ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

METODOLOGÍA PARA LA CLASIFICACIÓN SOCIOECONÓMICA DE LOS ESTUDIANTES DE LA ESCUELA POLITÉCNICA NACIONAL Y SU INCIDENCIA EN LA ASIGNACIÓN DE BECAS Y COBRO DE ARANCELES

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

PROYECTO DE INVESTIGACIÓN

WILSON EFRÉN ARROYO RUALES

w.arroyo.r@gmail.com

DIRECTOR: MAT. NELSON ALEJANDRO ARAUJO GRIJALVA, MSc.

alejandro.araujo@epn.edu.ec

QUITO, JUNIO 2017

DECLARACIÓN

Yo, Wilson Efrén Arroyo Ruales, declaro que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional, puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Wilson Efrén Arroyo Ruales

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Wilson Efrén Arroyo Ruales, bajo mi supervisión.

Mat. Nelson Alejandro Araujo Grijalva, MSc.

DIRECTOR DE PROYECTO

AGRADECIMIENTOS

Al Mat. Araujo por guiarme y auspiciarme en este proyecto.

A mis padres porque me han dado el ejemplo de trabajo, esfuerzo y dedicación.

A mi esposa Geoconda por todo su apoyo.

A mis amigos por sus consejos y recomendaciones.

A mis profesores por sus sabias enseñanzas.

Al Departamento de Matemática por toda la colaboración recibida.

A la Escuela Politécnica Nacional por las facilidades brindadas para llevar adelante este proyecto.

DEDICATORIA

Dedico este trabajo a mi familia y a mis padres.

CONTENIDO

RESUMEN 1
ABSTRACT 2
CAPÍTULO 1
INTRODUCCIÓN 3
CAPÍTULO 2
MARCO TEÓRICO 6
2.1 PRINCIPALES CONCEPTOS SOBRE CLASIFICACIÓN SOCIO ECONÓMICA 6
2.2 ANÁLISIS DE COMPONENTES PRINCIPALES CATEGÓRICO (ACPC) 10
2.2.1 Relación entre coordenadas principales y componentes principales 12
2.3 ANÁLISIS DE CONGLOMERADOS 13
2.3.1 Algoritmo de partición: K-MEDIAS 14
2.4 ÁRBOL DE CLASIFICACIÓN 15
2.4.1 Índices de asociación y medidas de impureza 16
2.4.2 Algoritmos para desarrollar árboles 18
2.5 REGRESIÓN LOGÍSTICA MULTINOMIAL 19
2.5.1 Modelos Logit para respuestas nominales 19
2.6 MAQUINAS DE VECTORES DE SOPORTE 20
2.6.1 El caso linealmente separable 21
2.6.2 El caso linealmente no separable 22
2.6.3 Uso de la funcion nucleo (KERNEL) para el caso linealmente NO separable 23
CAPÍTULO 3
METODOLOGÍA DE CONSTRUCCIÓN DEL ÍNDICE SOCIOECONÓMICO DE LOS ESTUDIANTES DE LA ESCUELA POLITÉCNICA NACIONAL: ANÁLISIS Y RESULTADOS 25
3.1 ANÁLISIS DESCRIPTIVO DE VARIABLES 25
3.1.1 Variables de caracterización del estudiante 25
3.1.2 Variables de Ingreso Familiar 28

3.1.3 Variables de patrimonnio familiar del estudiante 29

3.2 CONSTRUCCIÓN DEL ÍNDICE SOCIECONÓMICO 29

3.3.1 Cálculo del Índice de Nivel Socioeconómico (INS) 37

3.3.2 Estratos del INS 39

CAPÍTULO 4
ASIGNACIÓN DE ESTRATOS SOCIOECONÓMICOS PARA ALUMNOS NUEVOS DE PREGRADO DE LA EPN 43

4.1 ÁRBOL DE DECISIÓN 43

4.2 REGRESIÓN LOGÍSTICA MULTINOMIAL 45

4.3 MÁQUINAS DE VECTORES DE SOPORTE 47

4.4 COMPARACIÓN DE LOS MÉTODOS DE CLASIFICACIÓN 49

CAPÍTULO 5 51

INCIDENCIA DE LOS ESTRATOS SOCIO ECONÓMICOS EN LA ASIGNACIÓN DE BECAS Y COBRO DE ARANCELES 51

CONCLUSIONES Y RECOMENDACIONES 59

CONCLUSIONES 59

RECOMENDACIONES 60

BIBLIOGRAFÍA 61

ANEXOS 64

ANEXO A.
CONCEPTOS DE COMPONENTES PRINCIPALES CATEGÓRICOS 64

ANEXO B.
ALGORITMO DE PARTICIÓN: K-MEDIAS 67

ANEXO C.
CONCEPTOS DE ÁRBOLES DE CLASIFICACIÓN 70

ANEXO D.
MODELOS ACUMULADOS PARA DATOS ORDINALES 74

ANEXO E.
TEOREMA KHUN – TUCKER Y NÚCLEOS 76

ANEXO F.
CÓDIGOS PARA GENERACIÓN DE MODELOS APLICADOS 78

F.1 CÓDIGO PARA GENERACIÓN DEL ACPC EN SPSS 78

F.2. CÓDIGO PARA GENERACIÓN DEL K – MEDIAS EN SPSS 78

F.3. CÓDIGO PARA GENERACIÓN DEL ÁRBOL DE CLASIFICACIÓN EN SPSS 79
F.4. CÓDIGO PARA GENERACIÓN DE REGRESIÓN LOGÍSTICA MULTINOMIAL EN SPSS 79
F.5. CÓDIGO PARA GENERACIÓN DEL SVM EN R 80
ANEXO G.
COMPARACIÓN DE QUINTILES : INGRESO PER CAPITA Y EQUIVALENTE 82

ÍNDICE DE TABLAS

Tabla 1. Variables de ingreso familiar y del estudiante 28
Tabla 2. Estadísticos descriptivos de las variables de patrimonio familiar del estudiante 29
Tabla 3. Resultados del Modelo Inicial 30
Tabla 4. Saturación en componentes del Modelo Inicial 32
Tabla 5. Resumen del Modelo Final 33
Tabla 6. Puntuaciones de saturación de las componentes 34
Tabla 7. Conglomerados a partir del algoritmo K-Medias con 5 grupos 39
Tabla 8. Conglomerados a partir del algoritmo K-Medias con 8 grupos 40
Tabla 9. Análisis descriptivo del INS por grupo 41
Tabla 10. Tabla de confusión estratos INS vs Árbol de clasificación 45
Tabla 11. Coeficientes de la regresión logística 46
Tabla 12. Tabla de confusión de la regresión logística multinomial 46
Tabla 13. Tabla de clasificación del SVM vs INS 49
Tabla 14. Tabla de comparación de métodos de clasificación automática 49
Tabla 15. Comparación de estratificaciones: Quintiles EPN – INS (número de casos) 54
Tabla 16. Comparación de estratificaciones: Quintiles EPN – INS (Porcentaje) 54
Tabla 17. Becas asignadas en la EPN por INS 55
Tabla 18. Becas asignadas en la EPN por QUINTIL 56
Tabla 19. Becas económicas por Quintiles EPN – INS 56
Tabla 20. Comparación de asignación de otras becas: Quintiles EPN - INS 57
Tabla 21. Pago de aranceles promedio en la EPN por INS 58
Tabla 22. Miembros de un hogar y su equivalencia en gasto 82
Tabla 23. Quintiles usando los miembros equivalentes: 83
Tabla 24. Quintil EPN vs Quintil Equivalente 83
Tabla 25. INS vs Quintil Equivalente 83
Tabla 26. Becas por Quintil Equivalente 84

ÍNDICE DE FIGURAS

Figura 1. Ejemplo del agrupamiento K-Medias 14
Figura 2. Ejemplo de un árbol de clasificación 15
Figura 3. Ejemplo de un árbol binario 16
Figura 4. Distribución de los estudiantes por provincia 26
Figura 5. Estado civil del estudiante y colegio de procedencia 26
Figura 6. Número de miembros del hogar del estudiante 27
Figura 7. Relación laboral del estudiante 27
Figura 8. Saturación en componentes para el Modelo Inicial 31
Figura 9. Gráfico de saturaciones Componente 1 vs Componente 2 35
Figura 10. Gráfico de saturaciones Componente 1 vs Componente 3 36
Figura 11. Gráfico de saturaciones Componente 2 vs Componente 3 37
Figura 12. Distribución de la población en los grupos con el algoritmo K-Medias por el promedio del INS 41
Figura 13. Población estudiantil por Grupos e INS 42
Figura 14. Árbol para clasificación de alumnos nuevos 44
Figura 15. Gráfico de dispersión entre Ingreso neto familiar y total de propiedades por clasificación SVM 48
Figura 16. Estratos socioeconómicos definidos por el INEC 51
Figura 17. Estratos socioeconómicos definidos por el INS 52
Figura 18. Importancia de las variables en la estratificación del INEC 53

RESUMEN

En el Ecuador el estado garantiza la gratuidad en la educación pública hasta el tercer nivel, en este contexto el Consejo de Educación Superior (CES) expide el REGLAMENTO PARA GARANTIZAR EL CUMPLIMIENTO DE LA GRATUIDAD DE LA EDUCACIÓN SUPERIOR PÚBLICA, el cual es de cumplimiento obligatorio y contiene la norma para la asignación de becas y el cobro de aranceles cuando un estudiante ha perdido la gratuidad de la educación pública.

Así, en este proyecto se ha definido una metodología para la estratificación socio-económica de los estudiantes de la Escuela Politécnica Nacional (EPN); para lograr este objetivo se utilizan el Análisis de Componentes Principales Categórico (ACPC) y el Cluster K-Medias que definen un índice socio económico y los correspondientes estratos, de tal manera que cada estudiante pertenece a uno de ellos. Con esta asignación y usando las técnicas de clasificación automática: Arboles de clasificación, Regresión logística multiclase y Maquinas de vectores de soporte se asigna a los estudiantes nuevos a uno de los estratos previamente establecidos por el índice construido.

