MODELO DE CLASIFICACIÓN Y PREDICCIÓN EN DOS ETAPAS: UTILIZANDO ÁRBOLES DE CLASIFICACIÓN Y EL ANÁLISIS DE REGRESIÓN MULTIVARIADA

Por

Yency Edith Choque Dextre

Tesis sometida en cumplimiento parcial de los requerimientos para el grado de

MAESTRÍA EN CIENCIAS

en

MATEMÁTICAS(ESTADÍSTICA)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ

2015

Aprobada por:

Dámaris Santana Morant , PhD. Miembro, Comité Graduado	Fecha
Edgardo Lorenzo Gonzalez, PhD. Miembro, Comité Graduado	Fecha
Edgar Acuña Fernández, PhD. Presidente, Comité Graduado	Fecha
Hilton Alers, PhD. Representante de Estudios Graduados	Fecha
Olgamary Rivera Marrero, PhD. Directora del Departamento	Fecha

Abstract of Disertación Presented to the Graduate School
of the University of Puerto Rico in Partial Fulfillment of the
Requirements for the Degree of Master of Sciences

TWO STAGE CLASSIFICATION AND PREDICTION MODEL USING CLASSIFICATION TRESS AND MULTIVARIATE REGRESSION TREES

By

Yency Edith Choque Dextre

May 2015

Chair: Edgar Acuña Fernández
Major Department: Mathematical Sciences

Currently there exists a great variety of methods and algorithms attempting to optimize the process of classification. However, these methods do not take into account the internal structure of the classification datasets. For this reason, this research work has the goal of developing a classification model in two stages using classification and regression trees (CART) and the multivariate regression trees (MRT). Taking into account also the presence of missing values. This model has been applied to datasets from the National Agrarian University La Molina (Lima-Perú) whithin the Faculty of Economy and Planification of the Department of Statistics and Informatics, with the goal of predicting if a student who is admitted to the university will be able to complete the required curriculum in the alloted timeframe. To develop the proposed model, it was considered the academic performance of the students during their first year of university studies. Considering only those students with an optimum performance, it the missing values were estimated means of two statistical techniques: Multivariate And Regression Trees and the k-Nearest Neighbor Imputation. Then, it was elaborated a statistical model using the CART's

technique, and finally, to validate the proposed model, it was used the methodology of resubstitution and the technique of cross validation. According to our results the first stage can be done automatically using clustering if the academic program does not require many courses with high level of mathematics.

`Abstract_eng.tex`.

Resumen de Disertación Presentado a Escuela Graduada
de la Universidad de Puerto Rico como requisito parcial de los
Requerimientos para el grado de Maestría en Ciencias

MODELO DE CLASIFICACIÓN Y PREDICCIÓN EN DOS ETAPAS: UTILIZANDO ÁRBOLES DE CLASIFICACIÓN Y EL ANÁLISIS DE REGRESIÓN MULTIVARIADA

Por

Yency Edith Choque Dextre

Mayo 2015

Consejero: Edgar Acuña Fernández
Departamento: Matemáticas

En la actualidad existe una gran variedad de métodos y algoritmos que tratan de optimizar el proceso de clasificación. Sin embargo, estos no toman en cuenta la estructura interna que tienen los datos. Por tal motivo, este trabajo de investigación tiene por objetivo desarrollar un modelo de clasificación en dos etapas usando árboles de clasificación (CART) y el análisis de regresión multivariada (MRT). Además se ha tenido en cuenta la presencia de valores perdidos. Este modelo ha sido aplicado a datos de la Universidad Nacional Agraria la Molina (Lima-Perú) del Departamento de Estadística e informática de la Facultad de Economía y Planificación, con el objetivo de predecir si un alumno que ingresa a la universidad terminará su carrera universitaria en los años establecidos por la entidad universitaria. Para elaborar el modelo propuesto, se consideró el rendimiento académico del alumno durante su primer año de estudios universitarios. Con los estudiantes que obtuvieron un rendimiento óptimo se estimaron los datos perdidos mediante dos técnicas estadísticas: El árbol de regresión y clasificación multivariada y la imputación por los k vecinos más cercanos. Seguidamente, se elaboró el modelo estadístico utilizando

la técnica del CART. Finalmente, para validar el modelo propuesto se utilizó la metodología de la resubstitución y la técnica de la validación cruzada.

De acuerdo a nuestros resultados, la primera etapa de clasificación puede desarrollarse automáticamente, utilizando el análisis de cluster cuando el programa de estudios no requiera de un alto nivel académico en los cursos de matemáticas.

`Abstract_esp.tex`.

Copyright © 2015

por

Yency Edith Choque Dextre

A Dios y a mi familia por ser mi brújula en este sendero de la vida.

AGRADECIMIENTOS

A Dios por ser mi cómplice en la realización de mis más anhelados sueños.

Al Dr. Edgar Acuña Fernández, mi asesor de tesis, por sus constantes sugerencias y apoyo en el desarrollo de este trabajo de investigación.

Al Departamento de Matemáticas de la Universidad de Puerto Rico Recinto de Mayagüez, por darme la oportunidad de seguir mis estudios de Máster.

Al Departamento de Estadística e informática de la Universidad Nacional Agraria la Molina por su confianza y motivación para seguir mis estudios universitarios fuera del país.

A todas las personas que de alguna u otra manera contribuyerón e hicieron posible la realización de mi tesis.

A mis hermanos de corazón que compartierón conmigo esta frase, el cual a sido mi fortaleza en este pasaje de mi vida: "Mira que te mando que te esfuerces y seas valiente; no temas ni desmayes, porque Jehová tu Dios estará contigo a donde quiera que tu vayas" (Josué 1:9).

TABLA DE CONTENIDO

pagina

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

CART	Árbol de clasificación y regresión univariada.
MRT	Árbol de clasificación y regresión multivariada.
SS	Suma de cuadrados de los errores.
A_MRT	Suma de cuadrados aditivos.
db_MRT	Suma de cuadrados basado en la distancia.
SS_MRT	Suma de cuadrados de los errores en el árbol de regresión univariado.
LAD_MRT	Suma de desviaciones absolutas alrededor de la media.
SSD	Suma de cuadrados dentro de los grupos.
MCPD	Modelo de clasificación y predicción en dos etapas.
UNALM	Universidad Nacional Agraria la Molina.
ADM	Administración general.
DIF	Cálculo diferencial.
LEN	Lengua.
CAL	Cálculo diferencial.
ECON	Economía general.
INTRO	Introducción a la ciencia de la computación.
MAT_B	Matemática básica.
MAT_C	Matemática para computación.
BASE	Base de datos.
CAL_EST	Cálculo avanzado para estadística.
PROB	Cálculo de probabilidades.
EXPER	Diseños experimentales avanzados.
AP1	Estadística aplicada I.
EST_GEN	Estadística general.
GEST	Gestión de la calidad.
MODLOS	Modelos lineales.
TM1	Técnicas de muestreo I.
TP1	Técnicas de programación I.
MAT_FI	Matemáticas financieras.
AD_RH	Administración de recursos humanos.
CAL_D	Cálculo diferencial.
C_GER	Contabilidad gerencial.
DE	Desarrollo empresarial.
FI3	Finanzas III.
FEP1	Formulación y evaluación de proyectos I.
FEP2	Formulación y evaluación de proyectos II.
INTR_D	Introducción al derecho.
INV_O	Investigación de operaciones.

LEG_TRI	Legislación Tributaria.
LID_ORG	Liderazgo en organizaciones.
MACR1	Macroeconomía I.
MARK	Marketing.
MICR1	Microeconomía I.
NEG_INT	Negociaciones Internacionales.
ORG_MET	Organización y Métodos.
PLA_EST	Planeamiento Estratégico.
SIG	Sistemas de Información Gerencial.
TNI	Técnicas de negociación internacional.
T_AGRIN	Tecnología agroindustrial.
T_AGROP	Tecnología agropecuaria.
AFIN	Auditoría financiera.
FIN_PUB	Finanzas públicas.
TC_D	Teoría del crecimiento y desarrollo.
SEM	Seminario de tesis.
MAT_FIN	Matemáticas financieras.
MAT_BAS	Matemática básica.
EPE	Escuelas del pensamiento económico.
ALG	Algebra lineal.
INTE	Cálculo Integral.
C_COST	Contabilidad de Costos.
C_GEN	Contabilidad General.
C_GER	Contabilidad Gerencial.
D_E	Derecho y Economía.
D_EMP	Desarrollo empresarial.
METRIA	Econometría.
E_AGRA	Economía Agraria.
E_INFO	Economía de la información.
E_REG	Economía de la regulación.
E_RN	Economía de los Recursos Naturales.
E_BIEN	Economía del bienestar.
E_MA	Economía del medio ambiente.
E_AN1	Estadística aplicada a la economía y los negocios I.
EST_GRL	Estadística General.
ESP	Evaluación social de proyectos.
FEP	Formulación y evaluación de proyectos.
HEPC	Historia económica del Perú contemporáneo.
INV_OP	Investigación de operaciones.
MACR2	Macroeconomía II.
MAT_ECO	Matemática para economistas.
MICR1	Microeconomía I.
NEG_IN	Negociaciones internacionales.
TC_D	Teoría crecimiento y desarrollo.
PE	Política económica.

LISTA DE SIMBOLOS

x_{ij}	Representa la i-ésima posición con la j-ésima variable.
x^*	Nueva observacción.
$\bar{x}$	Media del conjunto de observaciones.
x	Mediana.
τ	Función de impureza.
α	Parámetro de complejidad.
$\lvert\tilde{\tau}\rvert$	Número total de nodos terminales en τ o complejidad del árbol.
$R(\tau)$	Representa a la tasa de mala clasificación total.
$R_{\alpha(\tau)}$	Penalizador de un árbol con muchas ramificaciones y nodos terminales.
$D(x_i, x_j)$	Función de distancia.

CAPITULO 1
INTRODUCCIÓN

La educación universitaria se ha caracterizado por ser uno de los pilares más importantes en el ámbito social, cultural y económico de un país, que actualmente está atravesando por un proceso de reforma importante para el desarrollo académico de los estudiantes. En la última década la cantidad de universidades ha tenido una expansión acelerada del 40% en promedio en América Latina y el Caribe, esto según la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura conocida por sus siglas en inglés como UNESCO [1]. En el informe [2], entregado por la UNESCO sobre la educación universitaria se afirma que la deserción en las universidades es causado por tres aspectos importantes :

i) El nivel socioeconómico: Ingresos familiares, expectativas de egreso con relación al mercado laboral, los estudiantes de bajos recursos económicos tienen que trabajar para solventar sus gastos en la universidad, falta de mecanismos para el financiamiento como becas integrales o parciales.

ii) Aspectos académicos: Deficiente preparación previa, desconocimiento de la profesión, nivel educativo de los padres, elección de la carrera, dificultades de adaptación al entorno universitario.

iii) Aspectos personales: Grado de satisfacción de la carrera elegida, falta de aptitudes, aspiraciones y motivaciones personales, madurez emocional, el entorno familiar, peor aún si se vive en un hogar donde existe violencia.

Por otra parte, se revela que solamente en promedio el 15% de los estudiantes que ingresaron a la universidad logra completar de manera satisfactoria sus estudios

universitarios, en los años establecidos por la entidad educativa. Esto ha generado un elevado costo económico, ya que se estima que al año en América Latina y el Caribe se pierden millones de dólares como producto de la deserción de los estudios universitarios. El reporte también manifiesta que el objetivo además de aumentar la cantidad de ingresantes a las universidades debe ser también reducir la deserción en las universidades.

Este desafío exige que las entidades gubernamentales se concentren en la implantación de leyes que exigan a las instituciones académicas cumplir con la misión de brindar una alta calidad educativa de manera eficaz, los cuales se vean reflejados en el alto porcentaje de éxito. Es decir, de estudiantes que culminan sus estudios universitarios en los años establecidos de acuerdo a la profesión que eligieron.

El propósito de este trabajo de investigación es desarrollar un modelo de clasificación y predicción en dos etapas utilizando técnicas estadísticas multivariadas cuyo objetivo es predecir si un estudiante terminará su carrera universitaria en los años establecidos por la entidad educativa basado en su rendimiento académico del primer año de estudios universitarios. Para alcanzar nuestro objetivo utilizaremos el árbol de clasificación y regresión conocido como CART, el cual es útil para analizar datos complejos y el análisis de cluster para agrupar estudiantes. También, usaremos los árboles multivariados de regresión y clasificación (MRT) y la técnica de los k vecinos más cercanos para imputar datos perdidos. La validación cruzada y el método de resubstitución serán usados para validar la metodología propuesta por este trabajo de investigación.

1.1 Motivación

Las técnicas estadísticas de clasificación y predicción tienen un rol muy importante para el análisis estadístico de los datos en las diferentes áreas de investigación [3–6]. Existen una gran variedad de algoritmos y modelos que se han centrado en maximizar la precisión de la clasificación sin tomar en cuenta la estructura que

tienen los datos. Es decir, que para desarrollar sus algoritmos y modelos utilizan los datos como un solo conjunto, lo que implica que desarrollen sus algoritmos en una sola fase. Por ejemplo; el árbol de clasificación y regresión (CART), que mediante divisiones iterativas construye su modelo. El análisis discriminante que utiliza el enfoque de Fisher para encontrar una óptima función discriminante. El análisis de cluster que para formar los grupos utiliza las medidas de proximidad, entre otros.
En este trabajo de investigación nos proponemos desarrollar una metodología de clasificacion y predicción en dónde tomamos en cuenta la estructura que tiene el conjunto de datos con el objetivo de reducir la tasa de mala clasificación y mejorar la precisión. Esta metodología podrá ser aplicada en conjunto de datos tales como: Reclutamiento de personal, cuando una persona postula a un puesto de trabajo pasa por una primera etapa (etapa de pre-selección); si pasa esa etapa entonces será evaluado con entrevistas y pruebas psicológicas y podrá ser clasificado como apto o no apto para el cargo que esta postulando (etapa de selección). En medicina, un paciente con cáncer puede responder de forma positiva o negativa a la quimioterapia (etapa de pre-selección), de los que si pasaron de manera positiva, son llevados a recuperación intensiva en donde finalmente se sabrá si se curó completamente o no (etapa de selección). Existen muchos otros conjunto de datos con este tipo de estructura en donde se podrá utilizar la metodología propuesta es este trabajo de investigación.

1.2 Objetivos

1.2.1 Objetivo General

Desarrollar un modelo de clasificación en dos etapas usando árboles de clasificación (CART) y el análisis de regresión multivariada (MRT). Teniendo en cuenta además la presencia de valores perdidos.

1.2.2 Objetivos específicos

- Desarrollar un modelo de clasificación y predicción en dos etapas con el objetivo de conocer si un estudiante terminará su programa de estudios en los años establecidos por la entidad educativa, basado en su rendimiento académico del primer año de estudios.
- Identificar aquellos alumnos que pasaron a la segunda etapa de clasificación. Es decir, aquellos que obtuvieron un aprovechamiento académico óptimo durante su primer año de estudios en la universidad.
- Estimar datos perdidos en la segunda etapa de clasificación, solamente en aquellos alumnos que pasarón a la segunda etapa de clasificación.

1.3 Resumen de los Capítulos

En el Capítulo 2, se muestran definiciones y conceptos importantes para desarrollar el modelo de clasificación y predicción en dos etapas propuesta en esta tesis de investigación. En la Sección 2.2, se explica de manera detallada la estructura de un árbol de clasificación y predicción, la metodología y la construcción del árbol. En la Sección 2.3 se ilustra a la clasificación no supervisada mediante el análisis de cluster jerárquico y el algoritmo k-means. En la Sección 2.4, se define conceptos importantes acerca de los árboles de clasificación y predicción multivariada. En la Sección 2.5, se describe las técnicas que nos permiten validar un modelo estadístico de clasificación como el método de resubstitución y la técnica de validación cuzada y finalmente en la Sección 2.6, los estudios previos que se han realizado con clasificación en varias etapas.

En el Capítulo 3, se describe la metodología usada en este trabajo de investigación. La Sección 3.2, explica acerca de las características principales que tiene la demanda de postulantes en el Perú, luego en la Sección 3.3 describimos la base de datos de los estudiantes que estudian en la Universidad Nacional Agraria la Molina (UNALM) con la que se trabajo en este proyecto de investigación. En la Sección

3.4, nos explayamos en la manera de como se diseñó el modelo de clasificación y finalmente en las Secciones del 3.5 al 3.8, se explica el procedimiento para la primera etapa de clasificación y el procedimiento de la segunda etapa de clasificación y predicción y también los métodos que validan el modelo propuesto en este trabajo de investigación.

En el Capítulo 4, se describe los resultados que se obtuvieron al aplicar la metodología de clasificación y predicción en dos etapas a la base de datos de la Facultad de Economía y Planificación de la UNALM, la cual se divide en tres departamentos académicos: Estadística e Informática, Economía y Planificación y Gestión Empresarial. Finalmente, en el Capítulo 5, se presenta las conclusiones que se alcanzó aplicando este modelo de clasificación y predicción propuesto por este trabajo de investigación.

CAPITULO 2
ASPECTOS TEÓRICOS

2.1 Introducción

En el presente capítulo presentamos las definiciones, procedimientos y algoritmos de las técnicas estadísticas que se utilizarón para desarrollar el modelo de clasificación y predicción en dos etapas (MCPD). Primero, comenzamos con el árbol de clasificación y regresión conocido como CART, el cual modela utilizando divisiones repetidas y binarias en cada nodo del árbol. Seguidamente, explicamos el análisis de cluster, donde mencionamos el procedimiento usado para formar los grupos (o clusters) y definimos medidas de proximidad. Luego, introducimos los arboles multivariados de regresión (MRT, por sus siglas en inglés) y explicamos definiciones tales como la impureza, el error de predicción y las distancias de disimilaridad que se emplean. También, se explica la imputación para datos perdidos utilizando el k vecino más cercano. Finalmente, nos explayamos en las tecnicas de validación como son: La validación cruzada y el método de resubstitución.

2.2 Árbol de Clasificación y regresión (CART)

Es una metodología estadística para clasificación supervisada con una variable dependiente. Ha sido ampliamente utilizado en la minería de datos para diferentes ámbitos de investigación, tales como en la medicina, biología, agricultura, aprendizaje de máquinas, economía, entre otros. Esto debido a que posee una gran versátilidad al poder diseñar el modelo con distintos tipos de variables predictoras, su objetivo es clasificar una nueva observación a una determinada clase de la variable

respuesta. También es utilizado por su facilidad en la interpretación de los resultados. El CART es robusto frente a valores outliers y se caracteriza por su invariancia cuando se realizan tranformaciones monótonas a las variables predictoras.
El CART, fue desarrollado por Breiman [7]; quien lo diseñó mediante el aprendizaje inductivo, el cual muestra como resultado un árbol de decisión, que está compuesto por un nodo raíz o nodo madre, los nodos hijos, las ramas y los nodos terminales (u hojas). Las particiones utilizadas para construir el árbol son realizadas de forma recursiva y binaria para generar reglas de decisión respecto a una de las variables predictoras, hasta encontrar en su camino un criterio de parada. Breiman [7], señala que el objetivo de esta metodología es encontrar una manera sistemática de predecir a que clase pertenece una nueva observación.

2.2.1 Árbol de clasificación

El análisis de clasificación utilizando árboles, se lleva a cabo cuando la variable de respuesta es categórica con dos o mas niveles, según sea el caso. Además, las variables predictoras pueden ser categóricas o continuas. El procedimiento para diseñar este modelo empieza con la construcción del árbol, seguido de la poda del árbol y finalmente, se elige el árbol óptimo.

2.2.2 Construcción del árbol

La construcción del árbol de clasificación se realiza mediante un particionamiento binario y de manera recursiva para cada nivel del árbol. Empieza dividiendo el nodo madre (o nodo raíz) el cual contiene a todo el conjunto de datos iniciales, en dos nodos hijos, donde cada uno de estos nodos esta compuesto por un subconjunto del conjunto de datos iniciales. El objetivo principal es que cada nodo hijo presente datos que sean lo mas homogéneos posibles dentro de ese nodo y lo mas heterogéneo posible entre nodos, respecto a la variable predictora que esta discriminando. Esta división recursiva se realiza hasta encontrar nodos terminales; esto quiere decir que

en cada nodo los datos sean tan homogéneos que no aya la necesidad de seguir dividiéndolo en nodos hijos. Es importante que para cada división que se realiza con los nodos se identifique un criterio de decisión con las variables predictoras del conjunto de datos, el cual lo encontramos en cada rama del árbol.

El particionamiento de un nodo del árbol tiene como objetivo que la impureza del nodo hijo sea menor que la del nodo madre, esto traerá como consecuencia que la suma de cuadrados de los errores sea menor en el nodo hijo cuando lo comparamos con el nodo madre. La función de impureza esta definido de la siguiente manera:

$$i(\tau) = \phi(p[y = \frac{1}{\tau}]) \tag{2.1}$$

Donde ϕ es una función de impureza, el cual presenta dos características importantes:

i. $\phi \geq 0$

ii. Para cualquier $p \in (0,1)$,$\phi(p) = \phi(1-p)$ y $\phi(0) = \phi(1) < \phi(p)$

Además, se define el grado de reducción de la impureza cuando se pasa del nodo madre hacia los nodos hijos en la división s, como sigue:

$$\triangle I(t) = i(\tau) - P[\tau_L] * i(\tau_L) - P[\tau_R] * i(\tau_R) \tag{2.2}$$

Donde τ es el nodo madre del nodo izquierdo τ_L y del nodo derecho τ_R; y $P[\tau_L]$ y $P[\tau_R]$ las probabilidades de que una observación caiga dentro de los nodos τ_L y τ_R respectivamente. Se elige una división determinada tal que $\triangle I(t)$ sea lo máximo posible.

