrio en los porcentajes de falsos positivos y de falsos negativos, el modelo que presenta la mejor opción es el que corresponde a una probabilidad de 0.77, ya que para esta probabilidad estos dos parámetros estan alrededor del 30 %, pero el porcentaje de aciertos es menor al 70 %, para mejorar este parámetro se puede considerar como aceptable la probabilidad referencial de 0.75 que en los dos casos (de 10 y 20 dimensiones) es superior al 70 % y la diferencia entre los porcentajes de falsos positivos y falsos negativos es mínima.

La Regresión Logística utiliza generalmente como probabilidad referencial Pr. = 0.5 como parámetro para clasificar, para este estudio la probabilidad referencial será Pr. = 0.75 que es correspondiente aproximadamente al porcentaje de bachilleres que ingresaron a la Universidad en este estudio en particular, por lo que el esquema de clasificación quedará definido como lo indica la Figura 16.

Figura 16
Criterio de clasificación óptimo para la RL



Una vez definido el criterio de clasificación óptimo para los modelos de 10 y 20 componentes, se procede a la aplicación del modelo de RL a la matriz de prueba, lo que permite determinar la eficiencia de los modelos en el pronóstico de nuevas observaciones. Los resultados conjuntos para los modelos de 10 y 20 componentes con una probabilidad referencial de 0.75 para las matrices de entrenamiento y de prueba se presentan en la tabla 13

Tabla 13

Resultados de RL en las matrices de entrenamiento y de prueba con Pr=0.75

Resultados de la RL aplicada a las matrices de entrenamiento y de prueba				
	Matriz de entrenamiento		Matriz de prueba	
	10 Dimensiones	20 Dimensiones	10 Dimensiones	20 Dimensiones
Aciertos	70.45	70.67	74.67	74.84
Sensibilidad	71.92	72.08	82.21	82.31
Especificidad	65.40	65.88	48.94	49.34
Falsos positivos	34.53	34.12	51.06	50.66
Falsos negativos	28.08	27.92	17.79	17.69

Comparando los resultados de acuerdo al número de dimensiones se observa que para el mismo valor de probabilidad los parámetros de los modelos de 10, y 20 dimensiones tienen valores muy similares y no se encuentra mayor diferencia al aumentar las dimensiones en el cálculo de las componentes principales.

El modelo que presenta la condición más favorable para probabilidad referencial de 0.75, es el de 20 componentes, con mínimas diferencias sobre el de 10 componentes; es necesario en este caso hacer un análisis costo beneficio, considerando el esfuerzo de duplicar el trabajo generando 20 componentes, si tan solo con 10 componentes el resultado es muy similar.

Si se deseara minimizar el error tipo I (disminuir el porcentaje de falsos positivos) se consideraría una probabilidad de 0.77 si por el contrario se desea minimizar el error tipo II (disminuir el porcentaje de falsos negativos) se deberá considerar la probabilidad de 0.73.

Los resultados de la aplicación de RL a las matrices de prueba indican que para las dos matrices, de 10 y 20 dimensiones, los valores de los parámetros que miden la eficiencia del modelo son muy parecidos. El porcentaje de aciertos y la sensibilidad es algo mayor que en la matriz de entrenamiento, pero presenta cierta desventaja en la especificidad, por lo que se puede concluir que el modelo pronostica adecuadamente a nuevas observaciones

Análisis Discriminante

Para la determinación del modelo de AD aplicamos sobre las matrices de entrenamiento de 10 y 20 componentes, la función lda del paquete MASS, del programa estadístico R. Los resultados se presentan en las tablas 14 y 15

Tabla 14
Resultados del AD para 10 componentes

Coeficientes del modelo de AD para matriz de 10 componentes				
D1: 0.6377	D2: 0,5442	D3: 0.0362	D4: -0.5872	D5 0.2200
D6:0.0969	D7: 0.1099	D8:0.1937	D9: -0.0572	D10: -0.0870

Tabla 15
Resultados del AD para 20 componentes

Coeficientes del modelo de AD para la matriz de 20 componentes				
D1: 0.6313	D2: 0.5394	D3: 0.0193	D4: -0.4932	D5 0.3763
D6: 0.0794	D7: 0.1246	D8: 0.1818	D9: -0.0868	D10: -0.0959
D11: 0.0212	D12: 0.0521	D13: 0.0410	D14:0.0161	D15: 0.0653
D16: 0.0140	D17: 0.0577	D18:0.1008	D19: -0.0293	D20: -0.0207