Palabras claves: Análisis de Componentes Principales Categórico, Algoritmos K-Medias, Índice de Estratificación Socioeconómico, Becas, Matrículas.

ABSTRACT

In Ecuador, the State guarantees free public education up to the third level, in this context, the Council of Higher Education (CES, in Spanish) issues the REGULATION TO ENSURE THE COMPLIANCE OF THE LIBERTY OF PUBLIC HIGHER EDUCATION, which is mandatory and contains the standard for the allocation of scholarships and the collection of fees when a student has lost the gratuity of public education.

Thus, in this project has defined a methodology for the socio-economic stratification of the National Polytechnic School's Students (EPN, in Spanish), to achieve this objective, the Principal Component Analysis Categorical (ACPC) and Cluster K-Means are used to define a socio-economic index and their strata, in such a way that each student belongs to one of them. With this and using the automatic classification techniques: Classification Trees, Multiclass Logistic Regression and Support Vector Machines is assigned to new students in each of the strata previously established by the index.

Keywords: Categorial Principal Component Analysis, K-Means Algorithm, Socioeconomic Stratification Index, Scholarships, Fees.

CAPÍTULO 1
INTRODUCCIÓN

La Escuela Politécnica Nacional es una universidad pública regida por la Ley Orgánica de Educación Superior (LOES). Siendo universidad pública debe considerar también las disposiciones y reglamentos emitidos por el Consejo de Educación Superior (CES); por ejemplo, el Reglamento para garantizar el cumplimiento de la Gratuidad de la Educación Superior Pública (REGES) aprobado por el Pleno del CES mediante Resolución RPC-SO-25-No.258-2014, el cual establece en sus disposiciones generales primera y segunda lo siguiente:

> "PRIMERA.- En caso de pérdida temporal o definitiva de la gratuidad, las instituciones de educación superior sujetas al presente Reglamento podrán cobrar por concepto de matrículas y aranceles entre el diez por ciento (10%) y cincuenta por ciento (50%) del valor recibido por la institución de educación superior por cada estudiante en función del costo óptimo por tipo de carrera y modalidad de aprendizaje, en el año inmediato anterior a la matrícula, considerando obligatoriamente la situación socio-económica del estudiante y su hogar. La información relativa al costo óptimo será suministrada por la SENESCYT a las instituciones de educación superior.
>
> SEGUNDA.- Las instituciones de educación superior pública deberán observar de manera estricta el presente Reglamento. En caso de expedir normas internas, estas no podrán encontrarse en contradicción con este Reglamento."
>
> (CES, 2014)

Por lo tanto, es obligatorio que cualquier medida económica que incida en los estudiantes considere la situación socio económica del hogar del estudiante. Un ejemplo de los beneficios económicos a los que un estudiante se puede acoger es la beca estudiantil; por otro lado, con respecto a las sanciones se encuentra el incremento de costos y aranceles ocasionados cuando el estudiante pierde una o varias materias.

De acuerdo al Servicio de Bienestar Estudiantil de la EPN, existe una clasificación socioeconómica para los estudiantes de pregrado. En esta clasificación se han definido quintiles, ordenados de menor a mayor según valor del ingreso per cápita familiar calculado a

partir de dividir el ingreso familiar (declarado por el estudiante) para el número de miembros del hogar; así se tiene que los estudiantes en el primer quintil pagan menos que los del segundo quintil, y así sucesivamente. Sin embargo, se han detectado estudiantes asignados al último quintil que muestran características que los ubicarían en algún quintil inferior.

Por otro lado, a partir del año 2012 solamente los estudiantes que aprobaron el Examen Nacional de Educación Superior (ENES) son admitidos en una universidad pública. Antes, cada universidad tenía a su cargo la admisión de estudiantes nuevos.

Adicionalmente, la Constitución de 2008 garantiza la gratuidad de los estudios hasta el tercer nivel:

> "Articulo 28.- La educación pública será universal y laica en todos sus niveles, y gratuita hasta el tercer nivel de educación superior inclusive."
>
> (Asamblea Nacional Constituyente, 2008)

Ahora, al proponer y aprobarse una nueva clasificación socioeconómica se afectaría el cobro de aranceles, así como la asignación de becas. Este trabajo propone una metodología para encontrar grupos homogéneos de estudiantes considerando sus características socioeconómicas que permitiría realizar la asignación de becas y el cobro de aranceles de forma técnica.

Para determinar la clasificación propuesta en este trabajo, se van a utilizar variables sociales, económicas y demográficas de los estudiantes de pregrado de la EPN. Se utilizarán técnicas de estadística descriptiva, de agrupamiento y de clasificación; de tal forma que se ubiquen de la mejor manera posible a todos los estudiantes dentro de los diferentes grupos socioeconómicos presentes en la población estudiantil. Una vez determinados los grupos, se utilizan técnicas de predicción que ubicarán dentro de un grupo socioeconómico a los estudiantes nuevos.

La clasificación encontrada en este trabajo permitiría a la EPN enfocar la asignación de becas hacia los individuos con mayor necesidad socioeconómica; y, por otro lado, cobrar aranceles de forma diferenciada.

En el Capítulo 2 de este trabajo se presentan las definiciones básicas y necesarias para comprender la construcción y estimación de la clasificación socioeconómica de los estudiantes de la EPN; se definirán conceptos como: hogar, núcleo familiar, pobreza multidimensional, vivienda, los estratos económicos, ingresos familiares y per cápita, patrimonio familiar y otros. Por otro lado, se presentan las técnicas estadísticas a utilizarse, primero el Análisis de Componentes Principales Categórico (ACPC), que permitirá encontrar los factores que intervendrán en la construcción de un índice de nivel socioeconómico para dar un ordenamiento de los estudiantes. Luego, se definirán grupos de estudiantes por nivel socioeconómico a través de los conglomerados utilizando el algoritmo K-Medias. Finalmente, se desarrollará la definición y explicación de los procedimientos de clasificación multiclase: Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés), Árbol de Decisión y Regresión Logística Multinomial; que servirán para determinar la asignación de los estudiantes nuevos a los niveles socioeconómicos determinados anteriormente.

En el Capítulo 3 se desarrolla la construcción y estimación de un índice socioeconómico y se determinan los niveles correspondientes y se presentan los resultados encontrados sobre los niveles de clasificación que se propone para la EPN.

En el capítulo 4 se define la metodología a utilizarse para ubicar a un estudiante que ingresa a la universidad y se realiza el análisis de todas las implicaciones relacionadas con los niveles de clasificación determinados y el impacto que tendrán para la EPN en el cobro de aranceles y la asignación de becas.

Luego en la sección de Conclusiones y Recomendaciones, se exponen los principales aportes del trabajo y los posibles temas de investigación para continuarlo, de manera que puedan contribuir a conocer las dificultades encontradas en el desarrollo de este trabajo y mejorar la metodología para la clasificación.

CAPÍTULO 2
MARCO TEÓRICO

2.1 PRINCIPALES CONCEPTOS SOBRE CLASIFICACIÓN SOCIO ECONÓMICA

Los principales conceptos se detallan a continuación:

- **Vivienda.** Es el espacio delimitado por paredes y techo, de cualquier material de construcción, con entrada independiente, destinada para ser habitado por una o más personas; la misma que aunque no haya sido construida originalmente para tales fines, sea utilizada como vivienda (Instituto Ecuatoriano de Estadística y Censos, INEC 2010)

- **Hogar.** Es la unidad social conformada por una persona o grupo de personas que se asocian para compartir el alojamiento y la alimentación. Es decir, hogar es el conjunto de personas que residen habitualmente en la misma vivienda o en parte de ella (viven bajo el mismo techo), que están unidas o no por lazos de parentesco, y que cocinan en común para todos sus miembros (INEC 2010).

- **Seguro Social**. Es la cobertura de los sistemas de seguros de salud, por los cuales generalmente se paga una cuota mensual o periódica. El Seguro de Salud es el derecho que tienen o adquieren los miembros del hogar para el cuidado de su salud, sea en centros públicos o privados como son el Instituto de Seguridad Social, IESS que comprende el Seguro General y el Seguro Campesino; el seguro de las Fuerzas Armadas, ISSFA; Instituto de Seguridad Social de la Policía, ISSPOL (INEC 2010).

- **Vehículos de uso exclusivo**. Se refiere al número de vehículos que dispone el hogar para uso exclusivo, con su respectivo año, tipo/clase y valor comercial (INEC 2010).

- **Casa/ villa.** Es toda construcción permanente hecha con materiales resistentes, tales como: hormigón piedra, bloque, ladrillo, adobe, caña o madera. Por lo general tiene más de un cuarto, tumbado, abastecimiento de agua y servicio higiénico exclusivo (INEC 2010).

- **Departamento en casa o edificio.** Es un conjunto de cuartos que forman parte de un edificio de uno o más pisos, se caracteriza por ser independiente y generalmente, tiene abastecimiento de agua y servicio higiénico de uso exclusivo (INEC 2010).

- **Suite de lujo.** Es el conjunto de sala, alcoba y cuarto de baño, suelen brindar más espacio que las habitaciones tradicionales y suelen incluir mayor mobiliario (mesas, sillas, etc.). Dicha denominación suele implicar alojamiento de alta categoría (INEC 2010).

- **Cuarto(s) en casa de inquilinato.** Comprende uno o varios cuartos pertenecientes a una casa, con una entrada común y directa desde un pasillo, patio, corredor o calle, y generalmente no cuenta con servicio exclusivo de agua o servicio higiénico. Los cuartos pueden ser propios, cedidos, arrendados o recibidos por servicios (INEC 2010).