Las diferentes maneras de impureza (ver figura 2-1) mas usados son:

i. La entropía : $\phi(p) = -p * log(p) - (1-p) * log(1-p)$

ii. El índice de Gini: $\phi(p) = min(p, 1-p)$

iii. Mínimo error, conocido también como el error de Bayes: $\phi(p) = p(1-p)$

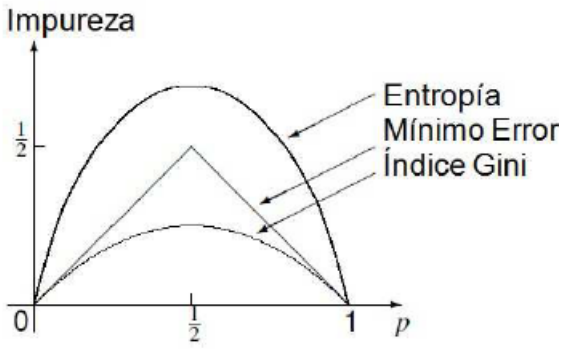

Figure 2–1: Funciones de impureza.

2.2.3 Poda del árbol

Uno de los objetivos principales cuando se elabora y se utiliza el análisis de clasificación mediante árboles, es encontrar un árbol que sea óptimo, eficiente y de fácil interpretabilidad. Para ello es importante observar como es el árbol que se obtuvo finalmente, en cuanto a la calidad que tienen sus nodos terminales y al tamaño del árbol. Si nos topamos con un árbol muy ramificado y complejo, éste puede traer consigo problemas de sobreajuste y por lo tanto se puede realizar la poda del mismo, el cual permite eliminar aquellas ramas del árbol, que son superfluas del árbol original. Cuando se decide llevar a cabo la poda del árbol se debe tomar en cuenta como es la calidad del árbol final (en cuanto a los nodos terminales y el tamaño del árbol final). Cuando se poda el árbol este estará inmerso en la existencia del costo de la mala clasificación, conocido también como costo-complejidad.

Supongamos que el árbol de clasificación construido con el CART es representado por $\tilde{\tau}$, su costo-complejidad se define como:

$$R_\alpha = R(\tau) - \alpha|\tilde{\tau}| \tag{2.3}$$

Donde:

i. $\alpha >= 0$, es el parámetro de complejidad.

ii. $|\tilde{\tau}|$, número total de nodos terminales en τ o también representa a la complejidad del árbol.

iii. $R(\tau)$, representa la tasa de mala clasificación total estimada por la validación cruzada.

Breiman [7], señala que si tenemos un árbol de clasificación τ_0, existe un único subárbol más pequeño que el τ_0 que mínimiza el costo-complejidad, para cualquier párametro de complejidad.

2.2.4 Elección del árbol óptimo

Aplicando el criterio de poda para el árbol de clasificación, esto dará como resultado varios árboles de diferentes tamaños y ramificaciones, el objetivo de este último análisis es encontrar de todos los árboles podados (o sub-árboles) a un árbol óptimo. El árbol será óptimo, cuando de todos los sub-árboles se encuentre un sub-árbol que presente la menor tasa de mala clasificación (o error asociado) en su respectivo proceso de aprendizaje y por la complejidad que el árbol óptimo posea. Para encontrar el árbol óptimo es recomendable utilizar el procedimiento de la validación cruzada, cuyo objetivo será estimar $R(\tau)$.

2.2.5 Árbol de regresión

El árbol de regresión se caracteriza porque su variable respuesta es una variable continua y sus variables predictoras pueden ser categóricas o continuas, al igual que el árbol de clasificación, éste presenta las mismas características para contruir el árbol, podarlo y la elección del árbol óptimo; la diferencia en comparación del árbol de clasificación radica en la manera de escoger la función de impureza que es utilizado cuando se divide un nodo; y el costo-complejidad cuando se decide podar el árbol para obtener el árbol óptimo.

2.3 Análisis de Clúster

El análisis de cluster es una técnica de clasificación estadística que tiene por objetivo conocer las agrupaciones existentes en los datos que estan siendo analizados. Una característica importante de esta técnica es que los grupos no se conocen apriori, por lo que también es conocido como clasificación no supervisada. El análisis de

cluster se plantea fundamentalmente lo siguiente: Dado un conjunto de datos de N observaciones, caracterizados cada uno de ellos con n variables; clasificamos las observaciones (con la información que se tiene de los datos) en grupos que sean lo más similares posibles, para ello es importante conocer las medidas de proximidad que se utilizarán para obtener los grupos o clusters; estas medidas de proximidad van a depender del tipo de variables, se elige el algoritmo de clasificación para determinar la estructura que tiene la agrupación de los datos y finalmente elaboramos un gráfico que nos represente los grupos (clusters) que se han generado.

2.3.1 Definición y procedimiento del análisis

Según Kendall y Buckland [8] definen el término cluster como un grupo de elementos contiguos de una población estadística, por ejemplo un grupo de personas viviendo en una casa, un conjunto de observaciones en una serie ordenada o un conjunto de parcelas adyacentes en un campo. Asimismo, Lang [9] lo define como una técnica estadística multivariante que tiene como objetivo generar grupos lo más homogéneos posibles, los cuales no estan establecidos con anterioridad.

Para Xi y Wunsh [10] el procedimiento (ver figura 2-2) que permite realizar el análisis de cluster consta de cuatro pasos:

1. Extracción o selección de las variables:

   Un punto de vista propuesto por Jain [11] y Bishop [12] nos dice que la elección de la selección de variables va distinguiendo a las variables de un conjunto de observaciones; mientras que la extracción de variables utiliza algunas transformaciones para generar prácticas y nuevas variables.

   Claramente la extracción de variables es potencialmente capaz de producir características que pueden ser de mejor uso en cubrir como es la estructura de los datos. Sin embargo, la extracción puede generar características que no son físicamente interpretables; mientras que la selección de variables asegura la retensión del significado general de las variables seleccionadas.

Ambos, la selección de variables y extracción de variables son muy importantes para la efectividad de las aplicaciones de cluster.

La selección o generación de variables salientes puede mejorar en reducir la gran cantidad de almacenamiento computacional, disminuir en cuanto a costo, simplificar los procesos del diseño y facilitar la interpretación de los datos.

Generalmente las variables ideales deberían usar patrones que se distingan por pertenecer a los diferentes cluster, inmune al ruido, fácil de obtener e interpretar.

La extracción de variables se presenta en el contexto de reducción de dimensionalidad y visualización de los datos.

2. Selección o diseño del algoritmo de cluster:

   Este paso usualmente consiste en determinar una medida de proximidad apropiada para ir contruyendo una función de criterio. Intuitivamente las observaciones son agrupadas dentro de diferentes "clusters" con la mayor similaridad posible y semejanza dentro de cada uno de ellos.

   Prácticamente todos los algoritmos de cluster están explícitamente o implícitamente conectados por alguna definición particular de medida de proximidad.

   Algunos algoritmos trabajan directamente con la matrix de proximidad. Una vez que la medida de proximidad es determinada, el cluster puede ser construido como un problema de optimización mediante una función de criterio. Los clusters obtenidos son dependientes de la selección de la función.

   Los algoritmos de cluster no son universales para resolver todos los problemas. Kleinberg [13] manifiesta que: "A sido una gran reto para desarrollar una estructura unificada pensando acerca de éste (cluster) como en una técnica fuerte y profunda en comparación de otros enfoques".

   Es importante para que se investigue cuidadosamente las características de un problema, seleccionar o diseñar de manera ordenada una apropiada estrategia de cluster.

3. Validación del cluster:

   Dado un conjunto de datos, cada algoritmo puede producir siempre una partición independientemente si existe o no existe realmente una estructura particular en los datos. Por otra parte, diferentes enfoques de agrupamiento usualmente conducen a diferentes clusters de los datos, e incluso para el mismo algoritmo la selección de un parámetro o la orden de presentación de patrones de entrada pueden afectar los resultados finales.

   Por lo tanto, una efectiva evaluación estándar y criterios son realmente importantes para proporcionar a los usuarios un grado de confianza para los resultados de cluster; estas evaluaciones podrían ser objetivas y no tener preferencias por cualquier algoritmo.

   Generalmente hay dos categorías de criterios de validación: Indices externos e indices internos.

   Los indices externos son basados en alguna estructura pre-específicada, que es el reflejo de la información apriori en los datos y es usado como un estándar para la validación de soluciones de cluster (agrupamiento). Las pruebas internas no dependen de la información externa (conocimiento apriori). En lugar de ello, examinan directamente la estructura de los cluster en los datos originales.

4. Interpretación de los resultados:

   El último objetivo del análisis de cluster es que los usuarios puedan obtener un significado profundo de los datos, y por lo tanto resolver efectivamente los problemas encontrados. Anderberg [14] habla del análisis de cluster como: "Una idea para sugerir hipoťesis", el también sugiere que: " Un conjunto de clusters no son un resultado final, pero si es un posible bosquejo".

   Expertos de diferentes campos de investigación son alentados para interpretar los cluster obtenidos integrando para ello evidencia experimental y su dominio en la información del área específica, sin restringir sus análisis y observaciones de cualquier

resultado específico de cluster, que como consecuencia en análisis y experimentos futuros pueden ser requeridos.

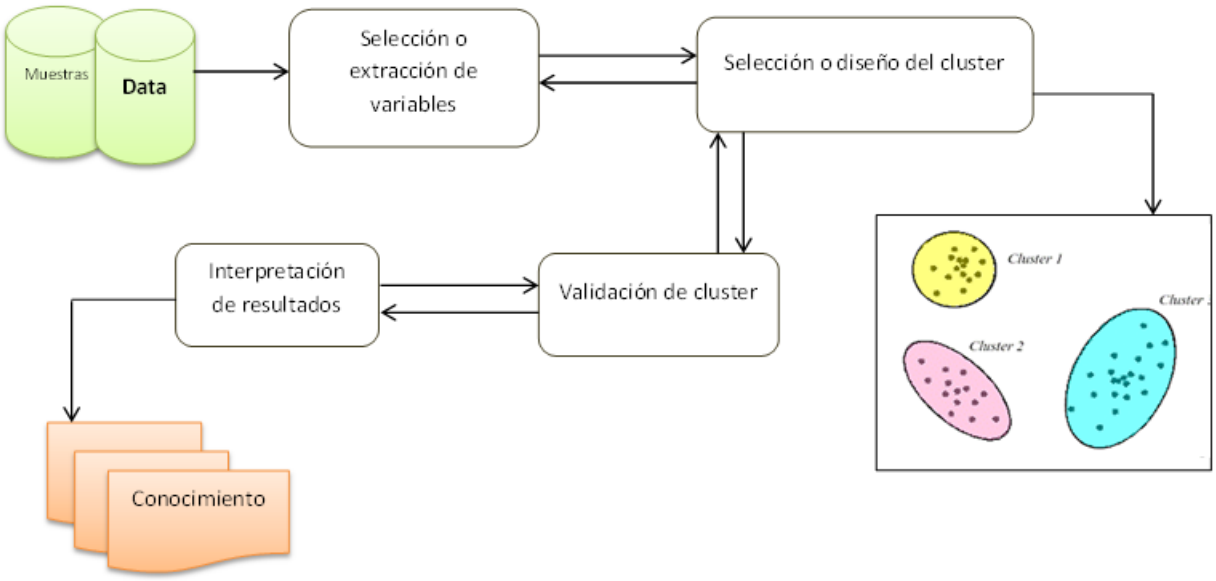

Figure 2–2: Procedimiento del análisis de cluster.

2.3.2 Medidas de proximidad

Los clusters son considerados grupos que contienen objetos de los datos, que son similares dentro de ellos, mientras que entre clusters no lo son. En consecuencia es natural hablar de que tipo de estándar podríamos usar para determinar la cercanía, o como esta medida de distancia (disimilaridad) o similaridad entre un par de objetos, un objeto y un cluster , o un par de clusters. Aquí revisaremos enfoques de medida entre individuos (u observaciones) pero no las que existen entre las variables.

- Niveles de medida y tipos de variables:

  Una observación (o un objeto) esta descrito por un conjunto de características o variables, usualmente representado como un vector multidimensional. Para un conjunto de datos de N observaciones con d variables, sea Nxd es la matriz patrón construida del vector correspondiente, donde cada fila de la matriz denota una observación mientras que cada columna representa a una variable.

Una propiedad de las variables es su nivel de medida, que refleja el significado relativo de números [15]. Los niveles de medida consisten en cuatro escalas: Nominal, ordinal, intervalo y de razón.

- Nominal: Las variables en este nivel son representados con niveles, estados o nombres.
- Ordinal: Las variables en este nivel también son nombres, pero implica cierto orden, sin embargo la diferencia entre sus valores no tiene significado.
- Intervalo: Las variables en este nivel ofrecen una interpretación significativa de la diferencia entre dos valores, sin embargo no existe un verdadero cero y la razón entre estos dos valores no tienen significado. La variable de la temperatura es un ejemplo de este tipo de variable.
- Razón: Este tipo de variable posee todas las propiedades de las variables arriba mencionados, pero si tiene un cero absoluto, además el ratio entre dos valores tiene un significado. Por ejemplo, el pulso es considerado una variable de razón.

- Definición de las medidas de proximidad:

  Proximidad es la generalización de ambos, disimilaridad y similaridad. Una disimilaridad o función de distancia dado un conjunto de datos X debería satisfacer la siguientes condiciones:

  (a) Simetría:

  $$D(x_i, x_j) = D(x_j, x_i) \tag{2.4}$$

  (b) Positividad:

  $$D(x_i, xj) \geq 0, \quad \forall x_i; x_j \tag{2.5}$$

  Si las condiciones de,

  (c) Desigualdad triangular:

  $$D(x_i, x_j) \leq D(x_i, x_k) + D(x_k, x_j), \quad \forall x_i; x_j; x_k \tag{2.6}$$

y

(d) Reflexividad:

$$D(x_i, x_j) = 0; \quad si \quad x_i = x_j \tag{2.7}$$

se mantienen, la medida es llamada una métrica; si solamente se cumple la desigualdad triangular, la medida es denominada semimétrico.

– Medidas de proximidad para variables continuas

Quizás el mas usado comúnmente en medidas de distancia es la distancia Euclideana, también conocido como L_2 norm, representado como:

$$D(x_i, x_j) = (\sum_{l=1}^{d}(|x_{il} - x_{jl}|)^{\frac{1}{2}})^2 \tag{2.8}$$

Donde x_i y x_j son las observaciones de los datos en $d-dimensiones$. Como la distancia euclidiana cumple con todas las condiciones en la ecuación 2.4-2.7, es una métrica. Además investigaciones muestran que la distancia euclidiana tiende a la forma de cluster hiperesférico, estos cluster se caracterizan porque son invariantes a las traslaciones y rotaciones en el espacio. Sin embargo si las características son medibles con unidades que son muy diferentes, las variables con varianzas y valores grandes tienden a dominar sobre las otras variables. Una manera posible de tratar con este problema es la de normalizar la data y hacer que cada variable contribuya equitativamente en la distancia. Comúnmente se usa el método de la estandarización de la data, donde cada variable tiene media 0 y 1 de variancia.

$$x_{il} = \frac{x^*_{il} - m_l}{s_l}, \quad i = 1, 2, ..., N; \quad l = 1, 2, ...d \tag{2.9}$$

donde m_l es la media y s_l la desviación estándar, los cuales se definen como:

$$m_l = \frac{1}{N}.\sum_{i=1}^{N}(x^*_{il}) \tag{2.10}$$

y

$$s_l = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{il}^{*} - m_l)^2} \tag{2.11}$$

respectivamente.

La distancia euclidiana es un caso especial de una familia de medidas, conocido como distancia de Minkowski o L_p norm,definida como:

$$D(x_i, x_j) = (\sum_{i=1}^{d} |x_{il} - x_{jl}|^{\frac{1}{p}})^p \tag{2.12}$$

Note que cuando $p = 2$, la distancia se convierte en la distancia euclidiana.

2.3.3 Análisis de clustering jerárquico

El tipo de clustering jerárquico agrupa los datos con una secuencia de particiones anidadas (clustering jerárquico) o también puede partir de un solo cluster donde este incluye a todas las observaciones (clustering aglomerativo), esta metodología organiza la data dentro de la estructura jerárquica basada en la matriz de proximidad, los resultados son representados por un árbol binario o un dendograma. El nodo raíz de un dendograma representa el conjunto de datos, y cada hoja del nodo es considerada como un subconjunto de los datos. Los nodos intermedios describen el grado de proximidad que hay en las observaciones que se encuentran en ese nodo, la altura del dendograma usualmente expresa la distancia entre cada par de clusters o de un punto de la data y un clúster. Finalmente, los resultados obtenidos con esta metodología pueden ser obtenidos cortando el dendograma en sus diferentes niveles, esta representación proporciona un muy buen análisis descriptivo y de visualización, especialmente cuando la relación jerárquica existe en los datos. El procedimiento general (ver figura 2-3) para desarrollar este tipo de metodología es como sigue:

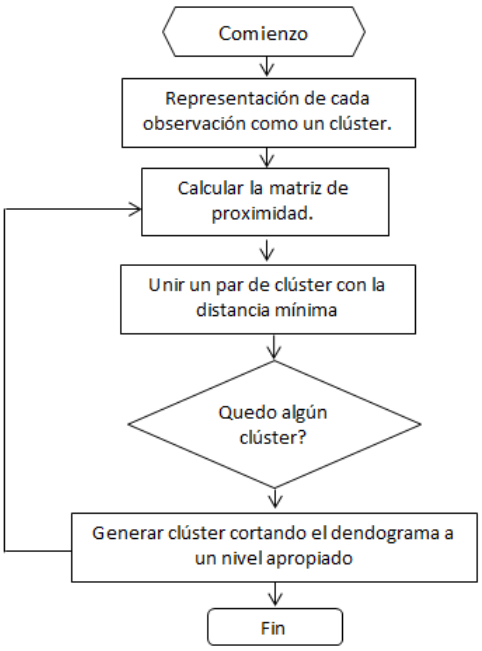

Figure 2–3: Procedimiento del análisis de agrupamiento jerárquico.

1) Empezar con N clusters (Nro. de observaciones). Calcular la matriz de proximidad (usualmente basado en la función de distancia) para los N clusters.
2) En la matriz de proximidad, buscar la distancia mínima $D(.,.) = minD(c_m, c_l)$, donde $D(.,.)$ es la función de distancia, y combina el cluster c_i y c_j para formar un nuevo cluster c_{ij}.
3) Actualizando la matriz de proximidad calculando las distancias entre los cluster c_{ij} y otros clusters.
4) Repetimos los pasos 2 y 3 hasta obtener un solo clúster.

2.4 Árboles de clasificación y Regresión Multivariada (MRT)

Es una técnica estadística multivariada, conocido por sus siglas en ingles como MRT (Multivariate regression trees); fue propuesto por Glenn De'Ath [16], quien lo expuso como una extensión de los árboles de regresión y clasificación (CART). Esta técnica se caracteriza por su versatilidad para trabajar con datos complejos y su modelo a diferencia del CART que tiene solo una variable

respuesta (o dependiente), posee varias; los cuales pueden ser categóricas o numéricas.

Otra peculiaridad de esta técnica es que no hace suposiciones acerca de las relaciones o formas de la distribución de las variables, por lo que se puede trabajar con diferentes tipos de datos; además por su flexibilidad, es usado para predecir explorar y describir un conjunto de datos. Una vez obtenidos los resultados utilizando esta técnica, para interpretarlos es necesario conocer los siguiente aspectos: identificar aquellas variables que estan fuertemente determinadas por las divisiones que se realizan en el árbol, elaborar plots para representar a las medias de los grupos encontrados y por último, encontrar aquellas variables que mejor caracterizan al grupo.

2.4.1 Impureza y error de predicción

Los árboles de regresión se caracterizan por ser modulares en el sentido de que la medida de la impureza, los criterios de división del nodo y el error de predicción son todos independientes de la poda y del crecimiento del árbol [7]. Esto facilita el desarrollo del MRT utilizando medidas de impureza, adicionales a las SS (suma de cuadrados) y medidas que corresponden a la disimilitud de las observaciones. Hay dos estrategias que se adoptan, los aditivos (A-MRT) y otro basado en la distancia (db-MRT).

Añadiendo la medida univariada de la impureza sobre la respuesta multivariada, pueden ser utilizados distintas medidas de suma de cuadrados alrededor de la media (SS-MRT). Por ejemplo, la suma de desviaciones absolutas alrededor de la mediana (LAD-MRT). Estas medidas de impurezas son ejemplos de distancias aditivas [17], que con cada variable respuesta multivariadas contribuyen a la impureza independientemente de otros.

Un aspecto útil del A-MRT es que son robustos a valores atípicos. La impureza del nodo se define como la suma de cuadrados dentro de los grupos, SSD. El

criterio de division maximiza la reducción de la impureza en una división (véase la tabla 2-1).

2.4.2 MRT basado en la distancia

El árbol multivariado se puede formular a partir de una matriz de disimilitud. Para el tratamiento de las disimilaridades mediante las distancias, los clusters se pueden formar con divisiones de los datos sobre los valores que minimicen el SSD (sumas de cuadrados de las distancias) dentro de los grupos.

SS-MRT se basa en la distancia euclidiana, Sin embargo, existen otras medidas que tambien son utilizadas. Por ejemplo, el análisis con la gradiente que depende ya sea explícita o implícitamente de una fuerte relación lineal entre alguna medida de disimilitud y la distancia. Los análisis basados en la distancia euclidiana a menudo fallan cuando se tienen valores altos de las gradientes, en comparación con otras alternativas como la estandarización de Bray-Curtis y disimilitud extendida, que solamente esta débilmente correlacionada con la distancia [18] [19]. Suponiendo que una medida particular ha sido elegido, además si la matriz resultante de disimilitudes de los clusters se trata con distancias, entonces estas distancias representan adecuadamente la estructura en los datos. MRT esta basado en la Distancia (db-MRT) que puede ser definido en términos de SSD dentro de los grupos (o cluster), independiente de la medida de disimilitud elegida para el análisis.