Una vez construido el modelo del AD, se procede a determinar la bondad del mismo ejecutándolo tanto sobre las matrices de entrenamiento como sobre las de prueba. La eficiencia del AD se determina mediante las tablas cruzadas que comparan los valores pronosticados con los valores reales de las observaciones, los resultados de la aplicación del AD a las diversas matrices se presenta en el Anexo 4, mientras que un resumen del mismo se lo puede ver en la tabla **??**

Tabla 16
Resultados del AD en las matrices de entrenamiento y de prueba

Resultados del AD aplicada a las matrices de entrenamiento y de prueba				
	Matriz de entrenamiento		Matriz de prueba	
	10 Dimensiones	20 Dimensiones	10 Dimensiones	20 Dimensiones
Aciertos	77.90	77.79	78.00	77.90
Sensibilidad	94.78	94.67	95.00	94.74
Especificidad	19.21	20.06	20.55	20.98
Falsos positivos	80.79	79.94	79.45	79.02
Falsos negativos	5.22	5.33	5.00	5.26
Pro. como grupo 1	91.72	91.34	91.45	91.15

Comparando los resultados del Análisis Discriminante de acuerdo al número de dimensiones se observa que los parámetros para 10, y 20 dimensiones tienen valores muy similares y no se encuentra ninguna mejora significativa al aumentar las dimensiones en el cálculo de las componentes principales.

Es necesario en este caso hacer un análisis costo beneficio, considerando el esfuerzo de duplicar el trabajo generando 20 componentes y dispersando la información, si tan solo con 10 componentes se obtiene un resultado muy similar.

La causa de que la variación mínima de los resultados entre los modelos de 10 y 20 componentes posiblememte se deba a que los coeficientes del modelo a partir de la onceava dimensión son muy pequeños en comparación con los primeros.

Comparando los resultados obtenidos sobre la matriz de prueba con los obtenidos en la matriz de entrenamiento, estos son muy similares, lo que representa que el modelo predice con la misma eficiencia las nuevas observaciones.

En este método el porcentaje de observaciones pronosticadas como pertenecientes al grupo 1 es superior al 91 %, lo que acompañado con el alto porcentaje de falsos positivos, indica la tendencia de este modelo de pronosticar a toda observación como perteneciente al grupo 1, lo que le convierte en un modelo no adecuado para el proceso de clasificación.

Máquina de Soporte Vectorial

Los resultados de la aplicación del método MSV para la clasificación de las observaciones tanto de las matrices de entrenamiento como la de prueba se presentan en el Anexo 5, el resumen de estos resultados se los puede observar en la tabla 17

Tabla 17
Resultados de la MSV en las matrices de entrenamiento y de prueba

Resultados de la MSV aplicada a las matrices de entrenamiento y de prueba				
	Matriz de entrenamiento		Matriz de prueba	
	10 Dimensiones	20 Dimensiones	10 Dimensiones	20 Dimensiones
Aciertos	78.98	82.14	77.77	77.81
Sensibilidad	98.28	98.26	97.50	95.92
Especificidad	13.04	27.08	10.15	15.90
Falsos positivos	86.96	72.92	89.85	84.10
Falsos negativos	1.72	1.74	2.44	4.08
Pro. como grupo 1	95.72	92.53	95.81	93.25

Comparando los resultados del modelo MSV de acuerdo al número de dimensiones se observa que los parámetros obtenidos para 10, y 20 dimensiones tienen valores muy similares y no se encuentra ninguna mejora significativa al aumentar las dimensiones en el cálculo de las componentes principales.

Los resultados obtenidos en la clasificación de las observaciones para las matrices de entrenamiento y de prueba son muy similares, lo que indica que el modelo pronostica con la misma eficiencia a las observaciones nuevas.

En este modelo al igual que en el AD, el porcentaje de las observaciones pronosticadas que pertenecen al grupo 1 es superior al 95 % para las matrices de 10 dimensiones y alrededor del 93 % para las de 20 dimensiones lo que sumado al alto porcentaje de falsos positivos, el bajo porcentaje de la especificidad indica que el modelo tiene la tendencia de clasificar como pertenecientes al grupo 1 a las observaciones, lo que le convierte en un método poco eficiente en el proceso de clasificar para nuestra data.

Posiblemente por la característica de la data, que no tiene bien definidos los grupos, ya que estos se sobreponen la eficiencia de pronóstico de este modelo no sea aceptable.

Modelo óptimo

Analizados los comportamientos de los métodos de clasificación en la data de este estudio, corresponde seleccionar el más adecuado que proporcione la mayor eficacia en la clasificación de las observaciones.

Comparando los resultados obtenidos al aplicar los modelos clasificatorios sobre las matrices con 10 y 20 componentes se obtienen resultados muy similares a pesar que teoricamente se incrementa la variabilida explicada de la data de 43.76 % con la matriz de 10 componentes al 63.84 % con la matriz de 20 componentes, no existe una mejora significativa en la eficiencia del los modelos clasificatorios, razón por la cuál se considera que lo más aconsejable es considerar la matriz de 10 componentes para representar la data de los factores asociados.