- **Mediagua.** Es una construcción de un solo piso, con paredes de ladrillo, adobe, bloque o madera con techo de teja, eternit, árdex o zinc y no tiene más de dos cuartos o piezas sin incluir cocina ni baño. Tiene como característica principal una sola caída de agua. Si tiene más de 2 cuartos considere como casa (INEC 2010).

- **Rancho.** Es una construcción rústica, cubierta con palma, paja, o cualquier otro material similar, con paredes de caña o bahareque y con piso de caña, madera o tierra, por lo habitual este tipo de vivienda se da en la Costa y en la Amazonía. En esta

categoría no entran los "ranchos" de las quintas ni fincas que generalmente tienen personas de ingresos altos, estos son considerados como casas (INEC 2010).

- **Covacha.** Es aquella construcción en la que se utiliza materiales rústicos tales como: ramas, cartones, restos de asbesto, latas, plásticos, etc., con piso de madera caña o tierra (INEC 2010).

- **Choza.** Es la construcción que tiene paredes de adobe, tapia o paja, con piso de tierra y techo de paja (INEC 2010).

- **Otro tipo de vivienda.** Son viviendas improvisadas o lugares no construidos para tales fines, como garajes, bodegas, furgones, carpas, casetas, conteiner, cuevas y otros (INEC 2010).

- **Servicio de Internet.** Se refiere al servicio de Internet utilizado dentro del hogar y será imprescindible que el hogar posea una computadora (INEC 2010).

- **Teléfonos celulares.** Hace referencia a que si el estudiante tiene uno o más celulares activados (INEC 2010).

- **Servicio de teléfono convencional.** Se entiende por tenencia de servicio telefónico, el acceso al servicio de telefonía fija, independientemente de la propiedad de la línea o del aparato (INEC 2010).

- **Correo Electrónico.** Se refiere a los estudiantes que utilizaron un correo electrónico personal, es decir que no sea de su trabajo. Es un servicio de red que permite a los usuarios enviar y recibir mensajes rápidamente (también denominados mensajes electrónicos) mediante sistemas de comunicación electrónicos. Su eficiencia, conveniencia y bajo coste (con frecuencia nulo) están logrando que el correo electrónico desplace al correo ordinario para muchos usos habituales (INEC 2010).

- **Miembros del hogar** (ENIGHUR[1] 2012). Se les considera Miembros de Hogar a las siguientes personas:

  a. Los residentes habituales que viven permanentemente en el hogar; es decir, que duermen la mayor parte del tiempo en él; incluyendo aquellos que se encuentran temporalmente ausentes por diferentes razones (trabajo, vacaciones, enfermedades, etc.), siempre que su ausencia sea por un período menor a seis meses.

  b. Las personas sin parentesco con el jefe del hogar o familiares de éste, que vivan habitualmente la mayor parte del tiempo en el hogar, siempre que no tengan otro lugar de residencia.

  c. Los servidores domésticos que son residentes habituales del Hogar y sus familiares que viven con ellos (puertas adentro).

  d. Personal de las Fuerzas Armadas destacado en el lugar y que vive habitualmente en el hogar la mayor parte del tiempo.

  e. Extranjeros que trabajan o estudian en el país desde hace seis meses, por lo menos, y que permanecerán viviendo la mayor parte del tiempo en el hogar en forma habitual.

  f. En el caso de la persona que sea reconocida como jefe (a) en dos o más hogares, deberá considerarla como miembro del hogar donde vive la mayor parte del tiempo.

  g. Las personas que estudian fuera de la ciudad donde vive el resto de la familia, serán considerados como miembros del hogar en las ciudades donde realizan sus

[1] ENIGHUR: Encuesta Nacional de Ingresos y Gastos de Hogares Urbanos y Rurales, realizada por el INEC.

estudios, independientemente de que estos regresen con frecuencia o no a sus hogares de origen.

- **Jefe del hogar**. Es la persona que siendo residente habitual, es reconocida como Jefe por los demás miembros del hogar, ya sea por una mayor responsabilidad en las decisiones familiares, por prestigio, relación familiar o de parentesco, por razones económicas o por tradiciones culturales (ENIGHUR 2012).

- **Perceptor de ingresos.** Es la persona que recibe ingresos de cualquier fuente u origen, sea proveniente del trabajo (asalariado o independiente), la renta de la propiedad (intereses, arriendos, etc.) o de transferencias u otras prestaciones recibidas (ENIGHUR 2012).

2.2 ANÁLISIS DE COMPONENTES PRINCIPALES CATEGÓRICO (ACPC)

El Análisis de Componentes Principales Categórico ACPC o Análisis de Componentes Principales No Lineal permite encontrar los componentes principales de un conjunto de datos donde las variables pueden ser continuas o categóricas. Para lograr esto, recurre al escalamiento óptimo de las variables categóricas. A continuación, se describe este método basado en el texto escrito por Daniel Peña (2002):

Se sabe que, dada una matriz X de orden $(n\ x\ p)$ se puede obtener variables centradas (media igual a cero) a través de la operación:

$$Y = \left(I - \frac{1}{n} 11'\right) X = PX$$

donde,

I: es la matriz identidad.

1: es el vector de unos de dimensión $(n\ x\ 1)$

$1'$: es el vector de unos de dimensión $(1\ x\ n)$

$$P = \left(I - \frac{1}{n}11'\right)$$

Entonces, a partir de la matriz Y se determinan dos matrices definidas positivas: la matriz de covarianzas S, y la matriz de productos cruzados o de similitud T, mediante:

$$S = \frac{Y'Y}{n} \qquad T = YY'$$

Considerando a la matriz T, los elementos de dicha matriz se escriben de la siguiente manera:

$$t_{ij} = \sum_{s=1}^{p} y_{is}y_{js} = y_i'y_j$$

donde, y_i' es la fila i de la matriz Y. Las distancias entre dos individuos se deducen de manera sencilla a partir de la matriz de similitud:

$$d_{ij}^2 = \sum_{s=1}^{p}\left(y_{is} - y_{js}\right)^2 = \sum_{s=1}^{p} y_{is}^2 - 2\sum_{s=1}^{p} y_{is}\, y_{ij} + \sum_{s=1}^{p} y_{js}^2$$

En términos de la matriz T, se puede escribir:

$$d_{ij}^2 = t_{ii} - 2t_{ij} + t_{jj}$$

Por tanto, a partir de la matriz Y se construye la matriz de similitud T y a partir de ésta se puede construir la matriz de distancias al cuadrado $D = \{d_{ij}^2\}$.

2.2.1 RELACIÓN ENTRE COORDENADAS PRINCIPALES Y COMPONENTES PRINCIPALES

Los vectores propios de S corresponden a las componentes principales; mientras que, los vectores propios de T son las coordenadas principales, se puede comprobar es que nS y T tienen el mismo rango y los mismos valores propios no nulos.

Sea w_i un vector propio de $Y'Y$ asociado al valor propio λ_i, entonces:

$$Y'Y\, w_i = \lambda_i w_i$$

multiplicando por Y a cada lado de la ecuación, se tiene:

$$YY'Yw_i = \lambda_i Yw_i$$

Lo que implica que Yw_i es un vector propio de YY' asociado al mismo valor propio λ_i. Además, si $n > p$ y la matriz $Y'Y$ tiene rango completo, entonces, tiene p valores propios no nulos que serán también valores propios no nulos de YY'. Es decir, los vectores propios de YY' son las proyecciones de la matriz Y en la dirección de los vectores propios de $Y'Y$.

Ahora, se puede representar a las componentes principales de Y por los p vectores propios, de la siguiente manera:

$$Z = YW$$

donde,

Z: es una matriz de orden $n\,x\,p$, que tiene en las columnas las componentes principales.

W: es una matriz de orden $p\,x\,p$, que contiene en las columnas a los vectores propios de $Y'Y$.

Por otro lado, la matriz de las coordenadas principales está definida por:

$$Y_\ell = [v_1, \dots, v_p] \begin{bmatrix} \sqrt{\lambda_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda_n} \end{bmatrix} = V\Lambda$$

donde,

v_i: es el vector propio de YY'

V: es una matriz de orden $n\ x\ p$ y contiene los p vectores propios de no nulos de $Y'Y$

Λ: es una matriz diagonal de orden $p\ x\ p$

De aquí que se puede concluir que $Y = V$, se puede ver, que aparte de un factor de escala, ambos procedimientos conducen al mismo resultado.

Las propiedades de las métricas euclídeas se presentan en el Anexo A.

2.3 ANÁLISIS DE CONGLOMERADOS

El objetivo del análisis de conglomerados (*clusters*) es agrupar elementos (individuos o variables) en grupos homogéneos en función de sus similaridades; estos se basan en métodos de clasificación automática utilizando criterios de similitudes.

Existen métodos de partición y jerárquicos, en este trabajo se utiliza el algoritmo de K-Medias, el cual tiene las siguientes propiedades:

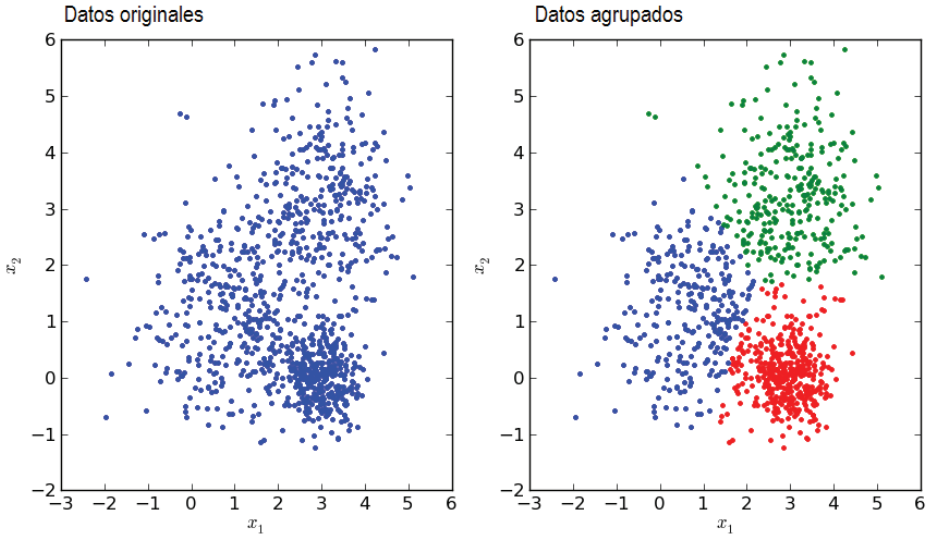
- Funciona de manera eficiente cuando el número de casos es grande.
- Forma grupos a partir de la similitud entre casos usando centroides.
- Los grupos que encuentra son distintos entre sí.
- No asume ninguna distribución específica para las variables.