Si las disimilitudes usadas en db-MRT son distancias euclideanas, entonces SS-MRT y db-MRT son exactamente equivalentes, como señalamos anteriormente, minimizando las distancias cuadráticas euclidianas de las posiciones sobre los centroides del nodo que es como minimizar las distancias dentro del nodo entre las posiciones mediante distancias euclídianas cuadráticas. MRT basado en la distancia extiende SS-MRT al igual que el análisis de redundancia basado en la distancia [20] que se extiende al RDA. En ambos casos, los

métodos basados en la distancia permiten que cualquier medida de disimilitud sea utilizado y no solamente la distancia euclidiana.

2.4.3 Comparación de árboles basado en la distancia y la aditividad

A pesar de la coincidencia de los SS-MRT y la euclidiana db-MRT, el A-MRT y db-MRT son muy diferentes. Por ejemplo, la impureza de A-MRT se puede definir de muchas maneras y se centra en la característica típica del nodo; por ejemplo, la media multivariante. Por el contrario, la impureza de db-MRT siempre se define como el SSD entre las posiciones dentro de los grupos (vase tabla 2-1) y la atención se centra en las posiciones individuales con todos los que contribuyen por igual a la impureza de un nodo. Sin embargo existe cierto vínculo entre A-MRT y db-MRT.

- En primer lugar, hay varias medidas de disimilitud que son equivalentes a las distancias euclidianas para datos con una escala adecuada; por ejemplo, chi-cuadrado y disimilitud Chord. Para tales medidas, db-MRT se puede determinar una escala apropiada para los datos y luego usar SS-MRT. Esto no es posible para disimilitudes basadas en sumas de desviaciones absolutas; por ejemplo, la disimilitud Bray-Curtis.
- En segundo lugar, para muchas disimilitudes, un análisis de coordenadas principales generará coordenadas de tal manera que las distancias euclidianas entre las observaciones sean exactamente proporcional a las disimilitudes (es decir, son euclidiana integrable) [17]. El análisis SS-MRT utilizando estas coordenadas como los datos de la respuesta nos dará un árbol idéntico como db-MRT de la disimilaridad original. Para estas dos situaciones, el análisis SS-MRT se puede utilizar para determinar el db-MRT de manera más eficiente; en particular para grandes matrices de

disimilitud.

Table 2–1: Medida de impureza y error de predicción para el árbol de regresión y clasificación multivariada (MRT).

Descripción del árbol	Impureza	Error de predicción
SS-MRT	$\sum_{i,j}(x_{i,j} - \bar{x}_j)^2$	$\sum_j (x^* - \bar{x_j})^2$
LAD-MRT	$\sum \lvert x_{ij} - \tilde{x}_j \rvert$	$\sum_j \lvert x^* - \tilde{x}_j \rvert$
db-MRT	$\sum_{i>k,k} d_{ik}^2$	$\sum_i \frac{(d^*)^2}{n} - \sum_{i>k} \frac{d_{ik}^2}{n^2}$

La notación x_{ij} denota a los datos en la $i - ésima$ posición de la $j - ésima$ variable; x^* es una nueva observación; $\bar{x}$ y x son la media y la mediana respectivamente; d_{ik}^2 y d_i^*2 denota la disimilaridad al cuadrado entre las posiciones i, k entre una nueva observación y posición respectivamente; n representa el número de casos en el grupo del error de predicción.

2.5 Imputación usando k vecinos más cercanos (KNN)

La imputación es un término que denota el procedimiento de reemplazar valores faltantes en un conjunto de datos incompleto, por valores estimados. El objetivo es emplear el conocimiento que se tiene del conjunto de datos para estimar los datos faltantes. El knn es un método para imputar datos faltantes tanto para variables continuas como categóricas, que utiliza la distancia de proximidad, o usualmente la distancia euclideana, para los vecinos más cercanos al dato faltante. Cuando la variable es continua el valor faltante se reemplaza mediante la media entre los vecinos más cercanos, y si la variable es categorica usa el valor mas frecuente entre los vecinos más cercanos.

Una de las ventajas de utilizar el knn es que no hay necesidad de crear un modelo predictivo por cada variable que tiene datos faltantes, porque el knn no construye modelos explícitos. El knn se puede adaptar fácilmente para trabajar con cualquier tipo de datos, simplemente modificando los atributos

que serán considerados en la distancia métrica; éste enfoque puede tratar fácilmente en conjuntos de datos con múltiples valores perdidos. El principal inconveniente de este enfoque es que siempre busca los casos más similares a través de todo el conjunto de datos. Esta limitación puede ser crítica, en el análisis de grandes cantidades de datos faltantes en un conjunto de datos.

2.6 Estimación del error de predicción por validación cruzada

La validación cruzada es un técnica estadística que es utilizada para validar un modelo estadístico. Es un procedimiento iterativo, lo que hace es dividir el conjunto de datos en k partes. Una de las partes, será el conjunto de prueba. Es decir, en este conjunto de datos se aplica el modelo obtenido con las k-1 partes restantes (conjunto de entrenamiento).

Lo ideal sería que si tuvieramos suficientes datos, tendríamos que separar una parte de los datos para la validación y utilizarlo para evaluar el desempeño de nuestro modelo de predicción; puesto que los datos a menudo son escasos, eso no puede ser posible. La validación cruzada usando $k - grupos$ utiliza parte de los datos disponibles para ajustar (entrenamiento) el modelo y la otra parte para ponerlo a prueba (validación).

La metodología de la validación cruzada se realiza de la siguiente manera:

(a) Dividimos los datos en k partes, aproximadamente del mismo tamaño. Por ejemplo cuando $k = 5$, como se muestra en la figura 2-4.

(b) Encontrar un modelo ajustado con las otras $k - 1$ partes de la data.

(c) Luego calcular el error de predicción; cuando se realiza la predicción de la kth $(k = 3)$ parte de la data.

Con $k = 10$, el estimador de la validación cruzada reduce de manera óptima el error de predicción.

Cuando $k = 5$, la validación tiene menor variancia, pero mayor es el sesgo, y

1	2	3	4	5
Entrenamiento	Entrenamiento	Validación	Entrenamiento	Entrenamiento

Figure 2–4: Validación cruzada cuando $k = 5$ grupos.

este podría ser un problema dependiendo de como el rendimiento del aprendizaje varía con el tamaño del conjunto de entrenamiento.

2.7 Estimación del error por resubstitución

Es una metodología estadística que se utiliza para validar modelos estadísticos; por ejemplo, supongamos que queremos validar un modelo de clasificación que se diseñó utilizando el árbol de clasificación. el procedimiento de esta técnica sería de la siguiente manera:

i. Se contruye el modelo estadístico, mediante el árbol de clasificación.

ii. Se utiliza el modelo del paso i para estimar los valores de la variables respuesta.

iii. Con la columna de la variable respuesta estimada, del paso ii, y la columna original se compara cada fila para saber si existió una predicción óptima de la variable respuesta.

iv. De la comparacion realizada del paso iii se calcula la tasa de mala clasificación.

Si ambos valores coinciden, el elemento es clasificado correctamente por el árbol, en caso contrario, se presenta un error de mala clasificación.

El error de resubstitución de T es la proporción de ejemplos de I que T clasifica incorrectamente.

2.8 Estudios previos con clasificación en varias etapas

Recientemente, Soma(2014) [21] propuso un algoritmo de decisión en múltiples etapas para generar reglas de decisión con los datos de retención de estudiantes. Soma trata de predecir que estudiantes están propensos abandonar sus estudios

acádemicos.

Soma [21] utilizó una metodología de clasificación en tres etapas, donde se aplicaron diferentes técnicas de clasificación estadísticas, tales como: el análisis de cluster, k-means, árboles de clasificación y regresión, entre otros. La metodología de clasificación y predicción fue realizada de la siguiente manera:

1.En la primera fase, las reglas se generaron utilizando diferentes algoritmos de árboles de decisión y diferentes métodos implementados por otros investigadores como Digangi(2010) [22]. Las reglas generadas como simples o complejas fueron utilizados en situaciones reales, donde se observó que no se cumplían las condiciones necesarias y suficientes, entonces las reglas fueron generadas a partir de un árbol de decisión controlada, en donde se observarón anomalías para algunas reglas, por lo que el estudio se extendió el uso de dos algoritmos diferentes: el árbol de decisión y a las reglas de asociación.

2.En la segunda fase, se utilizaron dos algoritmos diferentes: la metodología de múltiples etapas y el árbol de decisión, para la generación de las reglas. Los resultados utilizando el método de múltiples etapas generaron reglas más precisas en comparación de los empleados por el árbol de decisión.

3.En la tercera etapa, se amplió el estudio para verificar si el método realizado en la segunda etapa era eficiente, mediante la aplicación del método a diferentes conjuntos de datos, en éste análisis, surgió la siguiente inquietud: dependerá está generación de reglas sobre las características de los datos, para ello la técnica de generación de reglas se modificó para facilitar a los datos con dimensiones mas pequeñas o mas grandes. Finalmente, se encontró que los conjuntos de datos que tenían menos del diez por ciento de los datos en sus clases minoritarias no generaron reglas o estaban por debajo del umbral de precisión.

Van Lam,et.al [23], propusierón un modelo de clasificación en dos etapas para detectar páginas web maliciosas. Para alcanzar sus objetivos usan métodos conocidos en ciencia de la computación como el análisis de futuras páginas web potencialmente maliciosas, el cual puede distinguir las páginas web maliciosas y las categoriza dentro de dos tipos: las características estáticas y las características de tiempo de ejecución. La primera extrae los contenidos y las propiedades de las páginas web, sin tener la necesidad de un procesamiento total o la ejecución de la página web. La segunda consiste en extraer las características mediante la prestación de páginas web de forma completa y ejecutarlas en sistemas específicos.

La metodología para detectar páginas web maliciosas es interesante, porque elige a un conjunto o lista de URLs necesarios para ser inspeccionados y enviados hacia la extracción de características estáticas, el cual solamente extrae algunas web potenciales futuras mediante las características estáticas.

La metodología de clasificación es como sigue:

En la primera etapa de clasificación utiliza la extracción mediante las características estáticas para poder estimar las páginas web que se encuentran dentro de dos grupos (potencialmente maliciosas o beningnas). Solamente las páginas que son consideradas potencialmente maliciosas se envía a las páginas web con las carácterísticas de tiempo de ejecución, a continuación, se ejecuta para un sistema específico, donde serán monitoreados y capturadas durante este proceso.

En la segunda etapa de clasificación, aplican "Las características de tiempo de ejecución" y las "Características estáticas" o ambos, e identifican solamente el comportamiento malicioso. Una URL que se clasifica como benigna en la primera etapa de la clasificación, se le etiqueta como benigna. Una URL que está clasificado como malicioso en la segunda etapa de clasificación se etiqueta como malicioso.

2.9 Modelo de clasificación y predicción en dos etapas (MCPD)

El diagrama de flujo mostrado en la figura 2-5, nos muestra la metodología de clasificación y predicción en dos etapas (MCPD) propuesta por este trabajo de investigación. Supongamos que disponemos de un conjunto de datos en donde se toma en cuenta la estructura que éste presenta, como por ejemplo: datos para reclutamiento de personal, cuando una persona postula a un puesto de trabajo pasa por una primera etapa (etapa de pre-selección); si pasa esa etapa entonces será evaluado con entrevistas y pruebas psicológicas y podrá ser clasificado como apto o no apto para el cargo que esta postulando (etapa de selección). En medicina, un paciente con cáncer puede responder de forma positiva o negativa a la quimioterapia (etapa de pre-selección), de los que si pasaron de manera positiva, son llevados a recuperación intensiva en donde finalmente se sabrá si se curó completamente o no (etapa de selección); entre otros tipos de datos.

El modelo de clasificación y predicción en dos etapas, consiste en trabajar de la siguiente manera:

En la primera etapa: Se forman clusters (o grupos), utilizando para ello al análisis de cluster jerárquico o el método determinístico. Aquellas observaciones que cumplan con pertenecer al grupo de interés pasaran a la siguiente etapa de clasificación; de lo contrario esa observación no será de utilidad para el modelo.

En la segunda etapa: De las observaciones que pasaron a la segunda etapa de clasificación se verifica si existen valores faltantes (missing values). Si hay valores faltantes entonces estos valores seran imputados utilizando los árboles de regresión multivariada (MRT) o con la imputación de los k vecinos más cercanos.

Con los datos originales y los datos imputados en la segunda etapa de clasificación (con el MRT o el KNN) se aplica el árbol de clasificación y regresión (CART). Luego, obtenido la estimación de la clase a la que pertenece la variable respuesta,

este resultado se compara con la clase original y se calcula la tasa de mala clasificación. Finalmente, El MCPD se valida utilizando la técnica de la validación cruzada y el método de resubstitución teniendo como base la tasa de mala clasificación. Se comparan las tasas de mala clasificación obtenidos con los datos originales, y con los datos imputados durante la segunda etapa de clasificación.

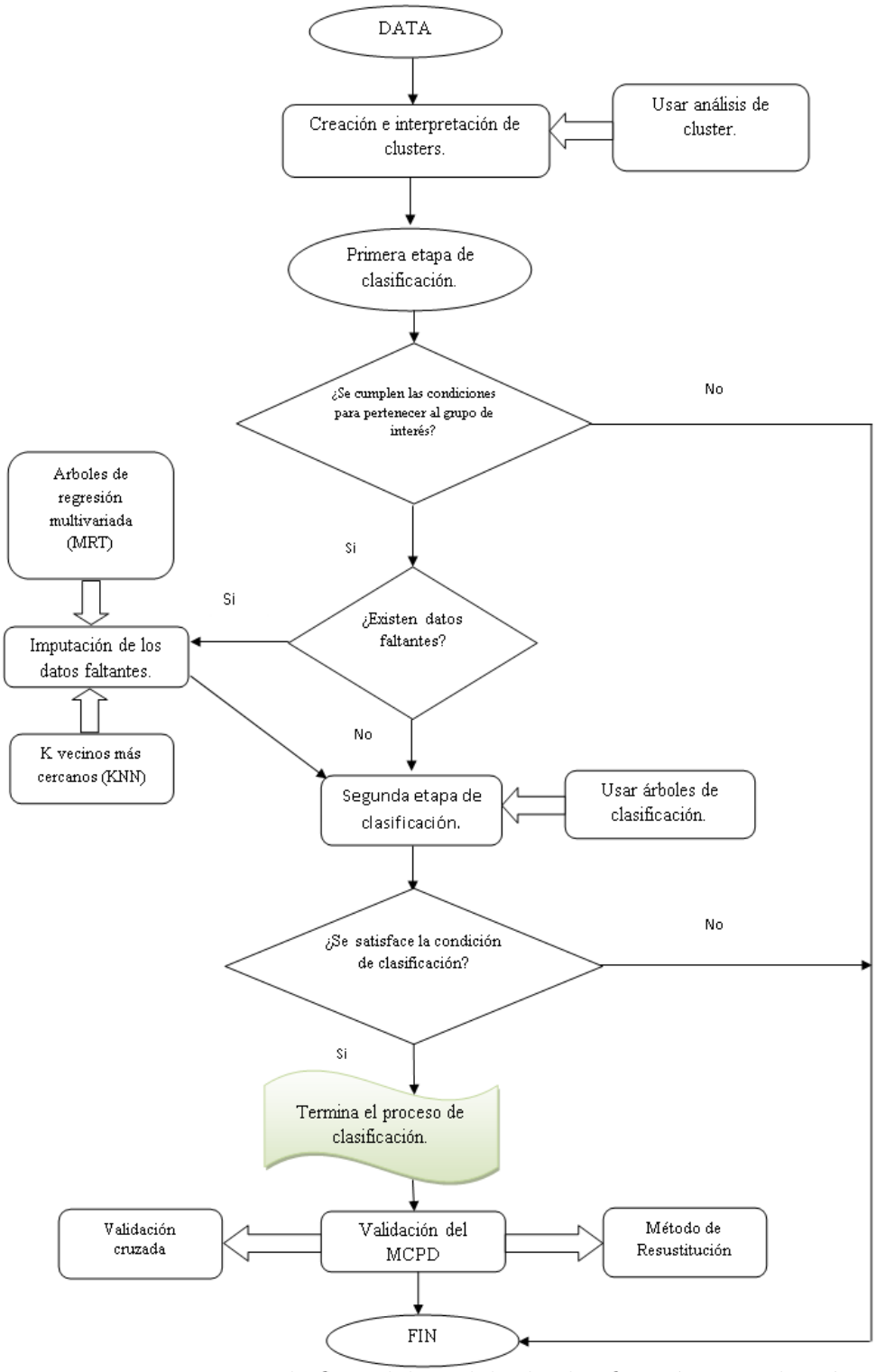

Figure 2–5: Diagrama de flujo del modelo de clasificación y predicción en dos etapas (MCPD).

CAPITULO 3
METODOLOGÍA

3.1 Introducción

El propósito de este trabajo de investigación es desarrollar un modelo de clasificación y predicción en dos estapas (MCPD) utilizando técnicas multivariadas como el MRT, CART y el Análisis de cluster. Seguidamente, validamos el modelo utilizando metodologías tales como la validación cruzada y el método de resubstitución. Las técnicas de clasificación supervizada y no supervizada definidas en el capítulo anterior trabajan con algoritmos. Por ejemplo, para realizar el análisis de Clúster utilizaremos el clustering jerárquico para desarrollarlo utilizaremos el procedimiento que se muestra en la Figura 2-3. En este capítulo preciso de explicar el perfil y la base de datos de los estudiantes universitarios evaluados en el estudio. Luego se explica el procedimiento de la construcción del MCPD y como son utilizadas las técnicas de validación del modelo detalladas en el Capítulo 2. Finalmente, para contruir el modelo de clasificación y predicción en dos etapas (MCPD) se elaboró el esquema de la Figura 3-1, el cual resume la metodología propuesta en este trabajo de investigación.

3.2 Perfil de los estudiantes universitarios en el Perú

La educación universitaria en el Perú atraviesa por un proceso de reforma importante para el desarrollo académico de los estudiantes, ello se debe a que en los últimos trece años la cantidad de universidades privadas se ha duplicado, generando preocupación por parte del Ministerio de Educación y demás entidades gubernamentales. Estas están trabajando en la implantación de leyes que exigan a las

instituciones académicas cumplir con la misión de brindar una alta calidad educativa de manera eficaz que se ven reflejados en el alto porcentaje de éxito. Es decir, del mínimo de estudiantes que culminan sus estudios universitarios en los cinco años establecidos por la ley universitaria en el Perú.

Sin embargo, la misión de las entidades universitarias son cuestionadas por el aumento en el "Índice de deserción de estudiantes" en donde se visualiza una disminución en la tasa de éxito, ya que según el Instituto Nacional de Estadiśtica e informática del Perú, conocido por sus siglas como INEI, menciona que de cada 10 alumnos que ingresan a la universidad solamente 5 de ellos logran culminar sus estudios universitarios, notando que el 50 por ciento de los alumnos desertan.

Todo ese panorama ha generado que las universidades pongan en marcha una serie de planteamientos y proyectos internos para identificar cuales son las principales deficiencias académicas, económicas, físicas y mentales de los estudiantes que han ingresado a la universidad. Para el caso de la Universidad Nacional Agraria la Molina (UNALM) se identificarón las siguientes razones: falta de interés del estudiante por su carrera, el estudiante ingreso a su carrera como segunda o tercera opción, fracaso en su aprovechamiento académico durante su primer año de estudios, ingresan a la carrera solamente con la intensión de trasladarse a otra, entre otros.

El propósito de este trabajo de investigación es desarrollar un modelo de clasificación y predicción en dos etapas que calcule la probabilidad de que un estudiante termine su carrera universitaria en los cinco años, tiempo que dura la curricula de estudios. Para alcanzar los objetivos usaremos el árbol de clasificación y regresión, los árboles multivariados de regresión y clasificación, el análisis de cluster. Así como, la validación cruzada y el método de resubstitición, estos dos últimos nos permitirán validar la metodología propuesta en este trabajo de investigación.

3.3 Base de datos de la Facultad de Economía y Planificación de la UNALM

Para elaborar el MCPD se trabajó con la base de datos de la Facultad de Economía y Planificación de la UNALM que consta de tres departamentos: el Departamento de Estadística e Informática, el Departamento de Economía y Planificación y el Departamento de Gestión Empresarial. Esta base de datos contiene toda la información del alumno, desde que ingresoó a la UNALM hasta que culminó sus estudios universitarios (donde obtuvo su grado académico). Además, para desarrollar este trabajó de investigación las variables con las notas en los cursos que tomó el alumno en su etapa académica, tomando en cuenta que cada departamento tiene sus cursos de concentración. En la Tabla 3-1, se muestra el número total de alumnos y la cantidad por departamentos, el cual representa a la totalidad de los alumnos que egresaron de la UNALM desde 2010-I hasta el 2013-I.

Table 3–1: Base de datos de la Facultad de Economía y Planificación.

Facultad de Economía y Planificación	
Departamento	Número de alumnos
Departamento de Estadística.	176
Departamento de Economía.	308
Departamento de Gestión empresarial.	416
Total.	900

Table 3–2: Cursos analizados durante la primera etapa de clasificación para el departamento de Estadística e informática.

Departamento de Estadística e Informática
Administración General.
Introducción a la Ciencia de la Computación.
Matemática Básica.
Cálculo Diferencial.
Economía General.
Matemática para Computación.

En la UNALM los cursos que toman los alumnos durante su primer año son denominados cursos generales[1] con excepciones de uno o dos cursos que son de concentración; estos cursos son importantes porque son pre-requisitos para que el alumno pueda llevar los cursos de concentración de acuerdo a su profesión . Los cursos generales en la facultad son los de Administración General, Cálculo Diferencial y Matemática Básica. En la Tabla 3-2 se pueden observar los cursos que se consideran en la primera etapa de nuestro estudio. Los cursos de concentración en el Departamento de Estadística e Informática son Introducción a la Ciencia de la Computación y Matemática para la Computación. En el departamento de Economía y Planificación (Ver Tabla 3-3) son Escuela del Pensamiento Económico y Matemática Financiera. Finalmente, en el Departamento de Gestión Empresarial (Ver Tabla 3-4) los cursos de concentración son Administración de Recursos Humanos y Matemática Financiera.