Del análisis de los tres modelos clasificatorios el AD y la MSV presentan la tendencia de clasificar en el grupo 1 a casi la totalidad de las observaciones, por lo que no son adecuados en el proceso de clasificación de la data seleccionada.

El modelo de RL considerando una probabilidad de referencia de 0.75 resulta el más adecuado, ya que a pesar de presentar un porcentaje de aciertos de alrededor del 70 % presenta un equilibrio en los porcentajes de falsos positivos y falsos negativos, además de un adecuado porcentaje de la sensibilidad y especificidad.

La tabla 7 indica los coeficientes que definen el modelo de RL, los valores más altos, positivos o negativos indican las dimensiones más influyentes en este método clasificatorio,

Las dimensiones significativas positiva son:

D1 0.6396

D2 0.4698

D5 0.2759

D8 0.1706

Las dimensiones significativas negativas son:

D4 - 0.5760

D10 -0.1002

Dentro de cada una de estas dimensiones, las variables que caracterizan positivamente la data se detallan en la tabla 18 y las variables de influencia negativa se detallan en la tabla 19

Tabla 18

Variables de influencia positiva en la aprobación del ingreso a la universidad

Variables de influencia positiva	
Dimensión	Variable y significado
D1	inte: tienes conección a internet en tu casa
	pisc: tipo de materiales de los pisos de tu casa
	nban: cuantos baños hay en funcionamiento en tu hogar
	nmic: tienes microondas en tu hogar
	ncel: cuantos celulares con conección a internet hay en tu hogar
D2	jhog: quien es el jefe de tu hogar
	ocst: cuál es tu principal ocupación
D5	agua: tienes agua potable o entubada en tu casa
	basu: tienes servicio recolección de basura en casa
	luze: tienes luz eléctrica en casa
D8	nlib: en tu casa tienes libros

Tabla 19

Variables de influencia negativa en la aprobación del ingreso a la universidad

Variables de influencia negativa	
Dimensión	Variable y significado
D2	nhij: tienes hijos
	vivy: vives con conyuge conviviente o pareja
D4	regi: región natural del Ecuador
	tpvi: en que tipo de vivienda habitas
D10	amoc: Te sientes amenazado por algún compañero

Variables significativas

Las variables significativamente positivas son:

- inte ...Tienes conexión a internet en tu casa
- pisc ...Tipo de materiales de los pisos de tu casa

- nban ...Cuantos de baños hay en funcionamiento en tu hogar
- nmic ...Tienes microondas en tu hogar
- ncel ...Cuantos celulares con conexión a internet hay en tu hogar
- jhog ...Quien es el jefe en tu hogar
- ocst ...Cual es tu principal ocupación
- agua ...Tiene agua potable en tu casa
- basu ...Tiene servicio de recolección de basura en casa
- Luze ...Tiene luz eléctrica en casa
- nlib ...En tu casa tienes libros

Las variables significativamente negativas son:

- nhij ...Tienes hijos
- vivy ...Vives con conyuge conviviente o pareja
- regi ...Región natural del Ecuador
- tpvi ...Tipo de vivienda
- amoc ...Te sientes amenazado por algún compañero

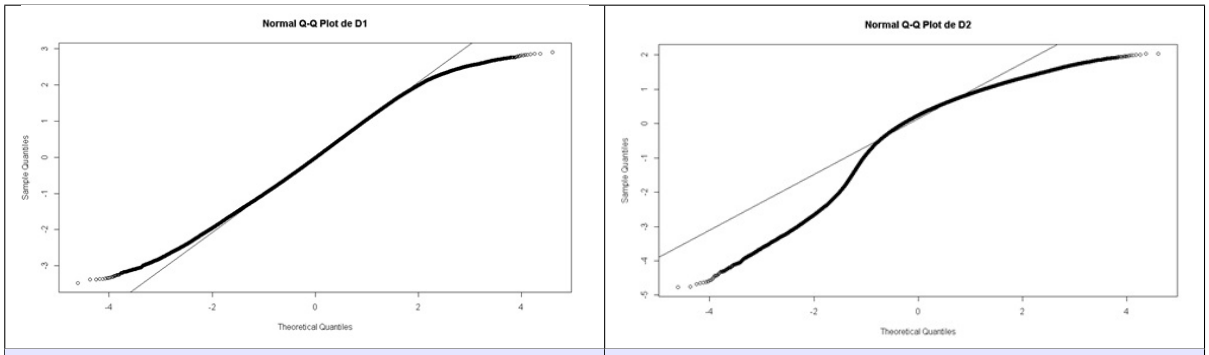
Supuestos de validación

La validación de los modelos está sujeta a varios supuestos, como son la homocedasticidad que indica que la varianza de los residuos es constante, la micronumerosidad se refiere al número de datos adecuado para aplicar el modelo, la presencia de valores atípicos, la normamalidad de las variables y la multicolinealidad la cual existe cuando se presenta una relación entre las variables independientes.