2.3.1 ALGORITMO DE PARTICIÓN: K-MEDIAS

Sea la matriz X de orden $n\ x\ p$. Se busca dividirla en K grupos. El *algoritmo k-medias* se esquematiza en las siguientes etapas:

1. Seleccionar K puntos como centros de los grupos iniciales. Esta selección puede ser: i) Asignando aleatoriamente los objetos a K grupos y tomando los centros dentro de cada uno; ii) tomando como centros los K puntos más alejados entre sí; o, iii) construyendo los grupos iniciales, con información *a priori,* y calculando sus centros.
2. Calcular las distancias euclídeas de cada elemento a los centros de los K grupos y asignar cada elemento al grupo cuyo centro sea el más próximo. La asignación es secuencial y se debe recalcular las coordenadas del centro del grupo con cada uno de los elementos que se va incorporando.
3. Definir un criterio de aglomeración óptima y comprobar si reasignando alguno de los elementos mejora el criterio.
4. Si no mejora el criterio de aglomeración óptima, se termina el algoritmo.

Figura 1. Ejemplo del agrupamiento K-Medias



Elaborado por: El Autor

En el Anexo B se muestra a detalle el esquema del algoritmo de partición K-Medias.

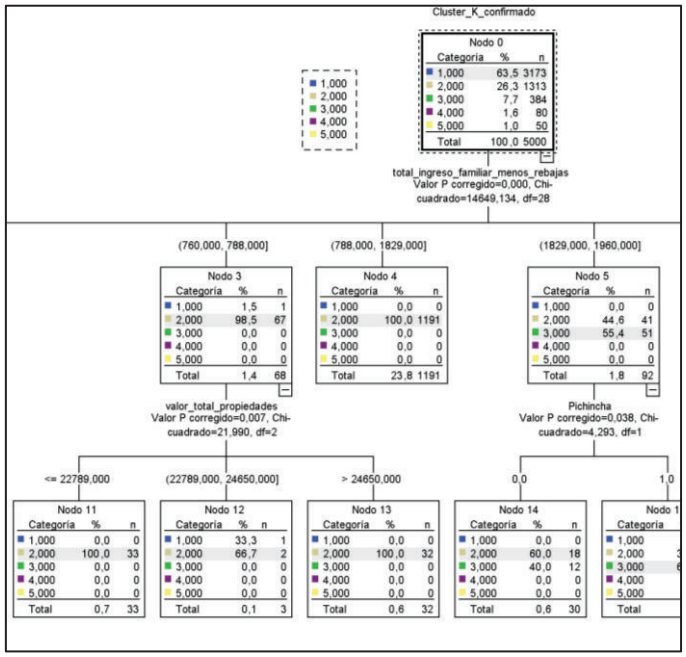
2.4 ÁRBOL DE CLASIFICACIÓN

Los árboles de clasificación son técnicas de exploración de datos que consisten en estudiar grandes cantidades de datos con el fin de encontrar patrones no triviales. Representan una serie de pautas basadas en ciertas variables explicativas que se muestran según se recogen en el árbol; esta sección corresponde a una adaptación del IBM Knowledge Center en el desarrollo de su herramienta Answer Tree 2015 y también se ha adaptado del documento Arboles de Clasificación y Regresión por José Manuel Rojo (2006).

Como se menciona en el documento de Rojo (2006), los árboles se construyen a través de un algoritmo que va dividiendo de forma recursiva los registros de la base de datos en grupos, representados por nodos, de manera que con cada subdivisión las frecuencias relativas de las categorías de la variable dependiente vayan tendiendo a 0 o a 1.

Un árbol es un grafo conexo que tiene un nodo inicial que se denomina *Nodo Raíz* y a partir de él se forman las aristas (ramas) con nodos que pueden formar otras aristas o ser nodos terminales.

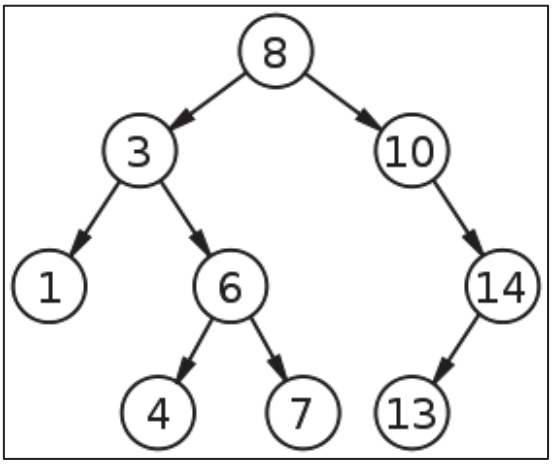
Figura 2. Ejemplo de un árbol de clasificación



Elaborado por: El autor

Árbol binario. Es un árbol en el que ningún nodo puede tener más de dos subárboles. En un árbol binario cada nodo puede tener cero, uno o dos subárboles (hijos). Se conoce el nodo de la izquierda como hijo izquierdo y el nodo de la derecha como hijo derecho.

Figura 3. Ejemplo de un árbol binario



Elaborado por: El autor

2.4.1 ÍNDICES DE ASOCIACIÓN Y MEDIDAS DE IMPUREZA

De acuerdo al documento de Rojo (2006), construimos árboles de clasificación al realizar subdivisiones sucesivas de tal forma que en cada nueva subdivisión mejora la clasificación con respecto a la variable criterio (variable del nodo raíz). Así, en cada subdivisión se evalúa la clasificación con los siguientes índices:

- Coeficiente Chi – cuadrado (Ji-cuadrada)
- Índice de Gini
- Índice binario
- Entropía

En este trabajo se utiliza el coeficiente Chi cuadrado, los otros índices: Gini, binario y entropía se describen en el Anexo C.

2.4.1.1 Coeficiente Chi – Cuadrado

Esta sección se ha adaptado de Rojo (2006), el coeficiente Chi – cuadrado (χ^2) trata de medir la asociación entre dos variables nominales, ordinales o una combinación de ambas partiendo de una tabla de contingencia, a partir de dos variables y sus categorías.

Se define las variables A y B con sus modalidades $(A_1, \dots, A_k)$ y $(B_1, \dots, B_p)$, respectivamente; de esta manera se obtiene una tabla $k\ x\ p$, el número de toda la población n es la suma de las frecuencias n_{ij} del interior de la tabla.

Por otro lado, se hace un recuerdo de algunos principios de probabilidades que son importantes para esta sección.

- Dos sucesos A y B son independientes, si: $P(A \cap B) = P(A) * P(B)$.
- Si un suceso A tiene una probabilidad de ocurrencia P y se realizan n repeticiones del experimiento aleatorio, entonces el número esperado de ocurrencias de este suceso será: $n * p$.

Si las categorías de la variable fila (I) y las categorías de la variable columna (J) son independientes, se debería cumplir con la condición:

$P(I \cap J) = P(I) * P(J)$.

Por lo tanto, la frecuencia esperada en cada celda de la tabla de doble entrada si las variables fueran independientes sería:

$$n'_{ij} = n * P(i \cap j\ ) = n * P(i) * P(j) = n * \frac{n_{i*}}{n} * \frac{n_{*j}}{n}$$

Por tanto, comparando las frecuencias observadas con las frecuencias esperadas si fueran independientes se tendrá una idea del grado de asociación existente entre las variables, el coeficiente chi-cuadrado ayuda a esto y se define de la siguiente manera:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{p} \frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}}$$

donde,

n_{ij}: es la frecuencia observada de la celda $\{i, j\}$

n'_{ij}: es la frecuenica esperada de la celda $\{i, j\}$

2.4.2 ALGORITMOS PARA DESARROLLAR ÁRBOLES

Existen varios algoritmos para la construcción y desarrollo de los árboles, como son CHAID, CART y QUEST, pero se consideró un algoritmo que permite clasificar cuando la variable dependiente es de tipo multinomial, en este documento se menciona el que se utiliza en este proyecto.

2.4.2.1 Chaid

CHAID son las siglas en inglés de Detector Automático de Iteraciones Chi Cuadrado (Chi-Squared Automatic Interaction Detector), este algoritmo permite trabajar con variables de cualquier medida (cuantitativas o cualitativas), además que también clasifica variables multinomiales. Dependiendo del nivel de minuciosidad con que se vaya a unir las categorías, este algoritmo puede ser CHAID o CHAID Exhaustivo, como se menciona en Rojo (2006), en cualquiera de los dos casos el algoritmo realiza los siguientes pasos:

1. Tomar la variable a predecir o clase criterio.
2. Dada una variable predictora, se unen las categorías más homogéneas y se dejan las heterogéneas inalteradas.
3. De todas las variables predictoras posibles, se elige a aquella que tenga el mayor valor del coeficiente chi – cuadrado para formar la primera rama del árbol. Si la variable es continua, se utiliza la prueba de Fisher (F).
4. Se regresa al paso 3, hasta que todas las variables queden agrupadas.