Table 3–3: Cursos analizados durante la primera etapa de clasificación para el departamento de Economía y Planificación

Departamento de Economía y Planificación
Administración General.
Cálculo Diferencial.
Economía General.
Escuelas del Pensamiento Económico
Matemática Básica.
Matemática Financiera.

3.4 Modelo de clasificación y predicción en dos etapas (MCPD) aplicado a la base de datos de la UNALM

Para desarrollar el modelo de clasificación y predicción en dos etapas se elaboró el esquema de la Figura 3-1, que se explica detalladamente de la siguiente manera:

[1] Los cursos generales se refieren aquellas materias académicas que son dictadas en los tres departamentos de la Facultad de Economía y Planificación de la UNALM.

Table 3–4: Cursos analizados durante la primera etapa de clasificación para el departamento de Gestión Empresarial

Departamento de Gestión Empresarial
Administración General.
Administración de Recursos Humanos.
Cálculo Diferencial.
Economía General.
Matemática Básica.
Matemática financiera.

- La primera etapa de clasificación tiene como objetivo principal generar dos grupos (o cluster, según sea el caso); los cuales se elaborarón de manera determinística y utilizando la técnica del análisis de cluster. Para ello, se evaluó el rendimiento académico del alumno que denominamos "el aprovechamiento académico del alumno durante su primer año de estudios", tomándo como variables los cursos que el alumno llevó en este primer año. Los grupos encontrados se etiquetaron como sigue: los alumnos que obtuvieron un rendimiento académico óptimo y los que no obtuvieron un aprovechamiento óptimo.
- La segunda etapa de clasificación tiene como objetivo elaborar un modelo de clasificación y predicción con aquellos alumnos que obtuvieron un aprovechamiento académico óptimo durante la primera etapa de clasificación. Para lo cual, primero se observó si existía la presencia de datos faltantes (missing values), los cuales se trabajaron utilizando la técnica estadística del árbol de regresión y clasificación multivariada (MRT), seguidamente se elaboró el modelo de clasificación y predicción usando la técnica del CART. Finalmente, se aplicó el método de resubstitución y la validación cruzada con k=10 grupos al MCPD, con el objetivo de comparar la tasa de mala clasificación obtenido por ambas metodologías.

3.5 Procedimiento de la primera etapa

Para empezar la primera etapa de clasificación primero se identificó las variables con las que se iban trabajar, estas variables representadas por la nota de los cursos, que están por departamentos. Para el Departamento de Estadística e Informática (Ver Tabla 3-1); el Departamento de Economía y Planificación (Ver Tabla 3-3); el Departamento de Gestión Empresarial (Ver Tabla 3-4). Además, cada observación representa la nota que obtuvo el alumno en alguno de estos cursos. En el Perú, esta nota va del rango de cero a veinte puntos, si el alumno tuvo una nota menor a once; entonces, tendrá que repetir el curso, pero si esta en el intervalo de once a veinte entonces este aprobará el curso.

Como el objetivo de esta primera etapa de clasificación es encontrar dos grupos (o cluster), se realizaron de dos formas diferentes:

1. Clasificación determinística
2. Clustering jerárquico.

3.5.1 Clasificación determinística

Para construir los dos grupos con la metodología determinística se calculó el promedio aritmético de las seis variables (por Departamentos) para cada fila (ver Tabla 3-5), en donde se obtuvo el promedio del rendimiento académico de cada alumno en su primer año de estudios, seguidamente encontramos dos grupos los cuales etiquetamos como aquellos alumnos que obtuvieron un aprovechamiento académico óptimo (los que sacaron un promedio mayor o igual a once), y los alumnos que no obtuvieron un aprovechamiento óptimo (con notas menores o iguales a once).

3.5.2 Clustering jerárquico

Para encontrar los dos clusters (grupos) durante esta primera etapa de clasificación se utlizó el procedimiento para este tipo de análisis de cluster, el cual se puede visualizar en la Figura 2-3:

Table 3–5: Conjunto de datos y variables utilizadas en la primera etapa de clasificación

ESTUDIANTE	ADM	ADGEN	DIF	LEN	MATE	FIN	APROV
E1	15	14	12	13	14	15	SI
E2	16	16	14	17	11	13	SI
E3	13	15	20	17	11	17	SI
E4	14	13	15	14	14	12	SI
E5	14	15	13	13	15	13	SI
E6	14	12	10	14	9	6	NO
E7	12	14	3	11	13	10	NO
E8	11	12	4	12	13	8	NO
E9	11	9	6	8	10	7	NO
E10	10	9	8	19	10	9	NO

3.6 Procedimiento de la segunda etapa

En la segunda etapa de clasificación fuerón analizados solamente aquellos alumnos (observaciones) que obtuvieron un aprovechamiento académico óptimo durante la primera etapa de clasificación. Para empezar el procedimiento hay que observar que en las Tablas A-1, A-2, A-3 se muestran las variables con las que se trabajó durante esta segunta etapa. Estas están por departamentos, considerándose para el Departamento de Estadística e informática, Economía y Planificación y Gestión Empresarial; once, ventiuno y diecinueve variables respectivamente por cada departamento, tomando en cuenta que existen cursos de concentración de acuerdo al departamento académico. Luego se verificó que cada fila contenga los datos (observaciones) en el intervalo establecido de cero a veinte puntos. Si bien es cierto en algunas filas se encontraron datos faltantes (missing values), éstos fueron trabajados mediante el árbol de clasificación y regresión multivariada conocido por sus siglas en inglés como MRT. Finalmente, con el conjunto de datos se elaboró el modelo mediante el árbol de clasificación y regresión CART.

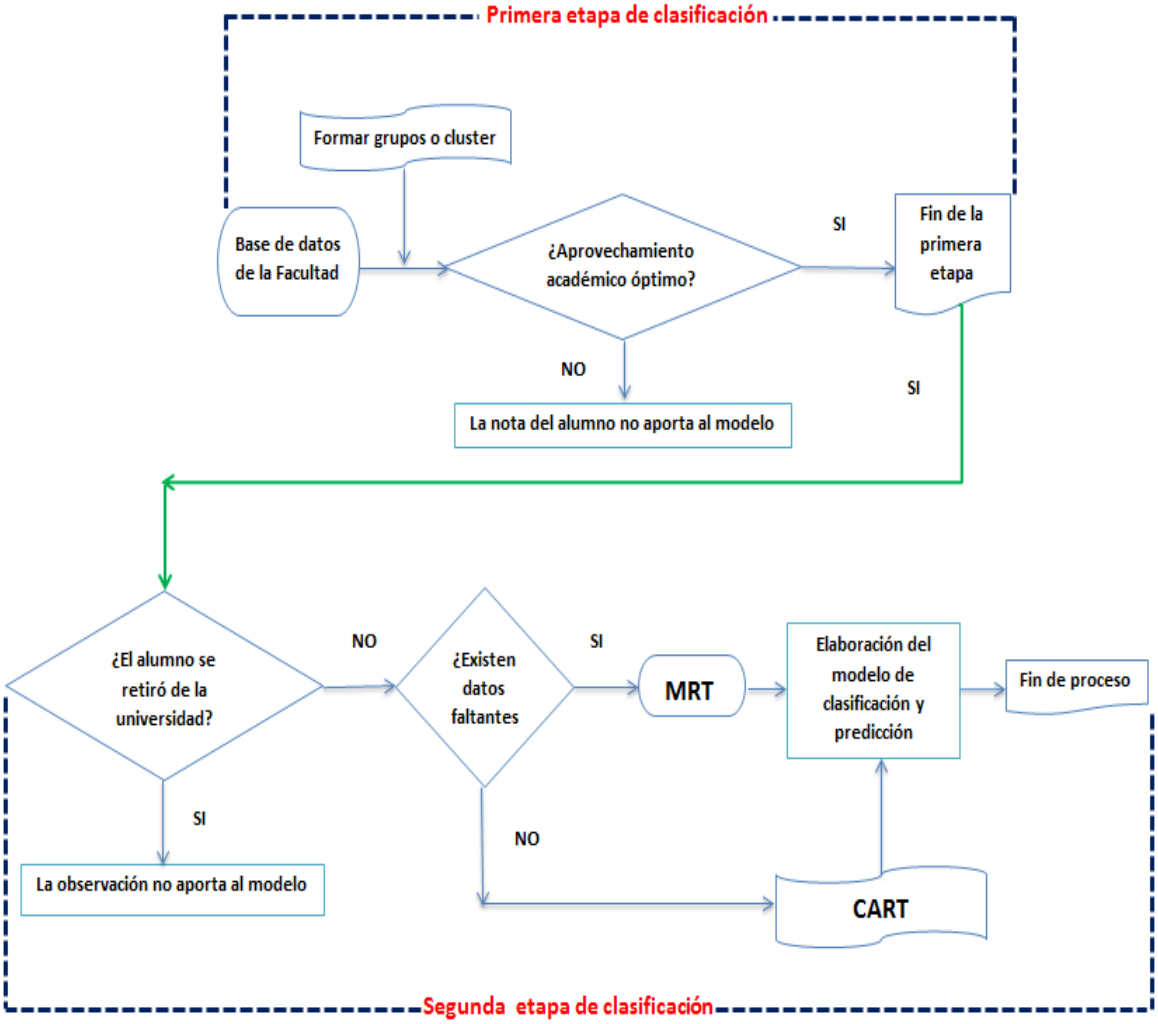

Figure 3–1: Diagrama del modelo de clasificación y predicción en dos etapas.

3.7 Validación cruzada

Como se había mencionado anteriormente para validar el modelo propuesto en este trabajo de investigación; se utilizó la metodología de la validación cruzada con k=10 grupos, que se desarrolló mediante los siguientes pasos:

1. En la primera etapa de clasificación se encontraron los dos grupos (clusters), los alumnos que obtuvieron aprovechamiento óptimo y los que no lo obtuvieron (mediante la clasificacion determinística o el clustering).

2. Con los alumnos (observaciones) que obtuvieron un aprovechamiento académico óptimo en la primera etapa y cada uno con sus respectivas variables en cada fila, se generaron diez grupos (G_1,G_2,...,$G_n; n = 10$).
3. El grupo uno (G_1) fue trabajado como la data de la validación, y al resto de los grupos (G_2,G_3,...,$G_n; n = 10$) como la data de entrenamiento.
4. Con los grupos (G_2,G_3,...,$G_n; n = 10$) se diseñó el modelo MCPD.
5. Luego se utilizó el modelo MCPD del paso 4, para estimar el error y la tasa de mala clasificación con el grupo uno (G_1).
6. Finalmente, se repetió los pasos tres, cuatro y cinco con los nueve grupos restantes, para luego calcular el porcentaje de mala clasificación del modelo propuesto en este trabajo de investigación.

3.8 Método de resubstitución

Para validar el modelo mediante el metodo de resubstitución, se realizó mediante los siguientes pasos:

1. Se forman dos grupos o clusters (según sea el caso).
2. En la primera etapa de clasificación, solamente el grupo de alumnos que obtuvieron un aprovechamiento académico óptimo pasaron a la segunda etapa de clasificación.
3. Con los alumnos que pasaron a la segunda etapa de clasificacion se modelo utilizando el CART.
4. Finalmente; una vez construido el modelo, se realizó la predicción para cada una de las filas con el modelo anterior.
4. Los resultados de predicción del paso anterior, fueron comparados con los datos originales, el cual se tiene como información previa de la base de datos de la facultad de economía y planificación.
5. Finalmente, se contrastó los resultados obtenidos y la información del conjunto de datos, y se calculó la tasa de mala clasificación. Este procedimiento

se realizó para validar el modelo de cada uno de los departamentos académicos analizados en este trabajo de investigación.

CAPITULO 4
RESULTADOS EXPERIMENTALES

A continuación se presenta el análisis del modelo de clasificación y predicción en dos etapas (MCPD) que se realizó de dos formas diferentes: Primero, utilizando la clasificación determinística y segundo, con el análisis de clustering. Ambas metodologías se validaron con el método de resubstitución y la técnica de la validación cruzada. Para desarrollar la metodología de clasificación propuesta en este trabajo de investigación se obtuvieron las tres bases de datos que constan de cada departamento de la Facultad de Economía y Planificación de la Universidad Nacional Agraria la Molina.

4.1 Análisis de deserción académica de la Facultad de Economía y Planificación

La deserción académica se refiere al abandono por parte de los estudiantes al programa académico que eligieron. En esta tesis de investigación se define a un desertor como aquel estudiante que no culminó de manera satisfactoria sus estudios universitarios o aquellos estudiantes que desertaron en continuar con la carrera que eligierón y deciden cambiarse a otra carrera u otra facultad.

Como se puede observar en la Tabla 4-1, la deserción académica afecta de manera considerable al Departamento de Estadística e Informática, principalmente porque los alumnos no cuentan con las aptitudes matemáticas que le requiere la carrera. Los cursos de concentración son considerados "los más difíciles" en comparación a otras facultades.

En el Departamento de Economía y Planificación se nota un porcentaje alto pero

menor en comparación del anterior lo cual indica que el estudiante luego de estudiar durante los dos primeros años su carrera universitaria opta por no finiquitar su programa de estudios. Sin embargo, en el Departamento de Gestión Empresarial se observa un porcentaje alentador en comparación de las dos anteriores lo que indica que los estudiantes sienten mayor satisfacción de elegir dicha carrera y no desertan, sino mas bien tienen la motivación académica de culminarla.

Table 4–1: Tabla de alumnos que desertaron de la Facultad de Economía y Planificación.

Facultad de Economía y Planificación			
Departamento	Total	Desertores	% Deserción académica
Estadística e Informática	308	139	45.12987
Economía y Planificación	433	175	40.4157
Gestión Empresarial	506	95	18.7747

También se analizó el porcentaje de retención, este señala la capacidad que tiene una entidad educativa de poder retener a sus estudiantes con el obejtivo de que terminen de manera satisfactoria la carrera que eligierón y para esta investigación lo conceptualizamos como la cantidad de estudiantes que terminan de manera satisfactoria la carrera que eligierón en el tiempo que establece la entidad. Se puede ver en la Tabla 4-2, que el porcentaje de retención en el Departamento de Gestión Empresarial es muy alto, esto refleja que los objetivos académicos en éste departamento se están cumpliendo en comparación del Departamento de Estadística que una vez más refleja una cifra desalentadora. En el Departamento de Economía y Planificación hay un cierto equilibrio comparado con los otros dos departamentos académicos.

Table 4–2: Tabla de alumnos que culminaron de manera satisfactoria su carrera en la Facultad de Economía y Planificación.

Facultad de Economía y Planificación			
Departamento	Total	Retencion	% Retención académica
Estadística e Informática	176	84	47.72727
Economía y Planificación	308	245	79.54545
Gestión Empresarial	416	384	92.3069

4.2 MCPD utilizando la metodología determinística

Para desarrollar el análisis de clasificación y predicción en dos etapas, se realizarón tres cortes [1] .A continuación presentamos los resultados del Departamento de Estadística e Informática, seguido del Departamento de Economía y Planificación y finalmente de Gestión Empresarial.

4.2.1 Departamento de Estadística e Informática

Para elaborar el modelo MCPD, en la primera etapa se utilizó la metodología determinística y en la segunda etapa se construyó un modelo con los cursos de concentración desde el segundo hasta el quinto año académico, conocidos también como variables predictoras; el modelo generado es el siguiente:

$$\text{GRAD} \sim BASE + GEST + MODLOS + TM1 + SERIES + SOBREV + \text{TP1} + \text{CAL\_EST} + \text{PROB} + \text{AP1} + \text{EST\_GEN}$$

Cuando el corte es mayor o igual a once en el primer año académico; hay una probabilidad de 0.95 que el estudiante termine de manera satisfactoria sus estudios universitarios si aprueba los cursos de estadística general y cálculo avanzado para estadísticos con una nota mayor o igual a catorce. Si el estudiante no logró alcanzar ese puntaje en Cálculo Avanzado y saca menos de ese puntaje, y además

[1] Un corte, se refiere al promedio de los cursos que el estudiante tomo durante su primer año de estudios académicos.

en los cursos de modelos lineales no alcanza a doce y en técnicas de muestreo 1 supera o iguala a doce, la probabilidad se reducirá a 0.67.

Continuando en la misma rama (ver la Figura 4-1) notamos que si en el curso de modelos lineales la nota fue menor a doce, en técnicas de muestreo 1 no superó a doce y en base de datos su puntaje es mayor igual a catorce la probabilidad será del 0.67 . En la Tabla 4-3, se resumen algunas reglas de decisión y probabilidad de graduación. En el mismo árbol de la Figura 4-1, se puede ver que en el nodo principal se encuentra el curso de estadística general, cuando éste es menor a catorce, la probabilidad será del 0.71, si en el curso de análisis de series de tiempo obtiene un puntaje mayor o igual a doce, además que en técnicas de programción 1 obtenga una nota aprobatoria mayor o igual a catorce.

Sin embargo, en la misma rama sino se logró alcanzar este puntaje en el curso de técnicas de programación 1 y en el curso de análisis de sobrevivencia saca una nota mayor o igual a doce, la probabilidad se reducirá en 0.04 comparándolo con la probabilidad anterior. Bajando en la rama donde se encuentra el nodo de análisis de sobrevivencia notamos que si en este curso no alcanzó a doce y en modelos lineales obtiene un puntaje mayor o igual a doce, la probabilidad aumenta en 0.08.

Si en el primer año académico el alumno obtuvo una nota mayor o igual a doce, el modelo es como sigue: en el nodo principal (Ver Figura 4-2) se encuentra el curso de estadística general que se dicta en el segundo año, según la curricula académica; si bajamos en la rama encontramos al curso de cálculo avanzado para estadística. Si en ambos cursos obtiene una nota mayor o igual a catorce, la probabilidad de terminar la carrera satisfactoriamente es del 0.97; pero sino logra alcanzar ese puntaje en cálculo avanzado; y en técnicas de muestreo 1, la nota es mayor o igual a

Table 4–3: Algunas reglas de decisión y probabilidad de graduarse con el MCPD-Determinístico del Departamento de Estadística e Informática con corte mayor o igual a 11.

Regla de decisión	Probabilidad
EST_GEN >= 14 & CALC_EST>=14	0.95
EST_GEN >= 14 & CALC_EST<14 & MODLOS <12 & TM1 >=12 & MODLOS <12 & TM1 >12	0.67
EST_GEN >= 14 & CALC_EST<14 & MODLOS <12 & TM1 <12 & BASE >=14& TM1 <12 & BASE >=14	0.67
EST_GEN <14 & SERIES>= 12 & TP1 >=14	0.71
EST_GEN <14 & SERIES>= 12 & TP1 <14 & SOBREV >=12	0.67
EST_GEN <14 & SERIES>= 12 & TP1 <14 & SOBREV <12 & MODLOS >=12	0.67

doce y se puede afirmar que la probabilidad se reduce a 0.75. En la misma rama del árbol, cuando en Técnicas de Muestreo 1 y en Cálculo de Probabilidades la nota no alcanza a doce la probabilidad de éxito será la misma que la anterior.

En la otra rama de la derecha del mismo árbol (ver Figura 4-2), cuando la nota final es menor a catorce en el curso de Estadística General y en Análisis de Series de Tiempo mayor o igual a doce. La probabilidad es de 0.76, si en Análisis de Sobrevivencia es mayor o igual a doce.

Continuando en esta misma rama cuando la nota en Análisis de Sobrevivencia es menor a doce y en el curso de Cálculo de Probabilidades consigue sacar un puntaje mayor o igual a doce la probabilidad se reducirá en 0.01, con respecto al anterior. Sin embargo, se puede ver que si en Análisis de Sobrevivencia y Cálculo de Probabilidades obtuvo menor a doce puntos en cada curso y además logra superar o igualar a catorce en Cálculo Avanzado para Estadísticos la probabilidad de graduarse de manera satisfactoria será del 0.67.

Table 4–4: Algunas reglas de decisión y probabilidad de graduarse con el MCPD-Determinístico del Departamento de Estadística e Informática con corte mayor o igual a 12.

Regla de decisión	Probabilidad
EST_GEN >= 14 & CALC_EST>= 14	0.97
EST_GEN >= 14 & CALC_EST <14 & TM1 >=12	0.75
EST_GEN >= 14 & CALC_EST <14 & TM1 <12 & PROB <12	0.75
EST_GEN <14 & SERIES>= 12 & SOBREV>= 12	0.76
EST_GEN <14 & SERIES>= 12 & SOBREV<12 & PROB>= 12	0.75
EST_GEN <14 & SERIES>= 12 & SOBREV<12 & PROB<12 & CALC_EST >= 14	0.67

Luego de calcular el rendimiento académico del estudiante en el primer año de estudios y realizando el corte en el promedio mayor o igual a trece, el modelo elaborado nos muestra (Ver Figura 4-3) en el nodo principal al curso de Estadística General, si nos movemos hacia la derecha donde la nota supera o iguala a catorce, se observa que si en Cálculo Avanzado alcanza un puntaje mayor o igual a catorce; entonces, la probabilidad de graduarse es de 0.97; en la misma rama del árbol cuando obtiene en Cálculo Avanzado para Estadística un puntaje menor a catorce y en el curso de Estadística Aplicada 1 un puntaje mayor o igual a doce la probabilidad de culminar de manera satisfactoria la carrera se reduce en 0.09.

En el lado izquierdo del nodo principal del árbol (Ver Figura 4-3) se puede ver, que el curso de estadística general es menor a catorce. Además, si en Técnicas de Muestreo 1 y en el curso de Técnicas de Programación 1 el puntaje supera o iguala a doce, la probabilidad es del 0.85 .