En la aplicación de los modelos de Regresión Logística, Análisis discriminante y Máquina de Vectores de Soporte la micronumerosidad no es problema ya que la data contiene una gran cantidad de observaciones, la multicolinealidad tampoco es problema ya que como variables se utiliza los resultados del Análisis de Componentes Principales lo que implica que las variables son ortogonales

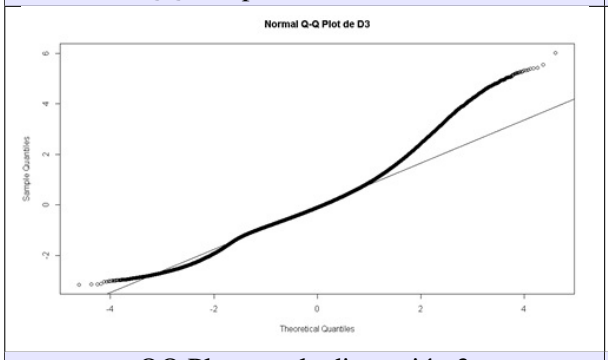
Figura 17

Diagramas QQ Plot para las cuatro primeras componentes principales

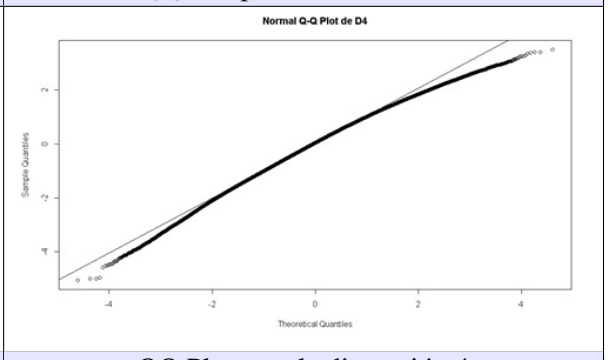


QQ Plot para la dimensión 1

QQ Plot para la dimensión 2



QQ Plot para la dimensión 3



QQ Plot para la dimensión 4

entre si y por tanto no están correlacionadas. Por lo expuesto anteriormente se pondrá énfasis en el estudio de la normalidad de las variables y en el estudio de la homocedasticidad en el modelo de Regresión Logística. Para el análisis de normalidad se procede al análisis de los gráficos qqplot para cada una de las variables los mismos que nos indican que no existe normalidad en estas. A continuación se presenta los gráficos para las primeras cuatro dimensiones.

Adicional al análisis gráfico de la normalidad de las variables explicativas se utiliza mediante el programa estadístico R, la prueba Jarque-Bera, que compara la simetría y la curtosis de la muestra con las de una distribución normal; los resultados de esta prueba indica que no existe distribución normal en las variables explicativas o independientes.

Para la determinación de la homocedasticidad se procede a graficar los errores al cuadrado versus la variable 1, el gráfico resultante no presenta una distribución uniforme de los errores por lo que se concluye la existencia de heterocedasticidad.

Adicionalmente utilizando el programa estadístico R mediante la prueba de Breusch-Pagan se confirma la heterocedasticidad de los errores.

Capítulo 4

Discusión y Resultados

Análisis de la data

El presente trabajo se sustenta en la información proporcionada por el INEVAL, referente al examen Ser Bachiller 2019; la misma está formada por dos matrices:

- Resultados del examen (Notas)
- Encuesta de Factores Asociados

En la matriz Factors Asociados, se eliminan las columnas que no contienen información en al menos el 50 % de las observaciones, y en la matriz Notas se eliminan las columnas duplicadas y las observaciones sin información, para posteriormente unir estas matrices, logrando una data, cuya dimensión es:

$$H = (299015X152)$$

La perfilación prioriza las variables más recurrentes, que se han utilizado en trabajos sobre accesibilidad a la universidad y rendimiento académico del bachillerato, realizados por universidades del continente e instituciones de relevancia mundial, lo que define 45 Factores Asociados significativos, que sumados a las variables de identificación de las observaciones (ID y código) y a la variable dependiente que es la nota promedio de las evaluaciones se completa 48 variables, la matriz resultante de este proceso contiene algunas celdas sin datos validos, por lo que se procede a eliminar las observaciones que poseen algún dato faltante, logrando una nueva matriz, en la cual todas las observaciones tienen información relevante en todas las variables, la dimensión de esta matriz es:

$$H = (265915X48)$$

Los modelos matemáticos predictivos requieren un número menor de variables, que representen el rendimiento académico de los bachilleres, por lo que, los 45 factores asociados que reflejan la tota-

lidad del estudio, es necesario reducirles a un número menor que caracterice el fenómeno estudiado con la menor perdida de información posible. El procedimiento que permite esta reducción es el Análisis de Componentes Principales cuya aplicación habrá reducido la dimensión del problema a costa de cierta perdida de información, así una reducción a:

- 10 dimensiones, explica el 43,86 % de la variabilidad total
- 20 dimensiones, explica el 63,72 % de la variabilidad total

El parámetro establecido por el INEVAL para acceder a la universidad, a partir de una prueba evaluada sobre 1000 puntos, es 700 puntos, esto es:

- Con un puntaje mayor o igual a 700 puntos se accede a la universidad
- Con un puntaje menor a 700 puntos no se accede a la universidad

Modelos de Clasificación

Regresión Logística.

Aplicado el modelo de Regresión Logística, para analizar su capacidad para predecir, se considera inicialmente la probabilidad referencial, Pr. = 0.5, que aplicado a la matriz de entrenamiento obtuvo los resultados descritos en las tablas 9 y 10.

Estos resultados indican que los porcentajes de observaciones asignadas al grupo 1 es alrededor del 92 %, es decir, el modelo bajo este valor de probabilidad referencial, lee en su gran mayoría a todas las observaciones como pertenecientes al grupo 1. Además el alto porcentaje de Falsos positivos, indican que este modelo no es un buen clasificador. Modificando el valor de la probabilidad referencial a 0.75, se obtiene resultados indicados en la tabla 13.

Estos resultados, a pesar de que ciertos parámetros como la sensibilidad y el porcentaje de aciertos han disminuido respecto al modelo anterior, este modelo con probabilidad de 0.75, presenta un equilibrio entre los falsos positivos y falsos negativos; la sensibilidad, la especificidad y el porcentaje de aciertos es aceptable, lo que convierte a este modelo en un buen clasificador.

Los resultados obtenidos al aplicar el modelo sobre la matriz de prueba, indican que este, es un buen clasificador para observaciones nuevas.

Análisis Discriminante.

En el Análisis Discriminante, la información de las variables explicativas o clasificatorias, se resume en la función discriminante, que es la que finalmente genera el proceso clasificatorio al encontrar un punto C de equilibrio, este valor es el elemento discriminante que separa las observaciones que pertenecen a cada grupo.

Los resultados de la aplicación del modelo Análisis Discriminante tanto para las matrices de entrenamiento como para las matrices de prueba se describen en la tabla 16, en esta se observa, que el porcentaje de observaciones asignadas al grupo 1 mediante este modelo, es superior al 91 % para todos los casos, así mismo el porcentaje de falsos positivos, está alrededor del 80 %, estos parámetros indican, que este modelo tiene la tendencia de calificar a casi todas las observaciones como pertenecientes al grupo 1. La especificidad para todos los casos esta alrededor del 19 % lo que significa que la mayoría de las observaciones del grupo cero fueron clasificadas erróneamente como del grupo 1. Estos resultados califican al modelo como mal clasificador.

Comparando los resultados de acuerdo al número de dimensiones se observa que los parámetros de los modelos de 10, y 20 dimensiones tienen valores muy similares y no se encuentra mayor diferencia al aumentar las dimensiones en el cálculo de las componentes principales.

Máquina de Soporte Vectorial (SVM).

Este procedimiento análiza las observaciones en un espacio R^n, siendo n la cantidad de variables que definen el perfil de cada observación, en este espacio R^n el sistema busca un hiperplano que separa y clasifica a las observaciones, si no encuentra este hiperplano, traslada todas las observaciones a espacios de mayor dimensión en donde pueda encontrar el hiperplano que separe eficientemente las observaciones.

Los resultados de la aplicación del modelo Máquina de Soporte Vectorial, a las matrices de entrenamiento y de prueba se presenta en la tabla 17, estos indican que el porcentaje de observaciones pronosticadas como pertenecientes al grupo 1 esta alrededor del 95 % para las matrices de 10 dimensiones, y alrededor del 93 % para las matrices de 20 dimensiones, esto acompaãdo a que el porcentaje de falsos positivos es superior al 80 %, la especificidad esta alrededor del 15 % y que casi no existen falsos negativos indican que el modelo tiene la característica de asignar al grupo 1 a casi todas las observaciones lo que convierte a este modelo en un mal clasificador.