2.5 REGRESIÓN LOGÍSTICA MULTINOMIAL

Esta sección es una adaptación de Econometric Analysis de Greene, 2007.

Sea Y una variable aleatoria categórica con más de dos categorías (nominal u ordinal) y sean $X_1, \dots X_n$ variables independientes que pueden ser cualitativas o cuantitativas. El modelo Logit Multinomial es una generalización del modelo Logit Binomial.

2.5.1 MODELOS LOGIT PARA RESPUESTAS NOMINALES

Se denota como K al número de categorías de la variable Y y $\{\pi_1, \dots, \pi_K\}$ a las probabilidad de distintas respuestas que satisfacen:

$$\sum_{k=1}^{K} \pi_k = 1$$

Supóngase que se tienen n observaciones independientes que se distribuyen en las K categorías, tal que la distribución de probabilidad del número de observaciones de las K categorías sea una distribución multinomial.

Como la variable dependiente es de respuesta nominal, el orden de las K categorías es irrelevante. Por tanto, se toma una de las categorías como respuesta ***base*** y se define un modelo logit con respecto a ella:

$$\log\left(\frac{\pi_k}{\pi_K}\right) = \alpha_k + \beta_k x \quad con\ \ k = 1, \dots, K-1$$

El modelo tiene $K-1$ ecuaciones con sus propios parámetros y los efectos varían con respecto a la categoría que se tome como base. Cuando $K = 2$ el modelo es equivalente a un logit binomial.

La ecuación general logit con respecto a la categoría base K determina al mismo tiempo los modelos logit correspondientes a cualquier pareja de categorías; así, si se considera dos categorías cualquiera (i, j) de las K posibles, se tiene:

$$\log\left(\frac{\pi_i}{\pi_j}\right) = \log\left(\frac{\pi_i}{\pi_K}\right) - \log\left(\frac{\pi_j}{\pi_K}\right) = (\alpha_i + \beta_i x) - (\alpha_j + \beta_j x) = (\alpha_i - \alpha_j) + (\beta_i - \beta_j)x$$

entonces,

$$\log\left(\frac{\pi_i}{\pi_j}\right) = \alpha_{ij} + \beta_{ij} x, \quad con\ \alpha_{ij} = (\alpha_i - \alpha_j)\ y\ \beta_{ij} = (\beta_i - \beta_j)$$

Para el caso de datos ordinales tenemos que la probabilidad de una variable Y es la probabilidad de que Y sea menor o igual que un determinado valor k. Así, para una categoría dada k se define la *probabilidad acumulada* como

$$P(Y \leq k) = \pi_1 + \cdots + \pi_k, \ \ con\ k = 1,2, \ldots, K$$

El detalle de los modelos acumulados para datos ordinales se presenta en el Anexo D.

2.6 MAQUINAS DE VECTORES DE SOPORTE (SUPPORT VECTOR MACHINES SVM)

Esta sección ha sido tomada y adaptada del trabajo hecho por Betancourt (2005), publicado en la Scientia et Technica Año XI, No 27, Abril 2005. UTP. ISSN 0122-1701.

La SVM es una técnica de clasificación diseñada originalmente para el análisis de clasificación binaria (en dos grupos). Es una máquina de aprendizaje que separa los datos con hiperplanos y si no es posible separar los datos, en el espacio que están inicialmente los datos, con un hiperplano lineal, se debe realizar una traslación de los datos mediante una aplicación no lineal con vectores de entrada a un nuevo espacio de dimensión más alta. En este nuevo espacio se construye una superficie de decisión lineal que tenga propiedades que garanticen que la capacidad de generalización de la máquina de aprendizaje sea alta. El desarrollo teórico de esta herramienta fue realizada por Vapnik (1995) para el caso de hiperplanos óptimos para clases separables. Se definió un hiperplano óptimo como una función de decisión lineal con el

margen de separación máximo entre los vectores de las dos clases. Entonces, para construir el hiperplano se debe tener en cuenta una cantidad pequeña de los datos de entrenamiento (vectores soporte), quienes determinaban que el margen sea máximo.

La maximización del margen (m) es un problema de programación cuadrática y puede ser resuelto a través de un problema dual con multiplicadores de Lagrange; sin el conocimiento de cómo están distribuidos los datos, el SVM encuentra el hiperplano óptimo utilizando el producto escalar con funciones de características de los *kernel*.

2.6.1 EL CASO LINEALMENTE SEPARABLE

Suponga que se tiene un conjunto de datos X, tal que $x_i \in \Re^n$ pertenece a un conjunto de etiquetas Y, tal que $y_i = \{-1,1\}$ para $i = 1, \dots, \ell$. En la mayoría de los casos, buscar el hiperplano adecuado en el espacio de entrada es demasiado restrictivo para que se pueda calcular en la práctica.

Una solución para este problema es mapear el espacio de entrada en un espacio de características de mayor dimensión; para esto se va a definir lo siguiente:

Sea $z = \phi(X)$ la notación del vector en el espacio de características con un mapeo $\phi(X)$ de $\Re^n$, a un espacio de características Z. Se debe encontrar el hiperplano:

$$p.z + b = 0$$

Tal que,

$$f(x_i) = sign(p.z_i + b) = \begin{cases} 1 & y_i = 1 \\ -1 & y_i = -1 \end{cases}$$

donde,
$p \in Z$ y $b \in \Re$

En conclusión, se dice que el conjunto X es separable linealmente si existe el par (p, b) tal que las inecuaciones:

$$\begin{cases} p.z + b \geq 1, & y_i = 1 \\ p.z + b \leq -1, & y_i = -1 \end{cases} \qquad i = 1, \dots, \ell$$

Sean válidas para todos los elementos del conjunto. En este caso, se puede encontrar la maximización del margen de las proyecciones de la base de entrenamiento.

2.6.2 EL CASO LINEALMENTE NO SEPARABLE

Si el conjunto X no es linealmente separable es necesario agregar variables no negativas $(\varepsilon_i \geq 0)$ tal que se obtenga la función modificada:

$$y_i(p.z + b) \geq 1 - \varepsilon_i, \quad i = 1, \dots, \ell$$
$$y_i(p.z + b) \leq -1 - \varepsilon_i, \quad i = 1, \dots, \ell$$

Los $\varepsilon_i \neq 0$ son aquellos en los que el punto x_i no satisface la inecuación del caso linealmente separable. Además, el término:

$$\sum_{i=1}^{\ell} \varepsilon_i$$

Se puede considerar como una medida del error de clasificación.

El problema del hiperplano óptimo se encuentra solucionando el problema:

$$\min \left\{ \frac{1}{2} p.p + C \sum_{i=1}^{\ell} \varepsilon_i \right\}$$

$$sujeto\ a\ las\ restricciones:$$

$$y_i(p.z + b) \geq 1 - \varepsilon_i, \quad i = 1, \dots, \ell$$

$$\varepsilon_i \geq 0, \qquad i = 1, \dots, \ell$$

donde,

C es una constante

A la constante C se le conoce como un parámetro de regularización y es el único que puede ajustarse en la formulación del SVM.

Utilizando un Lagrangiano para la maximización del margen y la transformación dual del problema, se puede obtener lo siguiente:

$$Max\ P(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j z_i \,.\, z_j$$

$$sujeto\ a\ las\ restricciones:$$

$$\sum_{i=i}^{\ell} y_i \alpha_i = 0,\ \ 0 \le \alpha_i \le C,\quad i = 1, \dots \ell$$

donde,

$\alpha = (\alpha_1, \dots, \alpha_\ell)$: es un vector de multiplicadores de Lagrange asociados a las contantes (b) de las inecuaciones del problema normal.

Este problema se resuelve aplicando el Teorema de Khun Tucker en el problema que se presenta en el anexo E.

2.6.3 USO DE LA FUNCION NUCLEO (KERNEL) PARA EL CASO LINEALMENTE NO SEPARABLE

Dado que no se tiene un conocimiento previo de la función ϕ, la resolución del problema se vuelve imposible. Para saltar este inconveniente, se utiliza una función $K(.,.)$ denominada *nucleo* o *kernel* con la que se obtenga el producto escalar de los puntos de entrada en el espacio de características Z; es decir:

$$z_i . z_j = \phi(x_i).\phi(x_j) = K(x_i, x_j)$$

Teorema de Mercer. Sea X un subconjunto compacto de $\Re^n$. Supongase que K es una función continua simétrica tal que el operador integral :

$$T_K: \quad L_2(X) \rightarrow \qquad L_2(X)$$

$$(T_K f)(.) = \int_X K(.,x) f(x) dx$$

Donde $L_2(X)$ es un espacio de Lebesgue y el operador integral es definido positivo, lo que implica que:

$$\int_{X\,x\,X} K(x_1, x_2) f(x_1) f(x_2) dx_1 dx_2 \geq 0, \quad \forall f \in L_2(X)$$

Se puede expandir la función $K(x_1, x_2)$ en una serie con convergencia uniforme sobre $X\ x\ X$ en términos de T_K funciones propias $\psi_j \in L_2(X)$, normalizado de tal forma que $\|\psi_j\|_{L_2} = 1$ y valores propios asociados positivos $\lambda_j \geq 0$, entonces:

$$K(x_1, x_2) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x_1) \psi_j(x_2)$$

La imagen de un vector x cualquiera del espacio de partida, a través del mapeo implícito definido por el kernel es:

$$\sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \psi_j(x)$$

Un kernel que satisface el teorema de Mercer es el gaussiano:

- **Kernel estándar gaussiano (radio basal)**

$$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$$

Otros núcleos (kernel) se mencionan en el anexo E.