En la misma rama, cuando en Técnicas de Muestreo 1 no alcanza a doce, y a eso le añadimos los cursos de Análisis de Series de Tiempo con puntaje de mayor o igual a doce y en Base de Datos no supera a catorce y en técnicas de muestreo 1 el puntaje es mayor o igual a doce, entonces la probabilidad de graduarse es del 0.57.

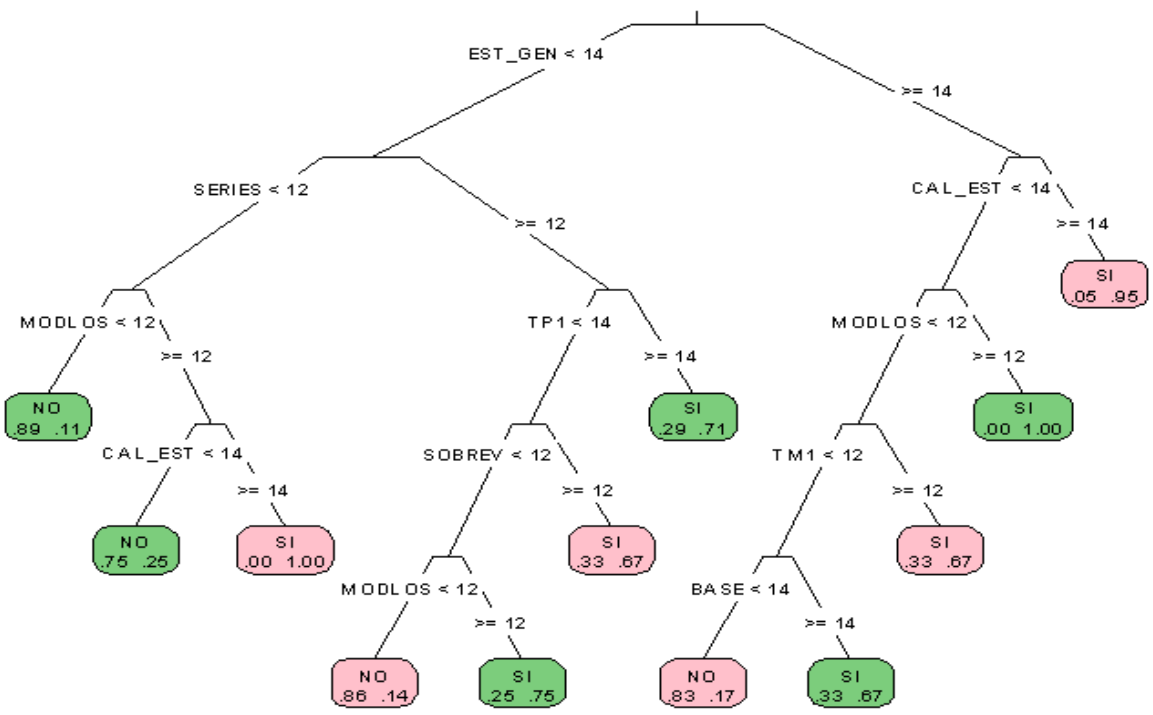

Figure 4–1: Árbol del MCPD-Determinístico del Departamento de Estadística e Informática con corte mayor o igual a 11.

4.2.2 Departamento de Economía y Planificación

Cuando se realizó el análisis estadístico con el modelo de clasificación y predicción en dos etapas (MCPD) para este departamento académico, se construyó el modelo que se muestra a continuación :

En el árbol de la Figura 4-4, se muestra el modelo cuando el corte es mayor o igual a 11. Se puede ver, que en el nodo principal se encuentra el curso de Finanzas Públicas, cuando éste es mayor mayor o igual a 14, y además en el curso de Macroeconomía 1 el puntaje es mayor o igual a 12; la probabilidad de graduarse es del 0.85. Continuando en la misma rama, cuando en Microeconomía 1 el puntaje es menor a 12 y en Contabilidad Gerencial mayor igual a 14, la probabilidad se reduce al 0.72. Siguiendo la misma rama, si en Contabilidad Gerencial el puntaje es menor a 14 y en Teoría del Crecimiento y Desarrollo mayor o igual a 14 la probabilidad es del 0.80.

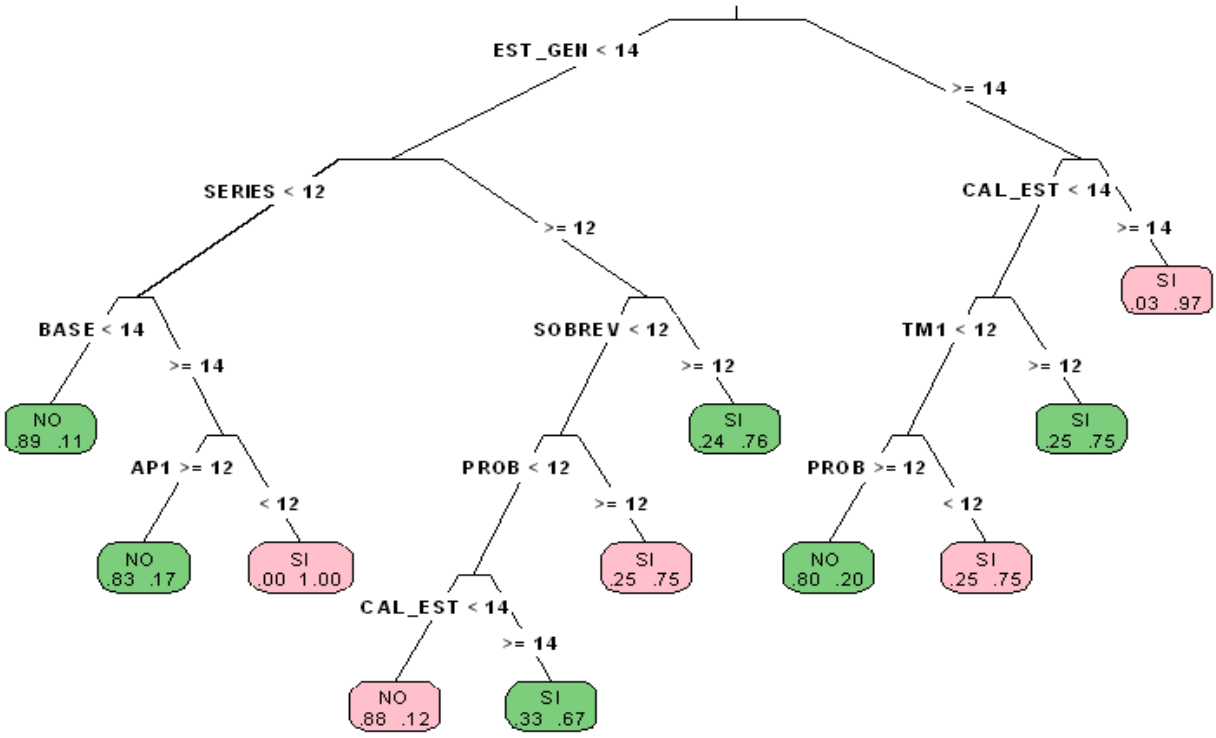

Figure 4–2: Árbol del MCPD-Determinístico del Departamento de Estadística e Informática con corte mayor o igual a 12.

En la misma figura; en la rama principal a la izquierda, cuando en el curso de Finanzas Públicas el puntaje es menor a 14 y en Teoría del Crecimiento y Desarrollo su nota es mayor o igual a 12; la probabilidad de graduarse como Economista es del 0.80. Continuando en la misma rama, cuando en Teoría del Crecimiento y Desarrollo es menor a 12, y además en Finanzas Públicas la nota es mayor o igual a 12, la probabilidad de graduarse se reduce en 0.02.

Realizando el corte mayor o igual a doce en la figura 4-5, si en el curso de Finanzas Públicas el puntaje es mayor o igual a 14 y en Macroeconomía el puntaje es mayor o igual a 12; la probabilidad de graduarse es de 0.88. Bajando en la misma rama cuando en Macroeconomía el puntaje es menor a 12 y en Contabilidad Gerencial la nota es menor a 14, y además en Contabilidad General y en Teoría de Crecimiento y Desarrollo el puntaje es mayor o igual a 14, entonces la probabilidad se reduce a 0.56.

Table 4–5: Algunas reglas de decisión y probabilidad de graduarse con el MCPD-Determinístico del Departamento de Estadística e Informática con corte mayor o igual a 13.

Regla de decisión	Probabilidad
EST_GEN >=14 & CALC_EST>=14	0.97
EST_GEN >=14 & CALC_EST<14 & AP1 >=12	0.88
EST_GEN <14 & TM1>= 12 & TP1>= 12	0.85
EST_GEN <14 & TM1<12 & SERIES= 12 & BASE <14 & TM1 >= 12	0.57

GRAD$\sim C_GEN + C_GER + DE + D_EMP + METRIA + E_MA$ + ESP+MACR1+ TC_D+ FIN_PUB

En la parte izquierda del nodo principal, si en el curso de Finanzas Públicas el puntaje es menor a 14 y en Teoría de crecimiento y desarrollo el puntaje es mayor o igual a 12, en Contabilidad General el puntaje es mayor o igual 14, podemos afirmar que la probabilidad de graduarse es de 0.85. En la misma rama, se puede ver que si en Contabilidad Gerencial el puntaje es menor a 14 y mayor o igual a 12, y además en Macroeconomía 1 el puntaje es mayor o igual a 12, entonces la probabilidad de graduarse es de 0.71.

Cuando el corte es mayor o igual a trece, en el árbol de la Figura 4-6, notamos en el nodo principal al curso de Contabilidad General. Si en este curso y en el Finanzas Públicas obtiene un puntaje mayor o igual a 12, la probabilidad de graduarse satisfactoriamente en este departamento será del 0.92.

Del nodo principal a la izquierda cuando en el curso de Contabilidad Gerencial obtiene un puntaje menor a 12 y en Teoría del Crecimiento y Desarrollo mayor o igual a 12, entonces se puede afirmar que la probabilidad de graduarse es del 0.82.

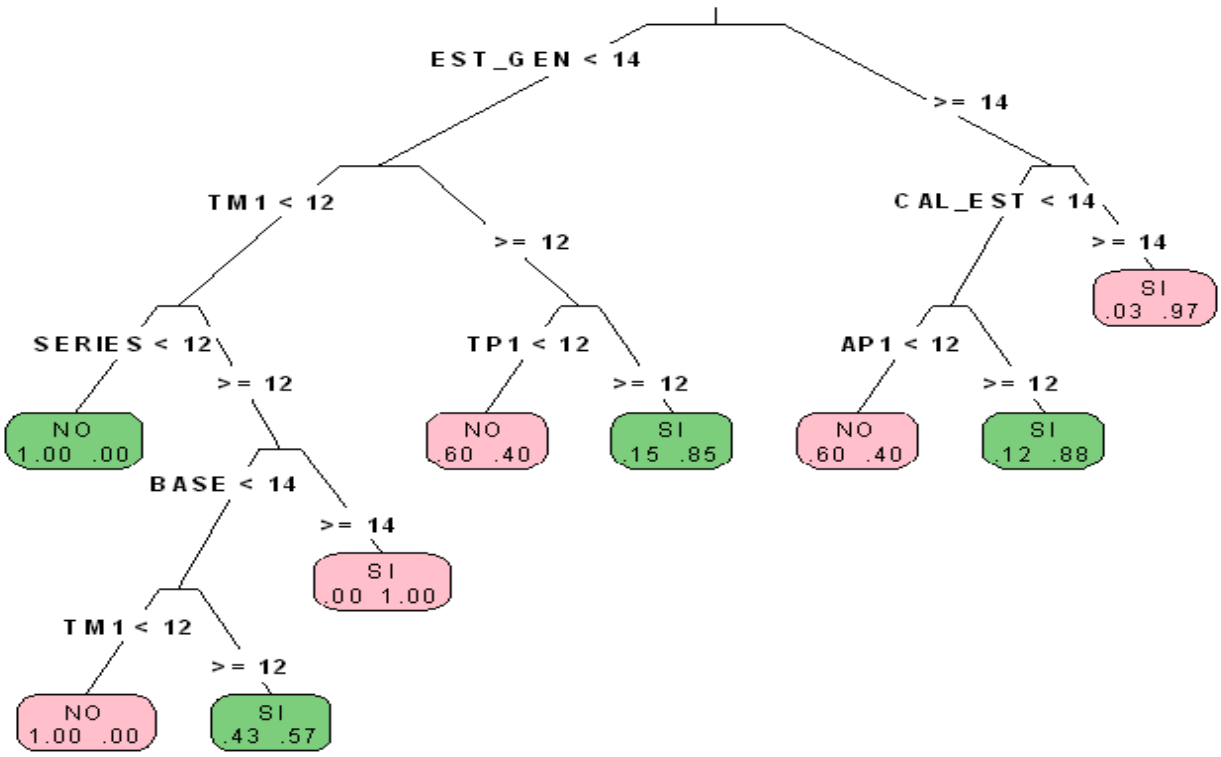

Figure 4–3: MCPD-Determinístico del Departamento de Estadística e Informática con corte mayor o igual a 13.

4.2.3 Departamento de Gestión Empresarial

El modelo para el Departamento de Gestión empresarial, se elaboró con las siguientes variables predictoras mostradas a continuación:

Cuando realizamos el corte de mayor o igual a once, se puede ver al curso de Finanzas 2 en el nodo principal del árbol. Cuando el puntaje en este curso es mayor o igual a 10 y además en los cursos de Negociosm Internacionales y Marketing el puntaje es mayor o igual a 14 y 12, respectivamente; podemos afirmar que la probabilidad de graduarse es 0.98. Continuando en la misma rama, cuando en el curso de Marketing no alcanza a 12 y en Contabilidad Gerencial el puntaje es menor a 16. La probabilidad de graduarse es 0.96.

Bajando por el nodo principal; cuando en Finanzas 2, su puntaje es mayor o igual a 10, y además en los cursos de Negocios Internacionales e Investigación de Operaciones obtiene un puntaje menor a 14, la probabilidad de graduarse de manera

Table 4–6: Algunas reglas de decisión y probabilidad de graduarse con el MCPD-Determinístico del Departamento de Economía y Planificación con corte mayor o igual a 11.

Regla de decisión	Probabilidad
FIN_PUB >= 14 & MACR 1 >= 12	0.85
FIN_PUB >= 14 & MACR 1 <12 & C_ GER >= 14	0.72
FIN_PUB >= 14 & MACR 1 <12 & C_ GER <14 & TC_D >= 14	0.80
TC_D <14 & C_GEN <16	0.57
FIN_PUB <14 & TC_D >= 12	0.80
FIN_PUB <14 & TC_D <12 & FIN_PUB >=12	0.78

Table 4–7: Algunas reglas de decisión y probabilidad de graduarse con el MCPD-Determinístico del Departamento de Economía y Planificación con corte mayor o igual a 12.

Regla de decisión	Probabilidad
FIN_PUB >= 14 & MACR 1 >= 12	0.88
FIN_PUB >= 14 & MACR 1 <12 & C_ GER <14 C_ GEN >= 14 & TC_D >= 14	0.56
FIN_PUB <14 & TC_D >= 12 & C_GEN >= 14	0.85
FIN_PUB <14 & TC_D >= 12 & C_GEN <14 & C_GEN >=12 & MACR 1 >=12	0.71

exitosa es del 0.92. En la misma rama, cuando en Investigación de Operaciones el puntaje es mayor o igual a 14 y en Finanzas 1 menor a 14, entonces la probabilidad se reducirá a 0.83.

Cuando el corte es mayor o igual a doce, la probabilidad es de 0.92; si en el curso de Finanzas 1, el puntaje es mayor o igual a 10.5, en Negocios Internacionales mayor o igual a 14, en Marketing menor a 12 y en Contabilidad Gerencial menor a 16.

Finalmente, cuando se realiza el corte a mayor o igual 13, para el análisis de este departamento académico, se encontró el árbol de la Figura 4-9, en el nodo principal tenemos al curso de Planeamiento estratégico, cuando el estudiante alcanza obtener

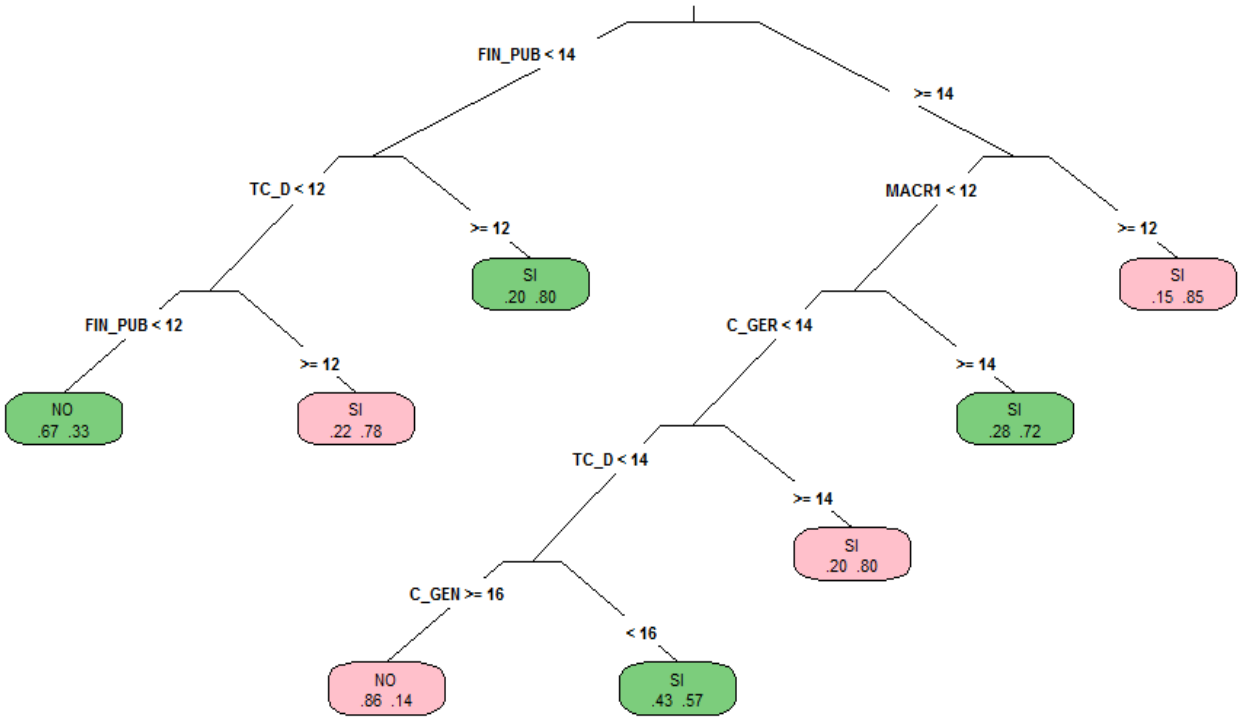

Figure 4–4: MCPD-Determinístico del Departamento de Economía y Planificación con corte mayor o igual a 11.

Table 4–8: Algunas reglas de decisión y probabilidad de graduarse con el MCPD-Determinístico del Departamento de Economía y Planificación con corte mayor o igual a 13.

Regla de decisión	Probabilidad
C_GER >= 12 & FIN_ PUB >= 12	0.92
C_GER <12 & TC_D >= 12	0.82

una puntaje que va del intervalo de [18-20> ; y si a eso le añadimos el puntaje en el curso de Contabilidad general con un puntaje menor a 18, la probabilidad de graduarse es del 0.88.

En la misma rama del árbol, si en el curso de Finanzas 2 y en Macroeconomía 1 el puntaje es menor a 14, la probabilidad es de 0.99.

4.3 MCPD utilizando la metodología jerárquica del análisis de cluster

El diseño del modelo de clasificación y predicción en dos etapas (MCPD) utilizando esta metodología fue utilizado para la primera etapa de clasificación, con

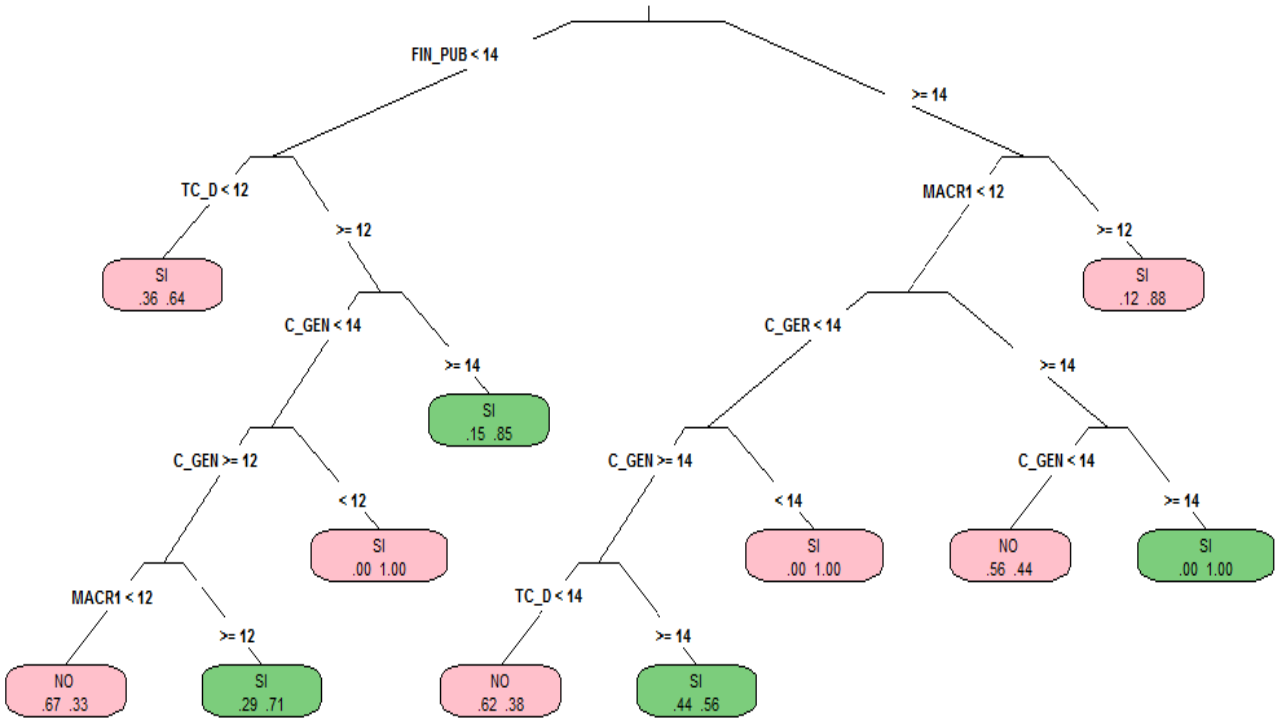

Figure 4–5: MCPD-Determinístico del Departamento de Economía y Planificación con corte mayor o igual a 12.

$$\text{GRAD} \sim AFIN + C\_GER + ECON + FI1 + INV\_O + MARK + \text{NEG\_INT} + \text{ORG\_MET} + \text{PLA\_EST} + \text{T\_PESQ}$$

las seis variables utilizadas anteriormente, según sea el departamento académico que esta siendo analizado; esto permitió encontrar dos grupos, los que fueron etiquetados como estudiantes con rendimiento óptimo y los estudiantes que no obtuvierón rendimiento óptimo. En la segunda etapa de clasificación se utilizarón dos metodologías estadísticas, como son: el CART, para elaborar el modelo; y el MRT para estimar los valores faltantes en el conjunto de datos.