Comparando los valores de los parámetros entre las matrices de 10 y de 20 componentes, estos

son muy similares, lo que se traduce en que es suficiente trabajar con la matriz de 10 componente donde es más facil deterninar las características de estas componentes.

De igual forma no existe mayor diferencia entre los resultados de las matrices de entrenamiento y las de prueba, lo que significa que clasifica con la misma eficiencia a las nuevas observaciones

Selección del mejor modelo clasificador

Comparando los resultados obtenidos al aplicar los modelos clasificadores sobre las matrices de 10 o de 20 componentes, revelan que no existen mayores diferencias, a pesar que el modelo de 10 componentes solo explica el 43,76 % de la variabilidad de la data mientras que la matriz de 20 componentes explica el 63.84 % de esta variabilidad, diferencia que no se traduce en generar un mejor modelo clasificador en cualquiera de las tres técnicas aplicadas, concluyendo que es conveniente usar la matriz con 10 componentes para definir el modelo óptimo.

Definido la mejor dimensión para el análisis clasificatorio, se compara los tres modelos predictivos para definir el que presente las mejores condiciones. La tabla 20 presenta los resultados obtenidos por las tres técnicas clasificatorias, cuando son aplicadas sobre la data con 10 componentes.

Tabla 20
Resultados de las técnicas clasificatorias sobre la data con 10 componentes

Resultados de los modelos clasificatorios						
	Matriz de entrenamiento			Matriz de prueba		
	RL	AD	MSV	RL	AD	MSV
Aciertos	70.45	77.90	78.98	74.67	78.00	77.77
Sensibilidad	71.92	94.28	98.28	82.21	95.00	97.57
Especificidad	65.40	19.21	13.04	48.94	20.55	10.15
Falsos positivos	34.53	80.79	86.95	51.06	79.45	89.85
Falsos negativos	28.08	5.22	1.72	17.79	5.00	2.44
Pro. como grupo 1	63.45	91.72	95.72	75.16	91.45	95.81

En la tabla 20 se observa que el porcentaje total de aciertos con las tres técnicas clasificatorias es muy similar, pero el porcentaje de observaciones clasificadas en el grupo 1 para el AD y MSV estan sobre el 92 %, porcentaje excesivo considerando que la proporción real de observaciones del grupo 1 es del 77 % lo que calificaría a estas técnicas como tendientes a clasificar en el grupo 1 a la mayoría de las observaciones. El parámetro falsos positivos para el AD y MSV están entre el 80 y 90 % lo que indica que a una gran cantidad de observaciones del grupo 0, las clasifica erróneamente en el grupo 1 por tanto la especificidad en estas técnicas es muy baja, entre el 13 y el 20 %. El estudio global de todos los parámetros de la Tabla 20 permite recomendar a la RL como la técnica óptima

para el proceso clasificatorio de la data.

La similitud de los resultados entre las matrices de entrenamiento y de prueba, permite determinar que la eficiencia de la prueba en pronosticar observaciones nuevas es similar a la obtenida con la matriz de entrenamiento que genera los modelos clasificatorios.

Variables significativas

Definido la Regresión Logística como la técnica clasificatoria óptima para esta data, y luego de determinar los coeficientes significativos en el modelo, se determinan las componentes o dimensiones más importantes o influyentes en este proceso de clasificación, y como asociada a cada componente se tiene sus factores asociados mas influyentes, se puede determinar un conjunto de estos factores asociados que influyen en forma positiva y negativa en el proceso clasificatorio, es decir en el ingreso del bachiller a la universidad pública. Estas variables se listan a continuación:

Variables significativamente positivas .

- inte ...Tienes conexión a internet en tu casa
- pisc ...Tipo de materiales de los pisos de tu casa
- nban ...Cuantos de baños hay en funcionamiento en tu hogar
- nmic ...Tienes microondas en tu hogar
- ncel ...Cuantos celulares con conexión a internet hay en tu hogar
- jhog ...Quien es el jefe en tu hogar
- ocst ...Cual es tu principal ocupación
- agua ...Tiene agua potable en tu casa
- basu ...Tiene servicio de recolección de basura en casa
- Luze ...Tiene luz eléctrica en casa
- nlib ...En tu casa tienes libros

Variables significativamente negativas.