CAPÍTULO 3
METODOLOGÍA DE CONSTRUCCIÓN DEL ÍNDICE SOCIOECONÓMICO DE LOS ESTUDIANTES DE LA ESCUELA POLITÉCNICA NACIONAL: ANÁLISIS Y RESULTADOS

3.1 ANÁLISIS DESCRIPTIVO DE VARIABLES

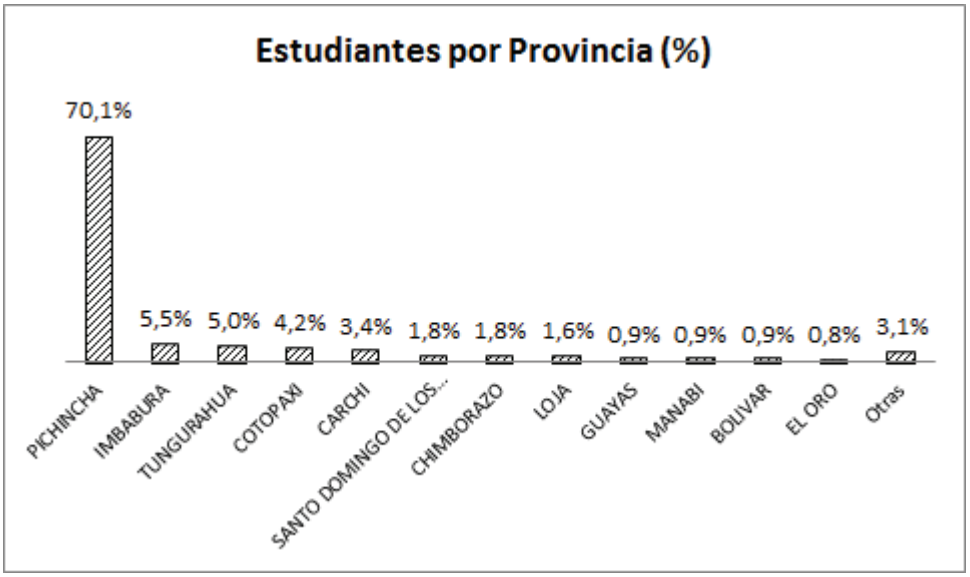
Dentro de este trabajo se han considerado 6.566 estudiantes de la Escuela Politécnica Nacional (EPN) matriculados en el período 2016 A; de los cuales se analizan las variables disponibles de los estudiantes proporcionada por la Dirección de Gestión, Información y Procesos (DGIP) de la EPN. Para el procesamiento y análisis de los datos se utiliza el *software* SPSS 20.

Para un mejor análisis de las variables se las va a dividir por grupos:

3.1.1 Variables de caracterización del estudiante

Dentro de la base de datos se tiene que la gran mayoría de los estudiantes (70,1%) son de la provincia de Pichincha, existe un 5,5% y un 5% que pertenecen a las provincias de Imbabura y Tungurahua, respectivamente. El 20% restante, se distribuye como se muestran en la figura siguiente:

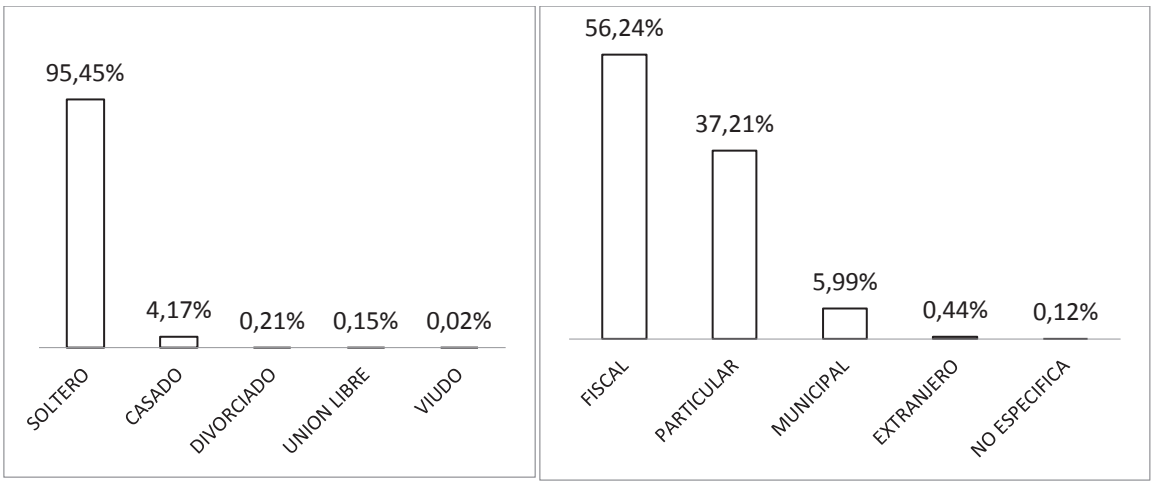
Figura 4. Distribución de los estudiantes por provincia



Fuente: DGIP
Elaborado por: El Autor

También, se tiene que el 95,45% de los estudiantes son solteros y el 4,17% casados. Además, el 56,2% de los estudiantes provienen de un colegio fiscal.

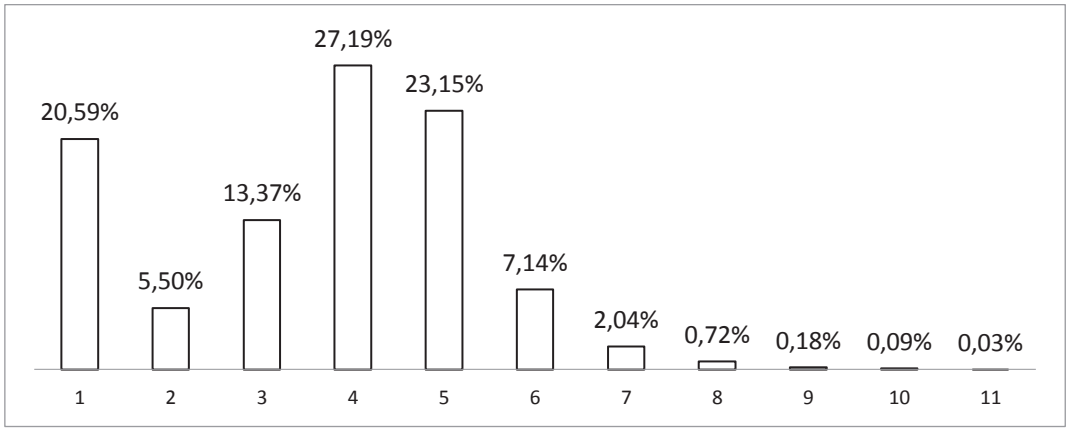
Figura 5. Estado civil del estudiante y colegio de procedencia



Fuente: DGIP
Elaborado por: El Autor

En cuanto a la composición familiar de los estudiantes, se tiene que el 27,2% tienen una familia que la conforman 4 miembros y un 23,2% que están en una familia de 5 miembros. Un 20,6% de estudiantes viven solos y un 18,9% tienen entre 2 y 3 miembros en su familia y el 10,2% están en una familia de 6 o más miembros.

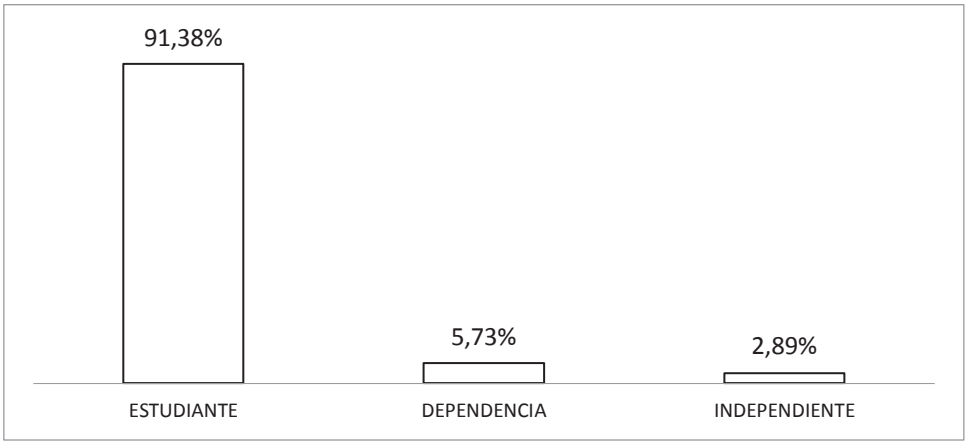
Figura 6. Número de miembros del hogar del estudiante



Fuente: DGIP
Elaborado por: El Autor

Por otro lado, existe un 8,6% de estudiantes que trabajan (5,7% bajo dependencia y un 2,9% de manera independiente) y el restante 91,4% son estudiantes exclusivamente.

Figura 7. Relación laboral del estudiante



Fuente: DGIP
Elaborado por: El Autor

3.1.2 VARIABLES DE INGRESO FAMILIAR

Tabla 1. Variables de ingreso familiar y del estudiante

	Mín	Media	Mediana	Máx	Desv. Est.
Ingreso declarado por el estudiante	5	745	600	12.460	622
Ingreso de la Madre	0	278	0	12.028	454
Ingreso del Padre	0	554	400	16.320	712
Ingreso por Arriendos	0	25	0	2.200	115
Otros Ingresos	0	45	0	11.008	215
Ingreso Per Cápita Familiar	0	331	140	5.994	373

Fuente: DGIP

El ingreso familiar que tenga el estudiante es una de las variables importantes al momento de realizar una clasificación socioeconómica. De los datos recolectados de los estudiantes de la EPN, se puede decir que los estudiantes en su ingreso a la universidad reportan tener un ingreso familiar promedio de USD 745, alcanzando máximos de hasta USD 12.460. Adicionalmente, existen hogares en los que el ingreso de la madre es de USD 0 y pueden llegar hasta USD 12.028. Algo similar sucede con los ingresos del padre.