4.3.1 Departamento de Estadística e Informática

Para modelar con el clustering jerárquico en este trabajo de investigación, primero se generarón los grupos utilizando la técnica estadística del clustering jerárquico; seguidamente, con los alumnos que obtuvierón un rendimiento óptimo en su primer año académico se les aplicó el árbol de clasificación y predicción.

En el árbol de la Figura 4-9, se puede ver en el nodo principal del árbol al curso

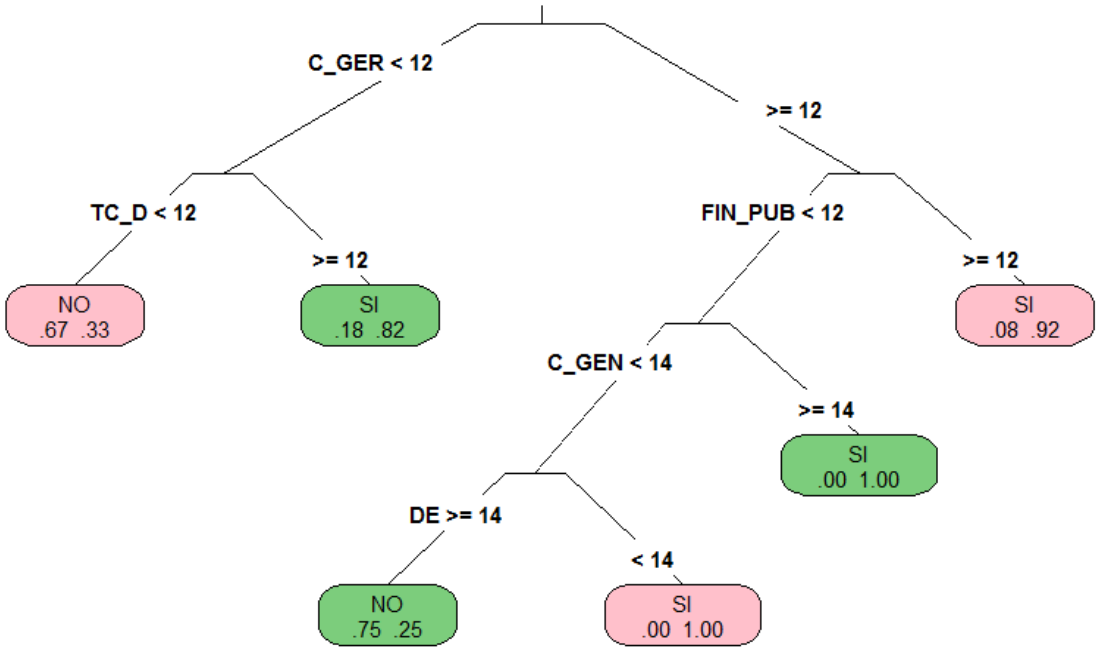

Figure 4–6: MCPD-Determinístico del Departamento de Economía y Planificación con corte mayor o igual a 13.

Table 4–9: MCPD-Determinístico del Departamento de Gestión Empresarial con corte mayor o igual a 11.

Regla de decisión	Probabilidad
FI 1 >= 10 & NEG_INT >14 & MARK >= 12	0.98
FI 1 >= 10 & NEG_INT >14 & MARK <12 & C_GER <= 16	0.96
FI 1 >= 10 & NEG_INT <14 & INV_O <14	0.92
FI 1 >= 10 & NEG_INT <14 & INV_O >14 & FI 1 <14	0.92

de modelos lineales, que lo dictan en el cuarto año de la carrera, si nos vamos hacia la derecha de ese nodo cuando el puntaje es mayor o igual a doce; y además en cálculo avanzado para estadísticos el puntaje es mayor o igual a catorce, la probabilidad de graduarse como Ingeniero estadístico e informático es del 0.93. Bajando en la misma rama encontramos a cálculo avanzado con un puntaje menor a catorce y cálculo de probabilidades con un puntaje mayor o igual a doce; donde la probabilidad de culminar de manera satisfactoria la carrera es de 0.88.

En el mismo árbol avanzando hacia el nodo izquierdo, cuando el puntaje en el

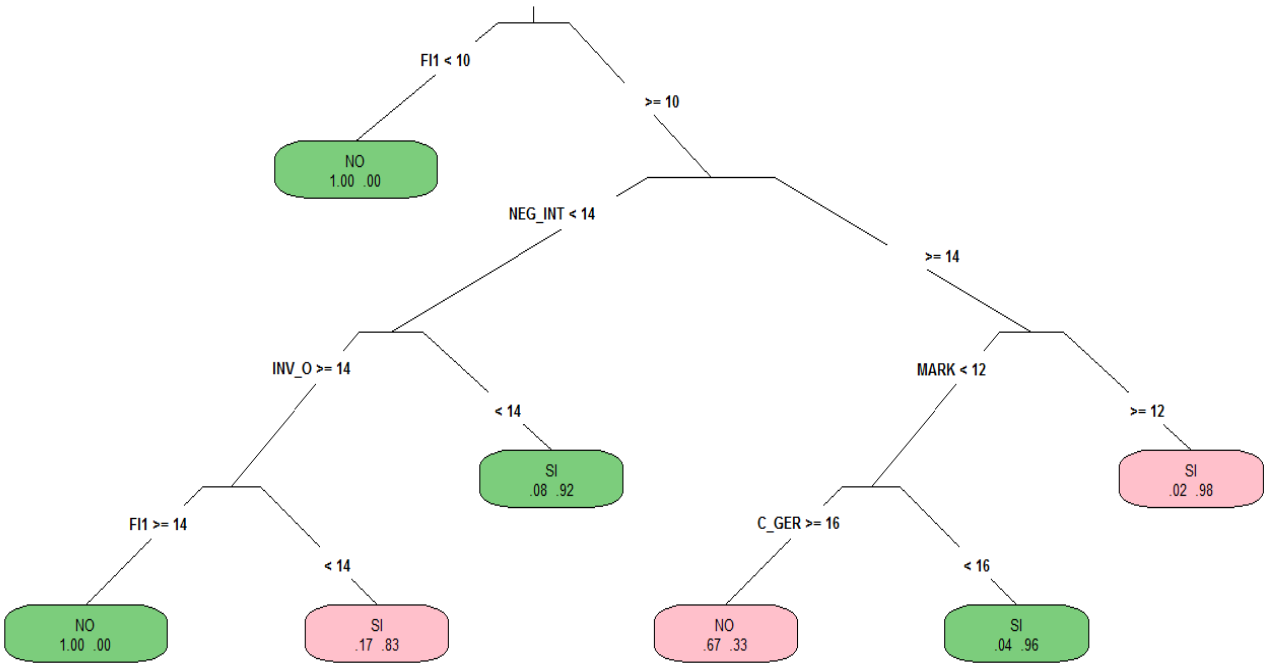

Figure 4–7: MCPD-Determinístico del Departamento de Gestión Empresarial con corte mayor o igual a 11.

Table 4–10: Regla de decisión y probabilidad de graduarse con el MCPD-Determinístico del Departamento de Gestión Empresarial con corte mayor o igual a 12.

Regla de decisión	Probabilidad
FI 1 >= 10 & NEG_INT >14 & MARK >= 12	0.98
FI 1 >= 10 & NEG_INT >14 & MARK <12 & C_GER <= 16	0.92

curso de modelos lineales es menor a doce; y asimismo, en los cursos de análisis de series de tiempo y análisis de sobrevivencia el puntaje es mayor o igual a doce, la probabilidad de éxito es del 0.8.

4.3.2 Departamento de Economía y Planificación

En el modelo elaborado con el MCPD para este departamento académico, se puede notar que en el nodo principal se encuentra el curso de Finanzas Públicas. Si en este curso obtiene un puntaje mayor o igual a 12, y además en Econometría mayor o igual a 10.5 entonces se puede afirmar que la probabilidad de graduarse es de 0.81. En el otro lado de la rama cuando en el curso de Finanzas públicas el

Table 4–11: MCPD-Determinístico del Departamento de Gestión Empresarial con corte mayor o igual a 13.

Regla de decisión	Probabilidad
NEG_INT >10 & FI 1 >8.5 & PLA_EST <= 20 INV_O >16 & ECON >12 & FI 1 <16	0.99

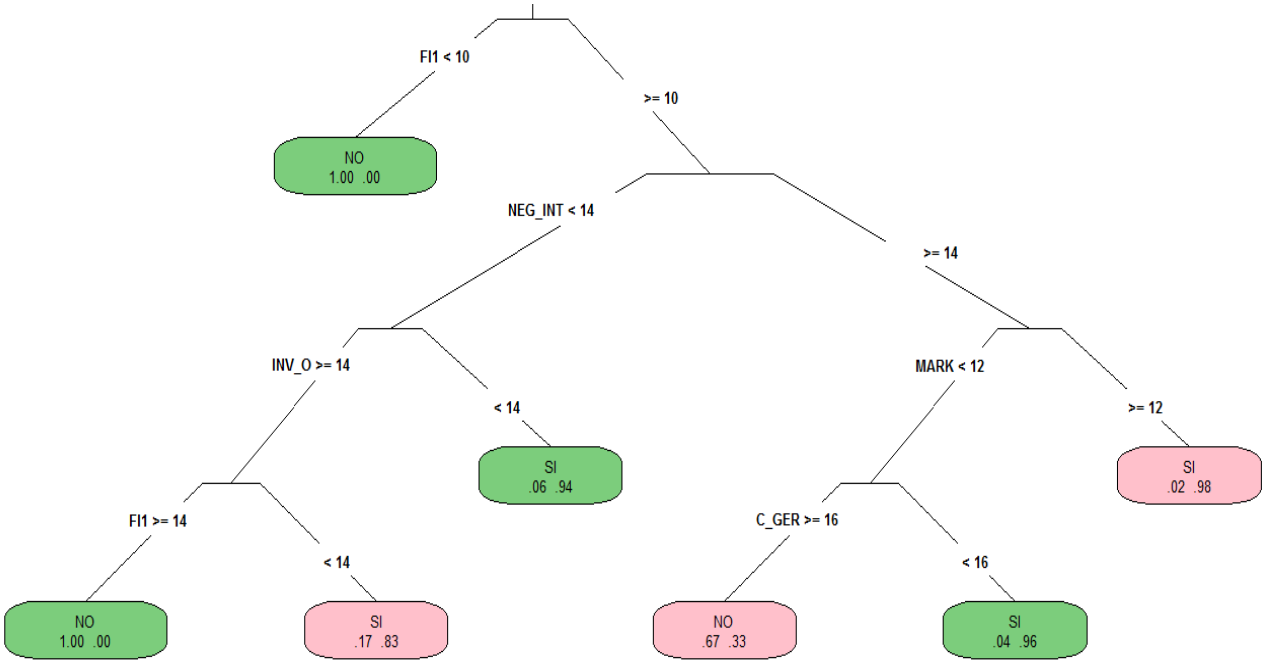

Figure 4–8: MCPD-Determinístico del Departamento de Gestión Empresarial con corte mayor o igual a 12.

puntaje es menor a 12 y en el curso de Estadistica general la nota es mayor o igual a 14, entonces la probabilidad de graduarse aumentará en 0.02 con respecto a la probabilidad anterior.

4.3.3 Departamento de Gestión Empresarial

Utilizando el clustering jerárquico para la primera etapa de clasificación y siguiendo la metodología presentada en este trabajo de investigación; obtenemos el árbol de la Figura 4-11, Se observa en el nodo principal al curso de Finanzas 1, si en este curso el puntaje es mayor o igual a 10.5 y en Negocios Internacionales mayor o igual a 14, además en Marketing mayor o igual a 12. La probabilidad de graduarse es de 0.98. Bajando en la misma rama, si en el curso de Marketing el puntaje es menor a 12 y en Contabilidad Gerencial el puntaje es menor a 16; la probabilidad

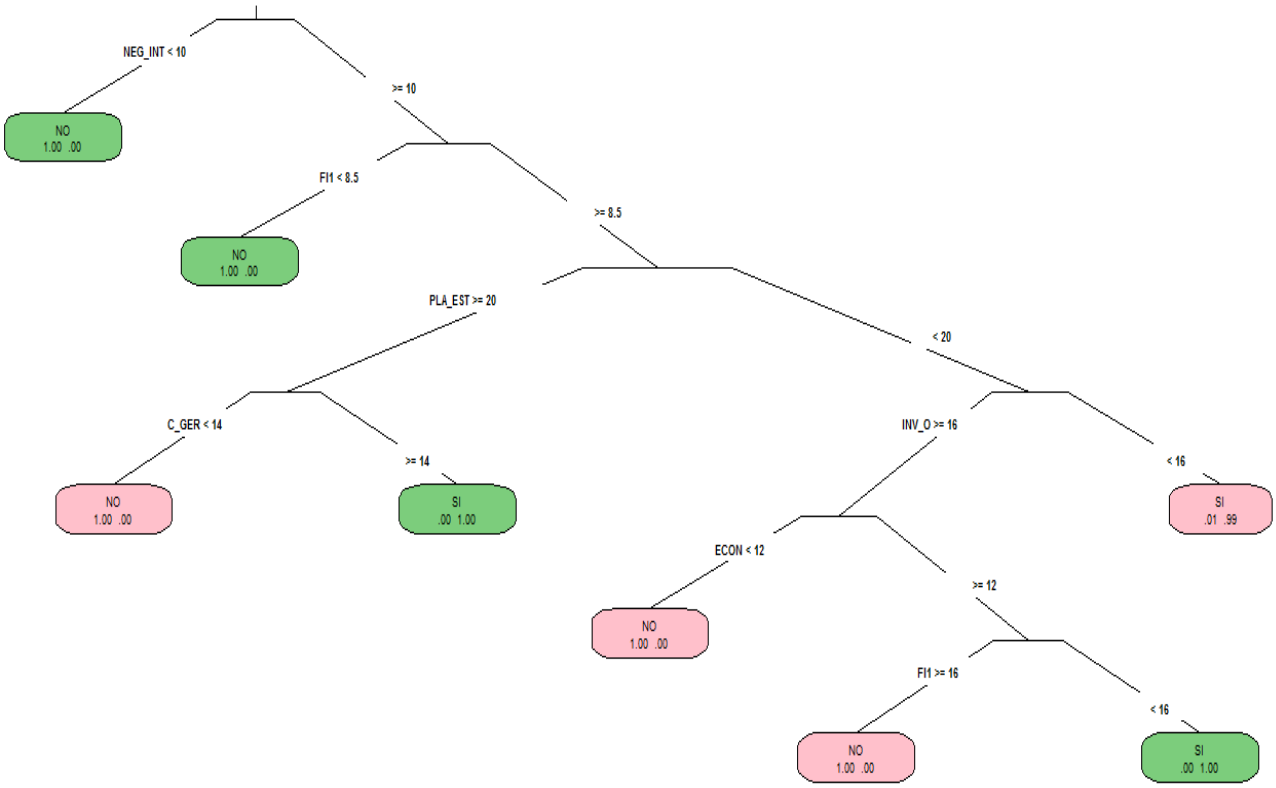

Figure 4–9: MCPD-Determinístico del Departamento de Gestión Empresarial con corte mayor o igual a 13.

es de 0.96. En la otra rama del árbol, cuando en el curso de Finanzas 1 el puntaje es mayor a 10.5 y en el curso de Negocios Internacionales el puntaje es menor a 14, y además en Investigacion de operaciones la nota es menor a 14, entonces la probabilidad de graduarse es de 0,93. Por último, si el puntaje en Finanzas 1 esta en el intervalo de [10.5-14], y en Negocios Internaacionales el puntaje es menor a 14 y en Investigación de Operaciones es mayor o igual a 14, entonces la probabilidad de graduarse es de 0.83.

Siguiendo en la misma rama del árbol, si en el curso de Contabilidad de costos y en Contabilidad general el puntaje supera o iguala a catorce la probabilidad se reduce en 0.12. Sin embargo, si en el curso de Organización y métodos el puntaje es menor a 12 y en Estadística general mayor o igual a 12, la probabilidad de graduarse es de 0.89.

Table 4–12: Regla de decision y probabilidad de graduarse con el MCPD con el clustering jerárquico del Departamento de Estadística e Informática.

Regla de decisión	Probabilidad
MODLOS >= 12 & CALC_EST>= 14	0.93
MODLOS >= 12 & CALC_EST<14 & PROB >= 12	0.88
MODLOS <12 & SERIES >= 12 & SOBREV >= 12	0.80

Table 4–13: Regla de decision y probabilidad de graduarse con el MCPD con el clustering jerárquico del Departamento de Economía y Planificación.

Regla de decisión	Probabilidad
FIN_PUB >= 12 & METRIA >= 10.5	0.81
FIN_PUB <12 & C_GRAL >= 14	0.83

4.4 Técnica de la validación cruzada para validar el MCPD

Con el objetivo de validar el modelo de clasificación y predicción en dos etapas (MCPD) propuesto en este trabajo de investigación, se utilizó la técnica estadística conocida como validación cruzada cuando k=10 grupos.

Se puede ver en la Tabla 4-15 que si utilizamos el MCPD-Determinístico para el Departamento de Gestión empresarial cuando el corte es mayor o igual a once, doce y trece los errores son 8.025%, 7.7719%, 2.996% respectivamente. Notamos que la tasa de error utilizando el MCPD para este departamento es menor que el 10%, este es un indicador que el modelo propuesto por este trabajo puede ser utilizado para estimar la probabilidad de que un alumno termine de manera satisfactoria su carrera universitaria en los años establecidos por la entidad educativa.

Para el Departamento de Economía y Planificación se puede notar que el error es del 11.3% cuando el corte es once. Cuando el corte es doce la tasa es de 12.647%; y por último, cuando el corte es trece la tasa de error es de 12.171%; por lo tanto hay confianza de utilizar el modelo MCPD para este departamento académico.

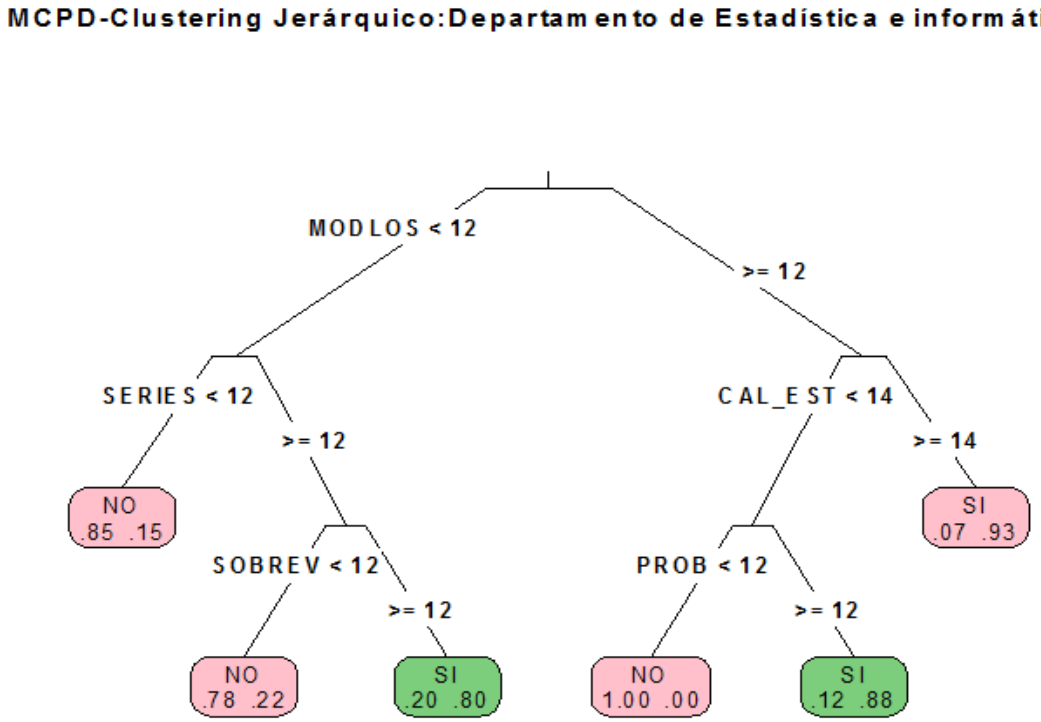

Figure 4–10: Árbol del MCPD con el clustering jerárquico del Departamento de Estadística e Informática.

Los resultados de los errores obtenidos para el Departamento de Estadística e Informática se puden ver en la Tabla 4-15; cuando el corte es once, doce y trece tenemos 29.1478%, 27.6% y 29.127%, respectivamente. Estos valores altos se deben principalmente porque en este departamento académico los alumnos tienden a retirarse o cambiarse de carrera, además hay una alta tasa de deserción académica y bajo índice de retención.