- nhij ...Tienes hijos
- vivy ...Vives con conyuge conviviente o pareja
- regi ...Región natural del Ecuador
- tpvi ...Tipo de vivienda
- amoc ...Te sientes amenazado por algún compañero

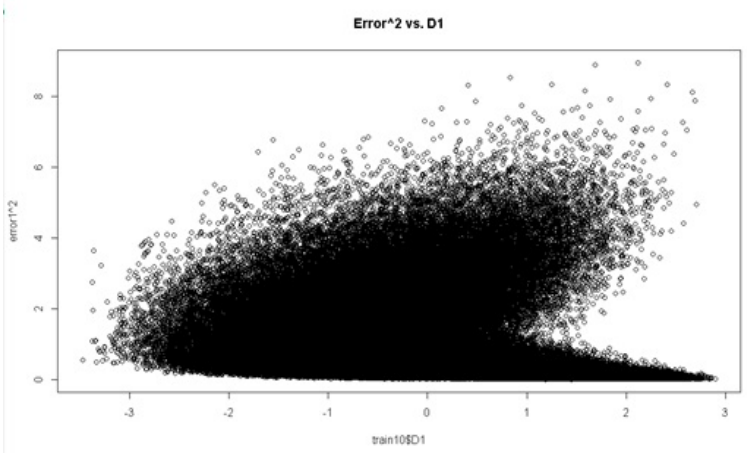
Supuestos de Validación

La prueba Jarque-Bera, que compara la simetría y la curtosis de la muestra con las de una distribución normal; los resultados de esta prueba indica que no existe distribución normal en las variables explicativas o independientes.

Para la determinación de la homocedasticidad se procede a graficar los errores al cuadrado versus la variable 1, el gráfico resultante no presenta una distribución uniforme de los errores por lo que se concluye la existencia de heterocedasticidad como se observa en la Figura 18.

El no cumplimiento de los supuestos de normalidad de las variables, y de hocedasticidad, repercute en la disminución de la confiabilidad de los resultaddos,la misma que en parte se neutraliza, por el número elevado de observaciones (265 915).

Figura 18
Distribución de los errores



Capítulo 5

Conclusiones y Recomendaciones

Conclusiones

De la elaboración del presente estudio se presenta las siguientes conclusiones:

- La data que contiene la encuesta de Factores Asociados contiene muchas columnas con muy poca información, lo que genera que se eliminen aproximadamente el 50 % de estas, entre las cuales estaban variables interesantes.
- La investigación de los procesos clasificatorios, permite conocer un mundo interesante y de actualidad como es el Análisis Multivariante, el mismo tiene repercusiones en diversos campos del conocimiento.
- La proporción de los estudiantes que obtienen un puntaje superior a los 700 puntos es del 77 %, lo que representa que aproximadamente uno de cada cuatro bachilleres no alcanza el nivel mínimo de conocimientos requeridos para una continuación del proceso educativo en el nivel superior.
- El alto porcentaje de bachilleres que no logra el nivel mínimo, indica que existe la necesidad de profundizar los estudios sobre esta problemática para encontrar soluciones al tema, las que al implementarse se traduzcan en un mejor nivel de vida en toda la sociedad.
- En la data estudiada el proceso predictivo que mejor funcionó fue la Regresión Logística, considerando una probabilidad de referencia de 0.75 la misma que tiene una eficiencia del 74 % en la data de prueba.
- Los Factores Asociados que el estudio determina como influencias positivas y negativas para el ingreso del bachiller a la universidad pública, permiten identificar situaciones críticas a las que están sometidos muchos estudiantes y en las que se debe poner atención para profundizarlas si estas son condiciones positivas o prevenirlas si son negativas.

- Esta estudio es producto de las competencias adquiridas en la MEMAT y se alinea con los objetivos planteados, tales como: elevar el nivel académico y científico mediante la aplicación de nuevas metodologías y herramientas tecnológicas; la aplicación de la matemática en el planteamiento y la solución de problemas de la vida real; utilizar programas computacionales acordes a los requerimientos y a las necesidades de la enseñanza actual de las matemáticas con el fin de potenciar la actividad del docente investigador.

Recomendaciones

La educación pública es el cimiento sobre el cual se desarrolla la sociedad, los resultados de las evaluaciones que se realizan en diversos niveles, indican la presencia de falencias en sus diversos etapas; esta situación debe traducirce en la implementación de estudios que detecten los factores que la afectan, para que, desde el estado se implemente políticas públicas tendientes a una continua mejora del sistema educativo nacional

Referencias

[1] ALDAS, JOAQUÍN y URIEL, EZEQUIEL, *Análisis Multivariante Aplicado con R*, Ediciones Paraninfo, Madrid, 2017.

[2] AMAT, JOAQUÍN, *Máquina de Vector Soporte*, Estadistica con R, publicado en: https://github.com/JoaquinAmatRodrigo/Estadistica-con-R/blob/master/PDF_format/34_Maquinas_de_Vector_Soporte _Support_Vector_Machines.pdf, 2017

[3] BAÍLLO, AMPARO y GRANÉ, AUREA, *100 Problemas Resueltos de Estadística Multivariante (Implementados en Matlab)*, Publicaciones Delta, Madrid, 2008.