Un indicador importante para el análisis es el Ingreso Per cápita Familiar, en el que hay familias con un ingreso per cápita de hasta USD 5.994 y otras con un mínimo de cero. El ingreso per cápita promedio es de USD 331 que resulta ser prácticamente un salario básico por cada persona. Sin embargo, considerando la mediana se puede observar que el 50% de las familias de los estudiantes tienen un ingreso menor a USD 140 por persona.

3.1.3 VARIABLES DE PATRIMONNIO FAMILIAR DEL ESTUDIANTE

Tabla 2. Estadísticos descriptivos de las variables de patrimonio familiar del estudiante

	Mín	Media	Mediana	Máx	Desv. Est.
Núm. total de propiedades	0	1	1	5	1
Valor de Propiedad principal	0,00	25.377	8.000	345.746	38.640
Valor de Terreno	0,00	4.064	0,00	116.680	10.491
Valor del Auto	0,00	2.528	0,00	40.000	5.676
Número total de vehículos	0,00	0,00	0,00	2	1
Valor total de vehículos	0,00	3.670	0,00	55.000	7.019
Valor total de propiedades y vehículos	0,00	32.142	17.475	359.769	43.146

Fuente: DGIP

En las familias de los estudiantes se pueden observar que hay algunas que no tienen ninguna propiedad en su patrimonio y algunas llegan a un máximo de cinco; aunque, el promedio es de una propiedad por familia. Existen familias que tienen hasta dos autos y a través de la mediana se puede concluir que por lo menos el 50% de las familias no tienen autos. Los valores referenciales de las propiedades se muestran en la tabla anterior.

3.2 CONSTRUCCIÓN DEL ÍNDICE SOCIECONÓMICO

Para la construcción del índice socioeconómico se va a utilizar el método multivariante denominado Análisis de Componentes Principales Categórico (ACPC) que permite reducir la dimensión de las variables que tienen una mayor correlación y genera factores lineales que explican la mayor cantidad de varianza en los datos.

Se inició considerando todas las variables presentadas en la sección precedente para poder determinar la influencia en cada una de ellas. Además, se integraron variables que se crearon para mejorar al modelo; estas son:

- Madre Trabaja: Es una variable dicotómica que tiene como valor 1 si el ingreso de la madre tiene un valor positivo y 0 en caso contrario.
- Colegio fiscal: Es una variable dicotómica que tiene como valor 1 si el estudiante proviene de un colegio fiscal y 0 en caso contrario.

A partir de lo expuesto anteriormente, y considerando dos factores, se obtuvieron los siguientes resultados (se denominará Modelo Inicial):

Tabla 3. Resultados del Modelo Inicial

Componente	Alfa de Cronbach	Total (Autovalores)	% de la varianza
1	0,727	3,237	16,184
2	0,541	2,057	10,283
Total	0,831	4,756	23,782

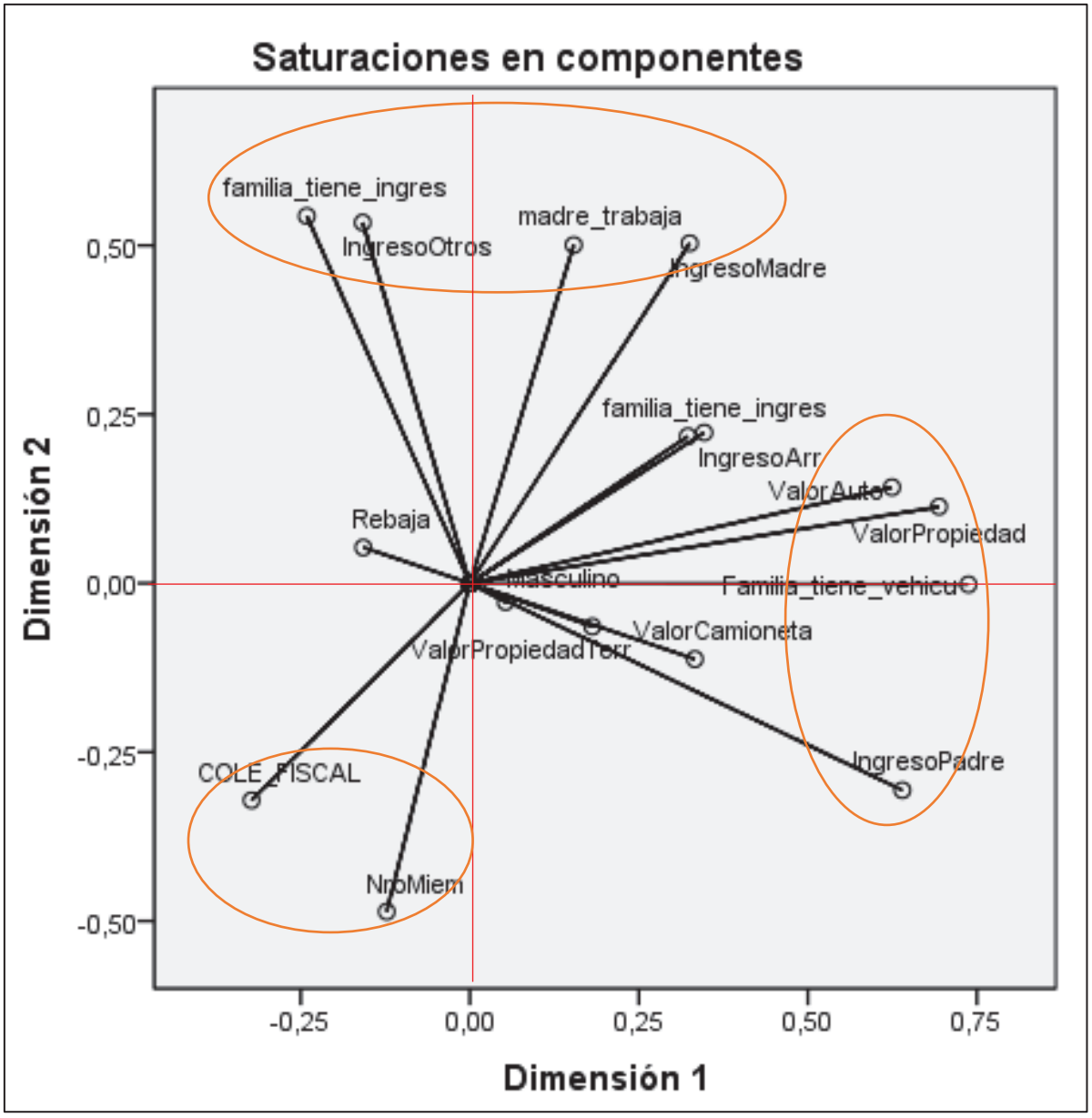
Elaborado por: El Autor

Como se muestra en la tabla anterior, el porcentaje total de varianza explicada por los factores es bajo (23,782%), lo que indica que con este modelo se explica menos de la mitad de la varianza de los datos. También se muestra el estadístico Alfa de Cronbach (0,831) que es una medida de confiabilidad que se maximiza cuando el modelo es "bueno", este estadístico está entre 0 y 1.

Se puede concluir que este modelo no es bueno y que se necesita reformularlo; sin embargo, se presentan los resultados para determinar las variables que se deben quitar por estar mal representadas para mejorar el modelo y por ende la metodología para determinar el nivel socioeconómico de los estudiantes de la EPN.

Un resultado importante dentro del análisis es la saturación en los factores (componentes) principales; en este caso, se muestra a continuación:

Figura 8. Saturación en componentes para el Modelo Inicial



Elaborado por: El Autor

Como se puede observar se tienen varias variables que están muy cerca del origen y que no se encuentran "bien" representadas en ninguna de las componentes; por tanto, es necesario eliminarlas para poder mejorar el modelo. Además, se puede inferir que es necesario agregar una componente adicional porque las variables forman tres grupos. Numéricamente las saturaciones se muestran en la siguiente tabla:

Tabla 4. Saturación en componentes del Modelo Inicial

	Dimensión	
	1	2
COLE_FISCAL	-0,323	-0,321
madre_trabaja	0,154	0,501
RProvinciaa		
RNivelacion		
RTipoPropiedad		
RRelacionLaboralPadre		
NroMiem	-0,123	-0,486
IngresoPadre	0,640	-0,306
IngresoMadre	0,325	0,503
IngresoArr	0,347	0,223
IngresoOtros	-0,158	0,534
Rebaja	-0,158	0,053
ValorPropiedad	0,694	0,113
ValorPropiedadTerr	0,181	-0,064
ValorAuto	0,624	0,142
ValorCamioneta	0,333	-0,112
familia_tiene_ingresos_x_arriendos	0,322	0,217
familia_tiene_ingreso_extra	-0,241	0,544
Masculino	0,053	-0,028
Familia_tiene_vehiculo	0,737	-0,002

Elaborado por: El Autor

Se sombrean en color amarillo las variables que no tienen puntuaciones en las dimensiones y aquellas cuyas puntuaciones están entre $\pm 0{,}20$ por no estar bien representadas.

Por otro lado, para mejorar el modelo se agrupan algunas variables ya que están contenidas directa o indirectamente en otras; estas son:

a. Total ingresos menos rebajas. Esta variable contienen a: Ingreso, Ingreso Padre, Ingreso Madre, Ingreso Arriendo, Ingreso Otros, Rebaja.
b. Valor total propiedades. En este caso, esta variable contiene a: Valor Propiedad, Valor Propiedad Terreno, Valor Auto y Valor Camioneta.
c. Total propiedades. Esta variable contiene a: Familia tiene Propiedad, Familia tiene terreno, Número total vehículos, es decir acumula el número de propiedades y vehículos.