Realizando la validación cruzada cuando el modelo de clasificación y predicción en dos etapas es elaborado con el clustering jerárquico durante la primera etapa de clasificación, obtenemos los resultados de la Tabla 4-16, donde se puede observar que la tasa de error es más baja en el Departamento de Gestión empresarial con el 10.3333%, que nos señala que el modelo predice de manera confiable la probabilidad que un estudiante se gradue exitosamente de este departamento académico; algo similar ocurre con el Departamento de Economía y Planificación porque su tasa de error es del 13.45638%, mientras que con el Departamento de Estadística e

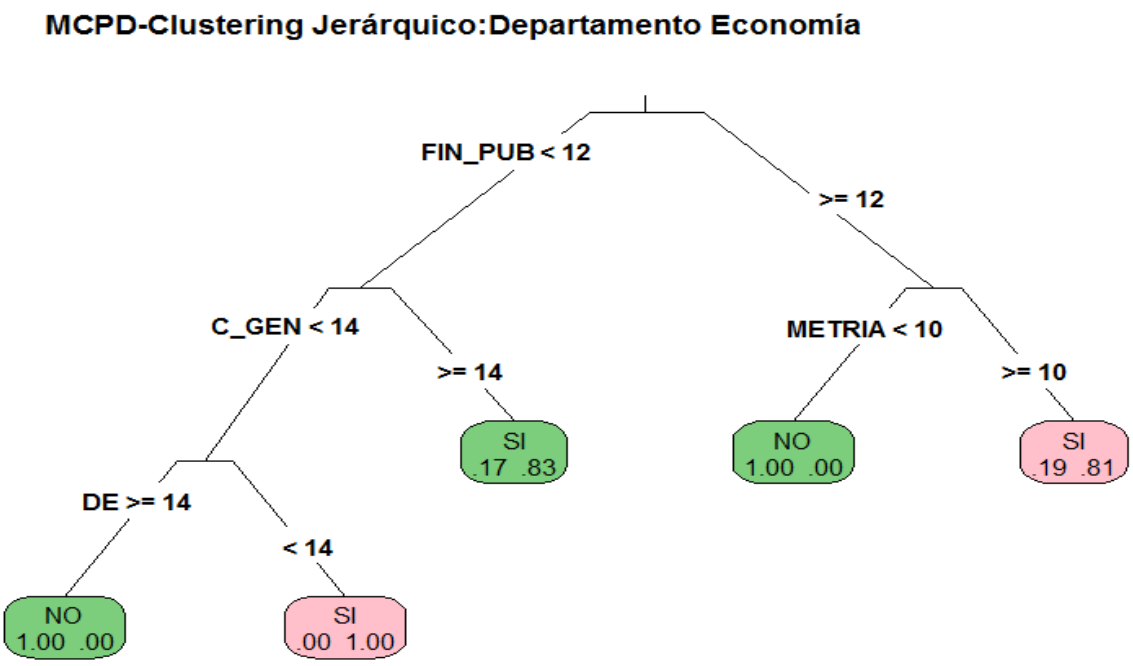

Figure 4–11: Árbol del MCPD con el clustering jerárquico del Departamento de Economía y Planificación.

Informática la tasa de mala clasificación es del 21.167%, el cual comparándolo con el método desterminístico, éste es el más fiable.

4.5 Método de resubstitución para validar el MCPD

Otra manera de validar el modelo propuesto por este trabajo de investigación es mediante el método de resubstitución. En la Tabla 4-17, se muestran los resultados obtenidos cuando se utiliza el MCPD-Determinístico para los tres departamentos de la facultad. La tasa de error en el departamento de Gestión empresarial es la menor para el punto de corte 11, y la tasa de error para el Departamento de Estadística e informática es la misma cuando se usan los puntos de corte 11 y 13. Con el clustering jerárquico(Ver Tabla 4-19) la tasa de error por resubstitución para el Departamento de Estadística es el doble comparado con los otros departamentos.

4.6 Estimación de los datos faltantes con el árbol de clasificación y predicción multivariada (MRT)

En esta sección estimamos los datos faltantes existentes en la base de datos de la Facultad de Economía y Planificación mediante el árbol de regresión multivariada

Table 4–14: Regla de decision y probabilidad de graduarse con el MCPD con el clustering jerárquico del Departamento de Gestión Empresarial.

Regla de decisión	Probabilidad
FI 1 >10.5 & NEG_INT >= 14 & MARK >12	0.98
FI 1 >10.5 & NEG_INT >= 14 & MARK <12 & C_GER <16	0.96
FI 1 >10.5 & NEG_INT >= 14 & INV_O <14	0.93
FI 1 >10.5 & NEG_INT >= 14 & INV_O <14	0.93
FI 1 >10.5 & NEG_INT <= 14 & INV_O >14 & FI 1 <14	0.83

Table 4–15: Tabla de errores utilizando el método de validación cruzada para MCPD-Determinístico.

Tasa de Errores del MCPD-Determinístico			
Corte	Dep.Gestión	Dep.Economía	Dep. Estadística
11	8.025316%	11.23377%	29.14773%
12	7.771883%	12.64706%	27.6%
13	2.996255%	12.17105%	29.127%

conocido como MRT, con el objetivo de obtener una base de datos completa. Los resultados a continuacioń muestran las tasas de errores después de estimar estos datos faltantes y modelar mediante el MCPD, para luego validarlo con el método de resubstitución y la validación cruzada cuando k=10.

4.6.1 Validación cruzada aplicada al MCPD en la base de datos completa y estimada con el MRT

En la Tabla 4-19, se muestra la tasa de errores para los tres departamentos cuando se estimo los datos faltantes con el MRT y luego se realizó el modelo de clasificación y predicción en dos etapas (MCPD) para los tres cortes de cada departamento. Se puede ver que en el Departamento de Gestión empresarial los errores se reducen a medida que se aumenta el punto de corte. Comparándolo con los resultados de la Tabla 4-15, los errores se reducen cuando los datos faltantes se estiman con el MRT.

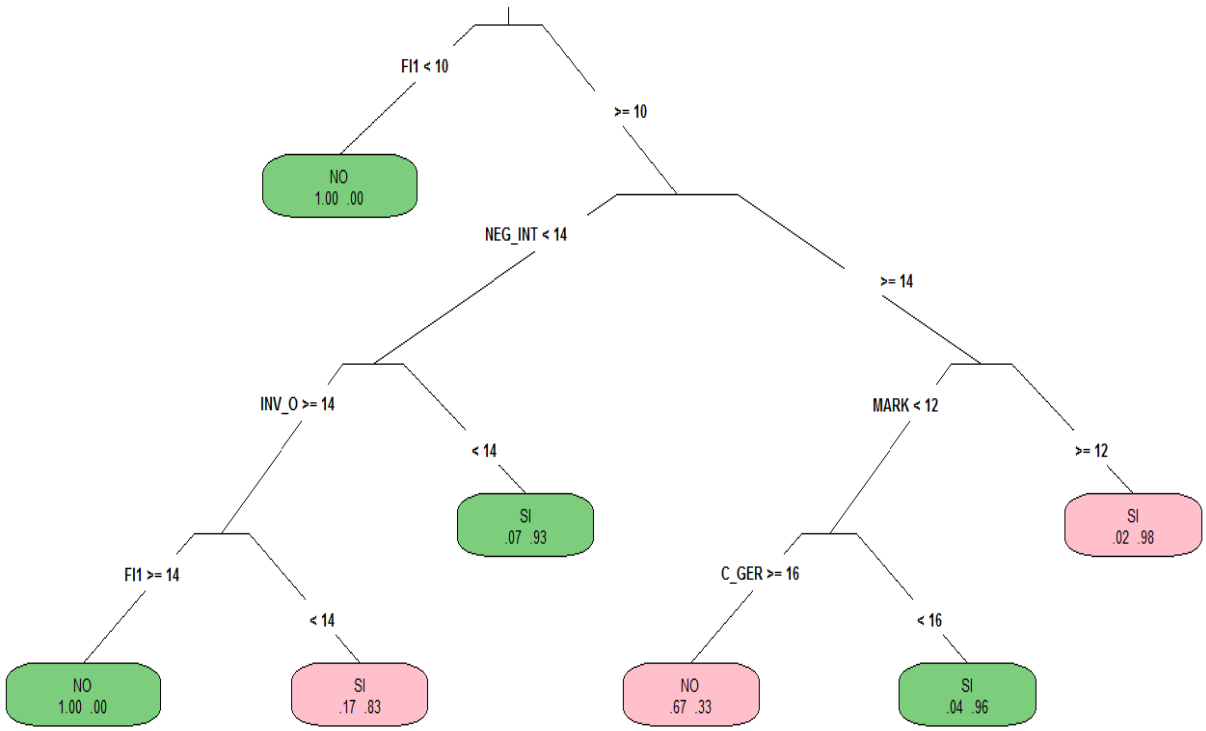

Figure 4–12: Árbol del MCPD con el clustering jerárquico del Departamento de Gestión empresarial.

Table 4–16: Tabla de errores utilizando el método validación cruzada para MCPD con el clustering jerárquico.

Errores del MCPD con el clustering jerárquico			
Método	Dep. Estadística	Dep.Economía	Dep.Gestión
Hclust	21.16667%	13.45638%	10.33333%

En el Departamento de Economía y Planificación la tasa de errores para los tres cortes se encuentran en el intervalo de 3.80% a 6.91%, estos valores son más peque˜as comparados con los resultados de la Tabla 4-15. Para el Departamento de Estadística e Informática, se puede ver que las tasas de errores son comparables con las tasas de errores de la tabla Tabla 4-15.

Estimando los datos faltantes del conjunto de datos de cada departamento con el MRT y seguidamente realizando el modelo de clasificación y predicción en dos etapas con el clustering jerárquico y finalmente, utilizando la técnica de la validación

Table 4–17: Tabla de errores utilizando el método de resubstitución para MCPD-Determinístico.

Errores del MCPD-Determinístico			
Corte	Dep. Gestión	Dep. Economía	Dep. Estadística
11	6.075949%	7.142857%	14.77273%
12	13.33333%	6.25%	6.100796%
13	13.09524%	12.17105%	2.621723%

Table 4–18: Tabla de errores utilizando el método de resubstitución para MCPD con el clustering jerárquico.

Errores del MCPD con el clustering jerárquico			
Método	Dep. Estadística	Dep.Economía	Dep.Gestión
Hclust	14.54545%	7.04698%	7.272727%

cruzada cuando k=10, obtenemos los resultados de la Tabla 4-23. Cuando se realiza la comparación de las Tablas 4-4 y 4-8, la primera tabla son las tasas de errores cuando se modela con los datos faltantes y la segunda tabla, la tasa de errores cuando se estima los datos faltantes con el MRT. Se puede ver que la tasa de error se reduce cuando se estima los datos faltantes para los departamentos de Economía y Gestión empresarial, mientras que en el Departamento de Estadística no sería conveniente estimar los datos faltantes, y lo recomendable es seguir trabajando con el conjunto de datos originales.

4.6.2 Método de resubstitución para validar el MCPD en la base de datos completa y estimada con el MRT

Si elaboramos el MCPD-Determinístico con los datos completos estimados y luego lo validamos con el método de resubstitución, se obtiene como resultado la tasa de errores para los tres cortes en los tres departamentos de la Tabla 4-24. Si realizamos una comparación de las Tablas 4-5, obtenidas con los datos originales y la Tabla 4-24 con los datos faltantes estimados, se puede ver que la tasa de error se reduce para los departamentos de Economía y Gestión empresarial, lo mismo ocurre para

Table 4–19: Tabla de errores con el MRT utilizando el método de validación cruzada para MCPD-Determinístico.

Tasa de errores utilizando MRT en MCPD-Determinístico			
Corte	Dep.Gestión	Dep.Economía	Dep. Estadística
11	4.014423%	3.831169%	30.05682%
12	3.984772%	6.544118%	28.73333%
13	2.392857%	6.907895%	30.2381%

Table 4–20: Tabla de errores utilizando el método validación cruzada para MCPD con el clustering jerárquico usando MRT.

Errores del MCPD con el clustering jerárquico usando MRT			
Método	Dep. Estadística	Dep.Economía	Dep.Gestión
Hclust	29.72477%	3.771626%	4.272727 %

el Departamento de Estadística e Informática, pero solamente cuando el corte es once; con el resto de los cortes la tasa de error aumenta.

Table 4–21: Tabla de errores utilizando el método de Resubstitución para MCPD-Determinístico.

Errores del MCPD-Determinístico utilizando MRT			
Corte	Dep. Gestión	Dep. Economía	Dep. Estadística
11	1.923077%	1.948052%	6.818182%
12	3.553299%	3.308824%	8%
13	1.071429%	5.921053%	7.142857%

Finalmente, cuando se modela con el clustering jerárquico con datos que estan completos y estimados con el MRT y validados con el método de resubstitución, los resultados de la Tabla 4-25 lo comparamos con los resultados obtenidos de la Tabla 4-21, y se puede ver que la tasa de errores se reduce cuando estimamos los datos faltantes con el MRT, para los tres departamentos académicos de la Facultad de Economía y Planificación.

Table 4–22: Tabla de errores con el MRT utilizando el método de validación cruzada para MCPD-Determinístico.

Tasa de errores utilizando MRT en MCPD-Clutering jerárquici			
Método	Dep. Estadística	Dep.Economía	Dep.Gestión
Hclust	7.33945%	2.0761 %	2.424242%

4.7 Imputación de los datos faltantes con el método del knn

Los datos faltantes de la base de datos de la Facultad de Economía y Planificación fuerón imputados mediante el método del vecino mas cercano (knn), obteniéndose una base de datos completa. Luego se le aplicó la metodología propuesta en este trabajo de investigación, definido como modelo de clasificación y predicción en dos etapas MCPD-Determinístico y con el clustering jearárquico. Por último se validarón los modelos utilizando el método de validación cruzada con k=10 y el método de resubstitución.

4.7.1 Validación cruzada aplicado al MCPD en la base de datos imputados por el KNN

Utilizando los datos completos e imputados con el knn, seguidamente utilizando el MCPD-Determinístico; y luego validando ese modelo con la validación cruzada cuando k=10, obtenemos la Tabla 4-26. Se puede ver que en el Departamento de Gestión Empresarial la tasa de errores se aproxima a la tasa de errores cuando se trabaja con el conjunto de datos faltantes (ver Tabla 4-3). En el Departamento de Estadística e Informática la tasa de errores también presentan una cercana aproximación, lo cual nos señala que si estimamos los datos faltantes con el knn las tasa de errores serán similares. En el Departamento de Economía y Planificación la estimación del conjunto de datos por MRT es mas optimista que la estimación por KNN, lo cual nos indica que los datos faltantes del conjunto de datos puede estimarse con el MRT porque la tasa de errores se aproxima a los errores cuando se trabaja con los datos sin imputar.

Table 4–23: Tabla de errores de la validación cruzada con el MCPD-Determinístico para la data imputada con el knn.

Errores del MCPD-Determinístico usando data imputada con el knn			
Corte	Dep.Gestión	Dep.Economía	Dep. Estadística
11	7.588832 %	2.922078 %	30.73864 %
12	7.563452 %	5.166052 %	27.33333 %
13	2.787135 %	3.289474 %	31.38095 %

Utilizando el MCPD-Clustering jerárquico y estimando los valores faltantes del conjunto de datos con el knn, el resultado nos muestra que en el Departamento de Gestión Empresarial la tasa de errores se reduce en 1% en comparación de la reducción cuando la imputación de datos se realiza utilizando el mrt con el 6%. Para los departamentos académicos de Estadística y Economía la tasa de errores nos indica que es indiferente realizar una imputación por el método de mrt o el knn.

Table 4–24: Tabla de errores de la validación cruzada con el MCPD-Clustering jerárquico para la data imputada con el knn.

Tabla de Errores del MCPD-Clustering jerárquico usando la imputación knn			
Método	Dep. Estadística	Dep.Economía	Dep.Gestión
Hclust	31.57143 %	3.114187 %	9.727273%

4.7.2 Método de resubstitución para validar el MCPD en la base de datos imputados con el KNN

Con el conjunto de datos imputados utilizando el knn, y modelando con el MCPD para luego validar el modelo con el método de resubstitución obtenemos la Tabla 4-13. Se epuede ver para el Departamento de Estadística e Informática cuando el corte es once se reduce en aproximadamente 3% en comparación del 8% cuando la estimación de los datos faltantes se realiza con el mrt, para los demás cortes de ese mismo departamento existe una reducción en la tasa de errores cuando se utiliza el KNN o el MRT. En el Departamento de Economía y Planificación, cuando el corte

es 13 hay una variación del 7% cuando la imputación se realiza con el knn que con el MRT; y por último, para el Departamento de Gestión Empresarial cuando el corte es once, se puede ver que no existe una variación significativa comparando la tasa de errores con el conjunto de datos faltantes o con la imputación knn.

Table 4–25: Tabla de errores utilizando el método de resubstitución para MCPD-Determinístico con la data imputada mediante el knn.

Errores del MCPD-Determinístico para los datos Imputados con el knn			
Corte	Dep. Gestión	Dep. Economía	Dep. Estadística
11	6.062418 %	11.65584 %	11.93182 %
12	6.064126 %	11.51292 %	10 %
13	2.657143 %	12.82895 %	11.71429%

Cuando se modela con el MCPD-Clustering jerárquico con los datos imputados por el knn, se puede en la Tabla 4-14 que la tasa de error es menor que si se hubiera trabajado con los datos faltantes para el Departamento de Estadística. Esta variación es similar en el Departamento de Gestión Empresarial. Sin embargo, para el Departamento de Economía, la tasa de error nos indica que es mas beneficioso no estimar los datos faltantes porque mejora la reducción en la tasa de errores.

Table 4–26: Tabla de errores utilizando el método de resubstitución para MCPD con el clustering jerárquico usando la data imputada con el knn.

Errores del MCPD-Clustering jerárquico con los datos imputados knn			
Método	Dep. Estadística	Dep.Economía	Dep.Gestión
Hclust	13.49206 %	11.93772 %	7.275758 %

CAPITULO 5
CONCLUSIONES

5.1 Introducción

En este trabajo de investigación se diseñó un modelo de clasificación y predicción en dos etapas (MCPD), de dos formas diferentes: MCPD-Determinístico y el MCPD-Clustering jerárquico. Ambos tienen como objetivo principal predecir si un estudiante que esta cursando sus estudios universitarios culminará de manera satisfactoria su programa académico de estudios en los años establecidos por la entidad educativa. Este modelo fue aplicado en tres fases diferentes: Primero, usando la base de datos original; segundo, imputando los datos faltantes con la metodología del árbol de regresión multivariada (MRT) y tercero, con la imputación mediante la técnica de los k vecinos más cercanos. Finalmente, en cada una de las fases mencionadas se realizó la validación de los modelos mediante el método de resubstitución y la validación cruzada.

5.2 Conclusiones

- El modelo de clasificación y predicción en dos etapas (MCPD), nos señala que los cursos que el estudiante lleva después de su primer año académico son de vital importancia para estimar si éste terminará o no sus estudios universitarios, tomando en cuenta que el alumno obtuvo un rendimiento óptimo en su primer año de estudios académicos, es decir si paso a la segunda etapa de clasificación.

- El MCPD-Clustering jerárquico nos muestra una tasa de mala clasificación que es menor comparándolo con el El MCPD-Determinístico en el departamento de Estadística e Informática. Esto ocurre cuando validamos el modelo propuesto con la técnica de la validación cruzada.
- El MCPD-Determinístico muestra como resultado la tasa de error mas pequeña en comparación del MCP-Clustering jerárquico, cuando la validación del modelo se realizó con el método de resubstitución cuando se realizó el corte en doce para los tres departamentos analizados.
- Cuando se realiza la imputación de los datos faltantes a los estudiantes que pasaron a la segunda etapa de clasificación, se concluye que la imputación por los k vecinos más cercanos (KNN) da como resultado una tasa de mala clasificación menor en comparación que la tasa de error por la imputación realizada con el MRT.
- Imputando los datos con el árbol de regresión multivariada (MRT) se puede ver que la tasa de error es más pequeña cuando se utiliza el MCPD-Determinístico para los tres departamentos académicos. Esto sucede si la validación del modelo fue realizado por el método de resubstitución cuando el corte realizado es de once. Sin embargo, cuando se valida con la técnica de la validación cruzada nos señala que la tasa de errores es mas pequeña en el departamento de Gestión empresarial para los tres cortes analizados.