[4] BALCÁZAR, JOSÉ, *Técnicas, procesos y aplicaciones de la Mineria de Datos*, publicado en: http://www.mavir.net/talks/105-jlbalcazar-feb2012, 2012.

[5] BELTRÁN, BEATRIZ, *Minería de Datos*, Universidad de Puebla, Puebla, 2008.

[6] CANO, JEEN *The V's of Big Data: velocity, volume, value, variety and veracity*, Rethink Maintenance, 2014.

[7] CUADRAS, CARLES, *Nuevos Métodos de Análisis Multivariante*, CMC Editions, Barcelona, 2014.

[8] GARSIDE, WILL y COX, BRIAN, *Big Data Storage*, John Wiley and Sons, Chichester, 2013.

[9] GAUTIER, EMILIO, *Educación de calidad. Comentarios a la nueva propuesta de OREALC/UNESCO*, REICE Revista Iberoamericana sobre calidad eficiencia y cambio en la educación, Madrid, publicado en: https://www.redalyc.org/pdf/551/55130505.pdf, 2007.

[10] HURWITZ, JUDITH et al., *Big Data*, John Wiley and Sons, Inc, Hoboken, 2013.

[11] INEVAL, *Preguntas frecuentes Ineval*, tomado de: https://www.evaluacion.gob.ec/preguntas-frecuentes-ineval-instituto-nacional-de-evaluacion-educativa/, Quito, 2020.

[12] INEVAL, *Historia Ineval*, publicado en: https://www.evaluacion.gob.ec/historia/, 2020.

[13] JAIN, ANIL, *The five Vs of Big Data*, Healthcare Data Analitics, 2016.

[14] JIMÉNEZ, ÁLVARO y ÁLVAREZ, HUGO, *Mineria de datos en la educación*, Legamés, Universidad Carlos III, Madrid, publicado en: https://www.it.uc3m.es/jvillena/irc/practicas/10-11/08mem.pdf, 2010.

[15] MADRID, TITO, *El sistema educativo de Ecuador: un sistema dos mundos*, Universidad Andina Simón Bolivar, Quito, publicado en: https://www.uasb.edu.ec/documents/2005605/2879782/MADRID+TAMAYO +TITO+LIVIO.+El+sistema+educativo+de+Ecuador+un+sistema

[16] MATHWORKS, *Statistics and Machine Learning Toolbox User's Guide*, The MathWorks, Inc, Natick Massachusetts, 2017.

[17] LABORATORIO LATINOAMERICANO DE LA EVALUACIÓN DE LA CALIDAD DE LA EDUCACIÓN, *Informe de resultados TERCE*, Biblioteca digital UNESCO, publicado en: https://unesdoc.unesco.org/ark:/48223/pf0000243532, 2015.

[18] LÓPEZ, JUAN, *Satisfacción laboral del maestro y rendimiento académico en matemáticas de los estudiantes a partir de los factores asociados y las notas utilizando las pruebas Ser Maestro Recategorización y Ser Bachiller 2014*, Escuela Politécnica Nacional, 2015.

[19] OCDE, ORGANIZACIÓN PARA LA COOPERACIÓN Y EL DESARROLLO ECONÓMICO , *El programa PISA de la OCDE que es y para que sirbe*, Pisa for Devolopment, Paris, publicado en: https://www.oecd.org/pisa/39730818.pdf, 2015.

[20] PATEIRO, BEATRIZ, *Introducción a lenguajes avanzados de computación: Matlab en la docencia en Química*, publicado en:http://mathgene.usc.es/matlab-profs-quimica/analisis-datos.pdf, 2016.

[21] PEÑA, DANIEL, *Análisis de Datos Multivariantes*, McGraw-Hill, Madrid, 2008.

[22] PEÑA, DANIEL, *Fundamentos de Estadística*, Alianza Editorial, Madrid, 2008.

[23] PESANTES, TANIA, *Construcción de un índice relacionado con el principio de pertenencia para las universidades nacionales utilizando datos del del proyecto SNIESE 2010 y la metodología de Análisis de Componentes Principales*, Escuela Politécnica Nacional, 2013.

[24] SERRANO, LUIS, *Simplificación de datos con análisis de componentes principales*, publicado en: https://www.youtube.com/watch?v=Gd5qzVpyiHw&abchannel=CTIMFESAcatlan.

[25] TORRES, MARIA JOSE, *Reformas educativas en América Latina, hoy*, II Congreso Internacional de Educación. publicado en: https://otra-educacion.blogspot.com/2016/06/reformas-educativas-en-america-latina.html, 2016

Anexos