APENDICES

APENDICE A
ALGORITMO DEL MODELO DE CLASIFICACIÓN Y PREDICCIÓN EN DOS ETAPAS (MCPD)

```
#****************************************
#DEPARTAMENTO DE ESTADISTICA E INFORMATICA
#****************************************
dataEST
str(dataEST)
data_NA=dataEST[,c(1,2,3,4,5,6,10,11,13,14,15,19,22,23,30,33,35,39)]
head(data_NA)
str(data_NA)
particion<-function(k)
{ c=c(rep(0,k))
for(i in 1:k)
{ if(sum(which(data_NA[i,]==111))>0)
{
c[i]=i
prueba1=data_NA[-c,]}
}
entren=prueba1
#str(entren)
```

```
prueb=data_NA[c,]
#str(prueb)
list(data_entren=entren,data_prueb=prueb,
st1=str(entren),st2=str(prueb),
l1=length(entren[,1]),l2=length(prueb[,1]))
}
k=length(data_NA[,1])
L<-particion(k)
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
entren_EST<-L$data_entren
prueb_EST<-L$data_prueb
str(prueb_EST)
str(entren_EST)
attach(entren_EST)
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
modelo_EST<-mvpart(data.matrix(entren_EST[,c(7:18)])~
ADM+CAL+ECON+INTRO+MAT_B+MAT_C,entren_EST)
modelo_EST
pred_EST<-predict(modelo_EST,prueb_EST,type="matrix")
str(pred_EST)
head(pred_EST)
pred_EST2=as.data.frame(pred_EST)
str(pred_EST2)
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
prueb_EST3<-prueb_EST[,c(7:18)]
head(prueb_EST3)
str(prueb_EST3)
```

```
k1<-length(prueb_EST3[,2])
k1
for(i in 1:k1) #filas
for(j in 1:12) #columnas{
if(prueb_EST3[i,j]==111){
prueb_EST3[i,j]=pred_EST2[i,j]}
}
str(prueb_EST3)
head(prueb_EST3)
estim_EST=cbind(prueb_EST[,c(1:6)],prueb_EST3[,c(1:12)])
estim_EST
EST_COMPL=rbind(entren_EST,estim_EST)
EST_COMPL
length(EST_COMPL[,2])
str(EST_COMPL)
head(EST_COMPL)
summary(EST_COMPL)
^^^^^^^^^^^^^^MCPD_METODOLOGIA_DETERMINISTICA^^^^^^^^^^^^^^^^^
First_Stage<-function(data,punt)
{ n<-length(data[,1])
prom<-c(rep(0,n))
for(i in 1:n)
{
v1=data[i,1]
v2=data[i,2]
v3=data[i,3]
v4=data[i,4]
```

```
v5=data[i,5]
v6=data[i,6]
promE<-(v1+v2+v3+v4+v5+v6)/6
prom[i]<-promE
}
aprov<-which(prom>=punt)
num=length(aprov)
list(NroAprov=num,aprovados=aprov)
}
pec1=First_Stage(EST_COMPL,11)$aprovados
length(pec1)
pec1
dataE1c<-EST_COMPL[pec1,]
attach(dataE1c)
modelE1<-rpart(GRAD~BASE+GEST+MODLOS+TM1+SERIES+SOBREV
+TP1+CAL_EST+PROB+AP1+EST_GEN,dataE1c,control=rpart.control(minbucket=2))
modelE1
prp(modelE1,box.col=c("pink", "palegreen3"),nn=F,type=3,
 extra=4,main="MCPD-Determinstica:Departamento Economa y planificacin" )
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
pre1c<-predict(modelE1,type="class")
length(pre1c)
ori1c<-EST_COMPL[pec1,18]
ori1c
aciertos1c<-which(ori1c==pre1c)
length(aciertos1c)
TMC_Ec1<-1-length(aciertos1c)/length(ori1c)
```

```
TMC_Ec1
error_dprep_est1<-crossval(EST_COMPL[pec1,c(7:18)],method="rpart",repet=10)
error_dprep_est1
#**********hclust***********
hc3<- hclust(dist(EST_COMPL[,c(1:6)])^2, "ave")
hc3
memb3 <- cutree(hc3, k =4)
memb3
length(memb3)
pec3<-which(memb3==1)
length(pec3)
dataE3c<-EST_COMPL[pec3,]
attach(dataE3c)
str(dataE3c)
modelE3c<-rpart(GRAD~BASE+GEST+MODLOS+TM1+SERIES+SOBREV+
TP1+CAL_EST+PROB+AP1+EST_GEN,
dataE3c,control=rpart.control(minbucket=2))
modelE3c
prp(modelE3c,box.col=c("pink", "palegreen3"),nn=F,type=3, extra=4,
main="MCPD-Hclust:Departamento Economa y planificacin" )
pre3c<-predict(modelE3c,type="class")
length(pre3c)
ori3c<-EST_COMPL[pec3,18]
ori3c
aciertos3c<-which(ori3c==pre3c)
length(aciertos3c)
TMC_Ec3<-1-length(aciertos3c)/length(ori3c)
```

```
TMC_Ec3
error_dprep_est3<-crossval(EST_COMPL[pec3,],method="rpart",repet=10)
error_dprep_est3



#****************************************
#DEPARTAMENTO ECONOMIA Y PLANIFICACION
#****************************************
dataEC
str(dataEC)
data_NA=dataEC[,c(1,2,3,4,5,6,39,9,10,11,13,14,15,17,18,19,20,21,22
,23,24,26,27,30,29,35,36,40)]
head(data_NA)
str(data_NA)
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
particion<-function(k)
{ c=c(rep(0,k))
for(i in 1:k)
{
if(sum(which(data_NA[i,]==111))>0){
c[i]=i
prueba1=data_NA[-c,]}
}
entren=prueba1
#str(entren)
prueb=data_NA[c,]
#str(prueb)
```

```
list(data_entren=entren,data_prueb=prueb,st1=str(entren)
,st2=str(prueb),l1=length(entren[,1]),l2=length(prueb[,1]))
}
k=length(data_NA[,1])
L<-particion(k)
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
entren_ECO<-L$data_entren
prueb_ECO<-L$data_prueb
str(prueb_ECO)
str(entren_ECO)
attach(entren_ECO)
modelo_ECO<-mvpart(data.matrix(entren_ECO[,c(7:28)])~ADM+DIF+ECON
+MAT_FIN+MAT_BAS,entren_ECO)
modelo_ECO
pred_ECO<-predict(modelo_ECO,prueb_ECO,type="matrix")
str(pred_ECO)
head(pred_ECO)
pred_ECO2=as.data.frame(pred_ECO)
str(pred_ECO2)
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
prueb_ECO3<-prueb_ECO[,c(7:28)]
head(prueb_ECO3)
str(prueb_ECO3)
k1<-length(prueb_ECO3[,2])
k1
for(i in 1:k) #filas
for(j in 1:22) #columnas{
```

```
if(prueb_ECO3[i,j]==111)
{prueb_ECO3[i,j]=pred_ECO2[i,j]}
}
str(prueb_ECO3)
head(prueb_ECO3)
estim_ECO=cbind(prueb_ECO[,c(1:6)],prueb_ECO3[,c(1:22)])
estim_ECO
ECO_COMPL=rbind(entren_ECO,estim_ECO)
ECO_COMPL
length(ECO_COMPL[,2])
str(ECO_COMPL)
head(ECO_COMPL)
summary(ECO_COMPL)
^^^^^^^^^^^^^^^MCPD_METODOLOGIA_DETERMINISTICA^^^^^^^^^^^^^^^^
First_Stage<-function(data,punt)
{
n<-length(data[,1])
prom<-c(rep(0,n))
prom

for(i in 1:n)
{
v1=data[i,1]
v2=data[i,2]
v3=data[i,3]
v4=data[i,4]
v5=data[i,5]
```

```
v6=data[i,6]
promE<-(v1+v2+v3+v4+v5+v6)/6
prom[i]<-promE
}
aprov<-which(prom>=punt)
num=length(aprov)
list(NroAprov=num,aprovados=aprov)
}
pec1=First_Stage(ECO_COMPL,13)$aprovados
length(pec1)
pec1
dataE1c<-ECO_COMPL[pec1,]
attach(dataE1c)
modelEc11<-rpart(GRAD~C_GEN+C_GER+DE+D_EMP+METRIA+
E_REG+E_MA+ESP+MACR1+TC_D+FIN_PUB,dataE1c
,control=rpart.control(minbucket=1))
modelEc11
prp(modelEc11,box.col=c("pink", "palegreen3"),nn=F,type=3, extra=4
,main="MCPD-Determinstica:Departamento Economa y planificacin" )
pre1c<-predict(modelEc11,type="class")
length(pre1c)
ori1c<-ECO_COMPL[pec1,28]
ori1c
aciertos1c<-which(ori1c==pre1c)
length(aciertos1c)
TMC_Ec1<-1-length(aciertos1c)/length(ori1c)
TMC_Ec1
```

```
error_dprep_eco1<-crossval(ECO_COMPL[pec1,c(7:28)],method="rpart",repet=10)
error_dprep_eco1

^^^^^^^^^^^^^MCPD_ANALISIS DE CLUSTER^^^^^^^^^^^^^^^
hc3<- hclust(dist(ECO_COMPL[,c(1:6)])^2, "ave")
hc3
memb3 <- cutree(hc3, k =4)
memb3
length(memb3)
pec3<-which(memb3==1)
length(pec3)
dataE3c<-ECO_COMPL[pec3,]
attach(dataE3c)
str(dataE3c)
modelE3c<-rpart(GRAD~C_GEN+C_GER+DE+D_EMP+
METRIA+E_REG+E_MA+ESP+MACR1+TC_D+FIN_PUB
,dataE3c,control=rpart.control(minbucket=2))
modelE3c
prp(modelE3c,box.col=c("pink", "palegreen3"),nn=F,type=3
,extra=4,main="MCPD-Hclust:Departamento Economa y planificacin" )
pre3c<-predict(modelE3c,type="class")
length(pre3c)
ori3c<-ECO_COMPL[pec3,28]
ori3c
aciertos3c<-which(ori3c==pre3c)
length(aciertos3c)
TMC_Ec3<-1-length(aciertos3c)/length(ori3c)
```

```
TMC_Ec3
error_dprep_eco3<-crossval(ECO_COMPL[pec3,c(7:28)],method="rpart",repet=10)
error_dprep_eco3


#****************************************
#DEPARTAMENTO DE GESTION EMPRESARIAL
#****************************************
dataG
str(dataG)
data_NA=dataG[,c(1,2,3,4,5,6,36,24,25,26,27,28,29,30,31,32,34,35
,37,38,39,40,41,42,44)]
head(data_NA)
str(data_NA)
particion<-function(k)
{ c=c(rep(0,k))
for(i in 1:k)
{
if(sum(which(data_NA[i,]==111))>0){
c[i]=i
prueba1=data_NA[-c,]  }
}
entren=prueba1
#str(entren)
prueb=data_NA[c,]
#str(prueb)
list(data_entren=entren,data_prueb=prueb,st1=str(entren),
st2=str(prueb),l1=length(entren[,1]),l2=length(prueb[,1]))
```

```
}
k=length(data_NA[,1])
L<-particion(k)
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
entren_GES<-L$data_entren
prueb_GES<-L$data_prueb
str(prueb_GES)
str(entren_GES)
attach(entren_GES)
modelo_GES<-mvpart(data.matrix(entren_GES[,c(7:24)])~ADM+LEN
+MAT_B+MAT_FI+AD_RH+CAL_D,entren_GES)
modelo_GES
pred_GES<-predict(modelo_GES,prueb_GES,type="matrix")
str(pred_GES)
head(pred_GES)
pred_GES2=as.data.frame(pred_GES)
str(pred_GES2)
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
prueb_GES3<-prueb_GES[,c(7:25)]
head(prueb_GES3)
str(prueb_GES3)
k1<-length(prueb_GES3[,2])
k1
for(i in 1:k) #filas
for(j in 1:19) #columnas
{
if(prueb_GES3[i,j]==111){
```

```
prueb_GES3[i,j]=pred_GES2[i,j]}
}
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
str(prueb_GES3)
head(prueb_GES3)
estim_GES=cbind(prueb_GES[,c(1:6)],prueb_GES3[,c(1:19)])
estim_GES
GES_COMPL=rbind(entren_GES,estim_GES)
GES_COMPL
length(GES_COMPL[,2])
str(GES_COMPL)
head(GES_COMPL)
^^^^^^^^^^MCPD_METODOLOGIA_DETERMINISTICA^^^^^^^^^^^
First_Stage<-function(data,punt)
{
n<-length(data[,1])
prom<-c(rep(0,n))
prom
for(i in 1:n)
{
v1=data[i,1]
v2=data[i,2]
v3=data[i,3]
v4=data[i,4]
v5=data[i,5]
v6=data[i,6]
promE<-(v1+v2+v3+v4+v5+v6)/6
```

```
prom[i]<-promE
}
aprov<-which(prom>=punt)
num=length(aprov)
list(NroAprov=num,aprovados=aprov)
}
peg=First_Stage(GES_COMPL,13)$aprovados
length(peg)
peg
dataG1g<-GES_COMPL[peg,]
attach(dataG1g)
modelG1g<-rpart(GRAD~AFIN+C_GER+ECON+FI1+INV_O+
MARK+NEG_INT+ORG_MET+PLA_EST+T_PESQ
,dataG1g,control=rpart.control(minbucket=1))
modelG1g
prp(modelG1g,box.col=c("pink", "palegreen3"),nn=F,type=3
,extra=4,main="MCPD-Determinstica:Departamento Gestin Empresarial" )
pre1g<-predict(modelG1g,type="class")
length(pre1g)
ori1g<-GES_COMPL[peg,25]
length(ori1g)
cbind(pre1g,ori1g)
aciertos1g<-which(ori1g==pre1g)
length(aciertos1g)
ECD_Gg<-1-length(aciertos1g)/length(ori1g)
ECD_Gg
error_dprep_gesk<-crossval(GES_COMPL[peg,c(7:25)],method="rpart",repet=10)
```

```
error_dprep_gesk
^^^^^^^^^^^^^MCPD_ANALISIS DE CLUSTER^^^^^^^^^^^^^^^^
pxg<-which(GES_COMPL[,44]=="SI")
pxg
#**********hclust***********

hc2g <- hclust(dist(GES_COMPL[,c(1:6)])^2, "ave")
hc2g
memb2g <- cutree(hc2g, k =4)
memb2g
length(memb2g)
p22<-which(memb2g==1)
length(p22)
dataG2g<-GES_COMPL[p22,]
#attach(dataG2g)
#str(dataG2g)

modelG2g<-rpart(GRAD~AFIN+C_GER+ECON+FI1+INV_O+
MARK+NEG_INT+ORG_MET+PLA_EST+T_PESQ,dataG2g
,control=rpart.control(minbucket=3))
modelG2g
prp(modelG2g,box.col=c("pink", "palegreen3"),nn=F,type=3
,extra=4,main="MCPD-Hclust:Departamento Gestin empresarial" )
pred1g<-predict(modelG2g,type="class")
length(pred1g)
#indices
ori1g<-GES_COMPL[p22,25]
```

```
length(ori1g)

aciertos1g<-which(ori1g==pred1g)

#aciertos1g

TMC_E1g<-1-length(aciertos1g)/length(ori1g)

TMC_E1g

error_dprep_ges3<-crossval(GES_COMPL[p22,c(7:25)],method="rpart",repet=10)

error_dprep_ges3
```

APENDICE B
CURSOS DE CONCENTRACIÓN POR DEPARTAMENTOS

Table B–1: Cursos de concentración del Departamento de Estadística e informática

Departamento de Estadística e Informática	
Cursos	Pre requisito
Algebra Matricial.	Cálculo avanzado para estadística.
Análisis de datos categóricos	140 créditos
Análisis de series de tiempo	140 créditos
Análisis de sobrevivencia y confiabilidad	Modelos Lineales
Base de Datos.	Análisis y diseño de sistemas.
Cálculo avanzado para Estadística	Cálculo Integral
Cálculo de probabilidades	Estadística general y Cálculo avanzado para estadística
Control estadístico de la calidad.	Gestión de la calidad.
Estadística general	Cálculo diferencial
Estadística aplicada I.	Estadística general
Estadística aplicada II	Estadística aplicada I
Inferencia Estadística	Cálculo de probabilidades
Métodos numéricos y simulación	Técnicas de Programación 2
Modelos de optimización	Métodos Numéricos y simulación
Modelos lineales.	Inferencia estadística y Algebra lineal.
Técnicas de muestreo I	Metodología para la invetigación e innovación.
Técnicas de muestreo II	Técnicas de muestreo 1
Técnicas de programación I	Matemática para computación.
Técnicas de programación II	Técnicas de programación 1.

Table B–2: Cursos de concentración del Departamento de Economía y Planificación

Departamento de Economía y Planificación	
Cursos	Pre requisito
Algebra lineal.	Matemáticas para economistas
Contabilidad de costos.	Contabilidad general.
Econometría.	Estadística aplicada a la economía y negocios I y Microeconomía II, Macroeconomía II
Economía agraria.	Microeconomía y macroeconomía II.
Economís de la información.	Microeconomía II
Economía de la regulación.	Economía del bienestar
Economía de los recursos naturales.	Economía del bienestar y Macroeconomía.
Economía del bienestar.	Microeconomía II.
Economía del medio ambiente.	Economía de los recursos naturales
Estadística aplicada de la economía y a los negocios.	Estadística general
Evaluación social de proyectos.	Formulación y evaluación de proyectos.
Finanzas Públicas.	Macroeconomía II
Investigación de operaciones.	Estadística general.
Macroeconomía I.	Economía general y Cálculo diferencial
Macroeconomía II.	MacroeconomÍ
Matemática para economistas.	Cálculo integral.
Método de investigación económica.	160 créditos
Teoría Monetaria.	Macroeconomía II
Política económica.	Teoría del conocimiento y desarrollo.
Cálculo Diferencial.	Ninguno.
Economía General.	Ninguno.
Escuelas del pensamiento económico	Ninguno.
Matemática Básica.	Ninguno.
Matemáticas financieras.	Ninguno.

Table B–3: Cursos de concentración del Departamento de Gestión Empresarial

Departamento de Gestión Empresarial	
Cursos	Pre requisito
Análisis e Investigación de Mercados.	Marketing.
Cálculo Integral.	Cálculo Diferencial.
Contabilidad general.	Matemática básica.
Contabilidad de costos.	Contabilidad general.
Contabilidad gerencial.	Matemática financiera y contabilidad de costos
Desarrollo empresarial.	160 créditos
Dirección estratégica II.	Dirección estratégica I, contabilidad gerencial y planeamiento estratégico
Estadística aplicada a economy a los negocios I.	Estadística general
Estadística General.	Ninguno.
Finanzas I.	Contabilidad gerencial.
Finanzas II.	Finanzas I
Finanazas III.	Finanzas II
Formulación y evaluación de proyectos I.	120 créditos
Formulación y evaluación de proyectos II.	Formulación y evaluación de proyectos I.
Marketing.	Estadística General y microeconomía.
Microeconomía I.	Economía General.
Seminario de tesis.	160 créditos.
Teoría básica de la negociación.	Comunicación
Tecnología forestal y pesquera.	Tecnologá agroindustrial

REFERENCIAS BIBLIOGRÁFICAS

[1] UNESCO. Situación educativa de américa latina y el caribe: Hacia la eduación de calidad para todos al 2015. 2013.

[2] IESALC. Repitencia y deserción universitaria en américa latina. *UNESCO*, 2006.

[3] A.Hamann; T.Gylander y P.Chen. Developing seed zones and transfer guidelines with multivariate regression trees. *Tree genetics and genomes*, 7:399–408, 2011.

[4] C.Fraley y A.Raftery. Model-based clustering,discriminant analysis and density estimation. *Journal of the american statistical association*, 97:611–631, 2002.

[5] P.Haldar; I.Pavord; D.Shaw; M.Berry; M.Thomas; C.Brightling; A.Wardlow y R.Green. Cluster analysis and clinical asthma phenotypes. *American Journal of respiratory and critical care medicine*, 178:218–224, 2008.

[6] E.Deconinck; T.Hancock; D.Coomans; D.Massart y Y.Vander. Clasification of drugs in absorption classes using the clasification and regression trees. *Journal of Pharmaceutical and biomedical analysis*, 39:91–103, 2005.

[7] L.Breiman; H.Friedman; R.Olshen y Ch. Stone. *Classification and regression trees*. Academic PRESS, the wadsworth statistics-probability series edition, 1984.

[8] M.G.Kendall y W.R.Buckland. *A Dictionary of Statistical Terms*. London. Longman Group, 1976.

[9] S. Lang. *Algebra*. Addison-Wesley Publishing Company, third edition, 1993.

[10] R.Xu y D.C.Wunsh II. *Clustering*. IEE-Press Editorial Board, first edition, 2009.

[11] A.Jain; M.Murty y P.Flynn. Data clustering: A review. *ACM computing surveys*, 31:264–323, 1999.

[12] C.Bishop. Naural networks for pattern recognition. *PRESS.Oxford University*, 1995.

[13] J.Kleinberg. An impossibility theorem for clustering. *NIPS Conference on advances in neural information processing systems*, pages 463–470, 2002.

[14] M.Anderberg. *Probability and mathematical statistics*. Academic PRESS, first edition, 1973.

[15] A.Jain y R.Dubes. *Algorithms for clustering data*. Englewood diffs, third edition, 1988.

[16] G.De'ath. Multivariate regression trees: A new technique for modeling species-environment relationships. *Ecology*, 83:1105–1117, 2002.

[17] J.Gower y D.Hand. Biplots. *Monoghaphs on statistics and applied probability*, 1996.

[18] D.Faith; D.Minchin y L.Belbin. Extended dissimilarity: A method of robust estimation of ecological distances from high beta diversity. *Vegatatio*, 69:57–68, 1997.

[19] G.De'ath. Extended dissimilarity: A method of robust estimation of ecological distances from high beta diversity. *Plant ecology*, 144:191–199, 1999.

[20] P.Legendre y M.Anderson. Distance-based redundancy analysis: Testing multispecies responces in multifactorial ecological experiments. *Ecological Monographs*, 69:57–68, 1999.

[21] M.C.A. Soma. A Multi-Stage decision Algorithm for rule generation for minority class. *Texas Tech University*, 2014.

[22] H.C. DiGangi y S.Jannasch. A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, Vol. 8:307–325, 2010.

[23] V. Lam; I. Welch; X. Gao y P. Komisarczuk. Two-Stage Model to Detect Malicious Web Pages. *International Conference on Advanced Information Networking and Applications*, Vol. 83:1105–1117, 2011.

DATOS BIOGRAFICOS

Aca debes colocar la biografia en espanol.

En el archivo: `Biography.tex`

MODELO DE CLASIFICACIÓN Y PREDICCIÓN EN DOS ETAPAS: UTILIZANDO ÁRBOLES DE CLASIFICACIÓN Y EL ANÁLISIS DE REGRESIÓN MULTIVARIADA

Yency Edith Choque Dextre
(787) XXX-XXXX
Departamento de Matemáticas
Consejero: Edgar Acuña Fernández
Grado: Maestría en Ciencias
Fecha de Graduacion: Mayo 2015

Este es el resumen para la audiencia general.

En el archivo: `GeneralAudienceAbstract.tex`