Por lo tanto, se va a realizar el análisis con 6 variables que engloban a todas las variables que se obtuvieron de la DGIP. Los resultados del modelo estimado con estas variables y con tres componentes se detallan a continuación (se denominará Modelo Final):

Tabla 5. Resumen del Modelo Final

Dimensión	Alfa de Cronbach	Varianza explicada	
		Autovalores	% varianza
1	0,656	2,207	36,788
2	0,292	1,322	22,026
3	-0,016	0,987	16,449
Total	0,934	4,516	75,263

Elaborado por: El Autor

En el modelo encontrado, se tiene un total de tres componentes un 75,26% de varianza explicada y su Alfa de Cronbach es del 0,93. Es importante recalcar que se han tomado tres dimensiones aunque el tercer valor propio (autovalor) tiene un valor inferior a uno (prácticamente es uno); sin embargo, al considerar también la tercera componente, la varianza explicada pasa de un 58,8% a un 75,263%; por tanto, el modelo es bueno y representativo de las variables originales.

Es importante analizar la representación de las variables en cada uno de los ejes seleccionados; la saturación en los ejes nos indica este resultado, se presentan a continuación:

Tabla 6. Puntuaciones de saturación de las componentes

	Componentes		
	C1	C2	C3
Total_Ingreso_Familiar_Menos_Rebajas	0,732	-0,002	0,220
Total_Propiedades_y_vehiculos	0,790	0,414	-0,111
Valor_Total_Propiedades_y_vehiculos	0,806	0,351	-0,129
Número_Miembros	-0,317	0,745	0,223
Cole_fiscal	-0.473	0.624	0.287
Madre_Trabaja	0,269	-0,287	0,882

Elaborado por: El Autor

Se puede dar "nombres" a las componentes a partir de las variables mejor representadas en cada una de las componentes mostradas en la tabla anterior:

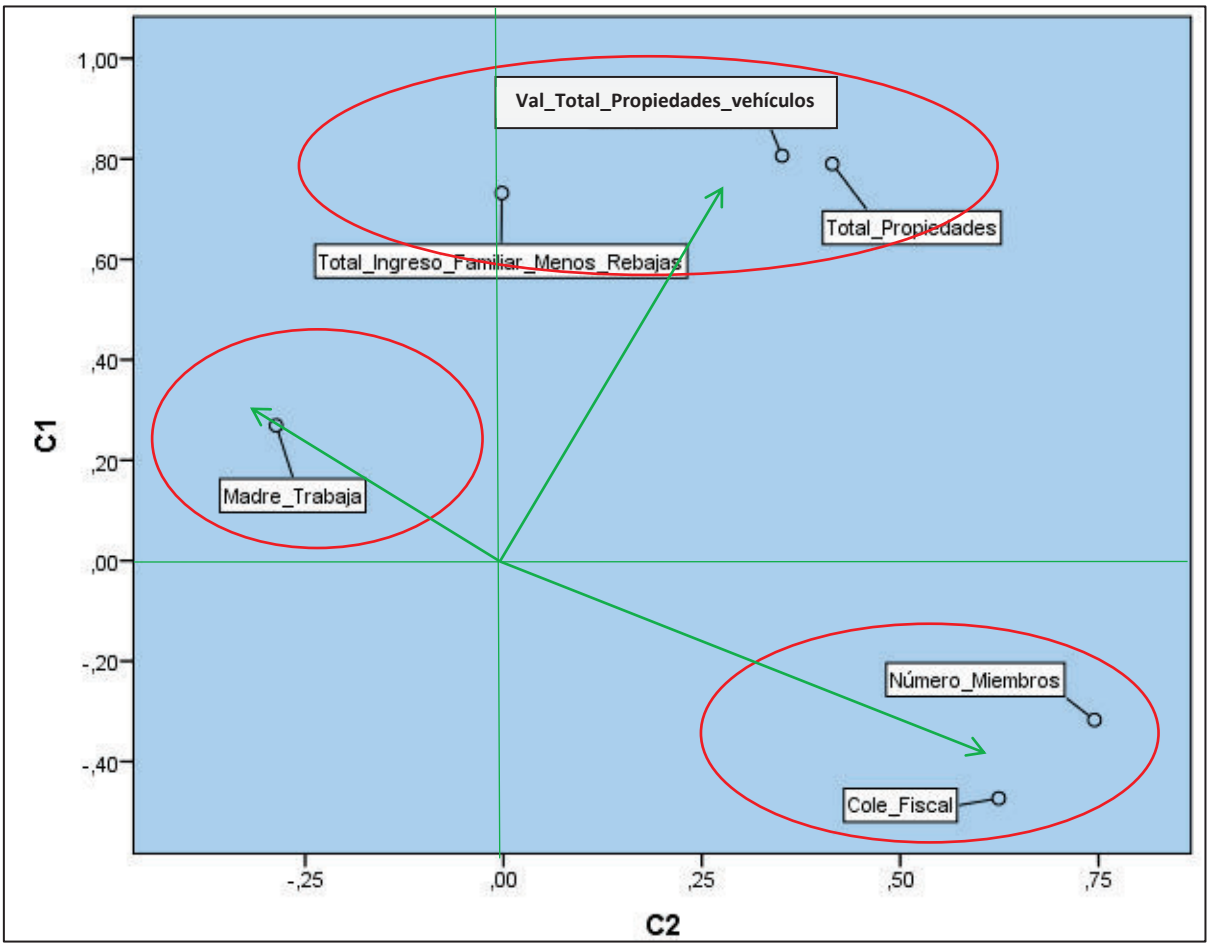
1. La primera componente explica el 36,788% y está compuesto principalmente por las variables: Total_ingreso_familiar_menos_rebajas, Total_Propiedades_y_vehiculos, Valor_Total_Propiedades_y_vehiculos. Por tanto, este eje se denominará ***componente de patrimonio familiar (C1).***

2. La segunda componente explica el 22,026% y se encuentra compuesta por las variables Número de Miembros y Cole_Fiscal . Por tanto, este eje se denominará ***componente de características del hogar (C2).***

3. La tercera componente explica el 16,445% de la varianza total de los datos y está representada por la variable: Madre Trabaja. Por tanto, este eje se denominará ***componente de estado laboral de la madre (C3).***

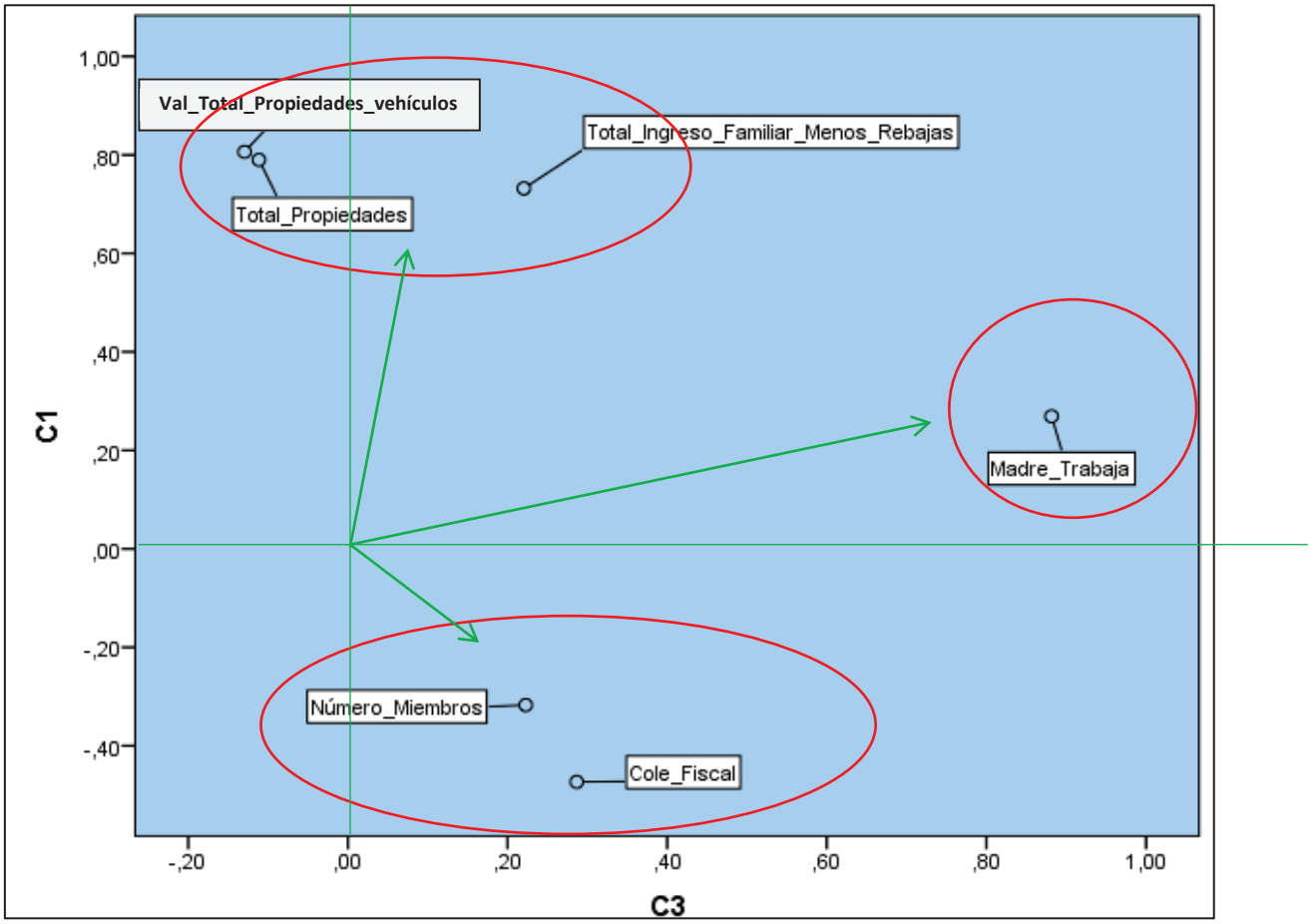
Gráficamente, los gráficos de dispersión entre las componentes se muestran a continuación:

Figura 9. Gráfico de saturaciones Componente 1 vs Componente 2



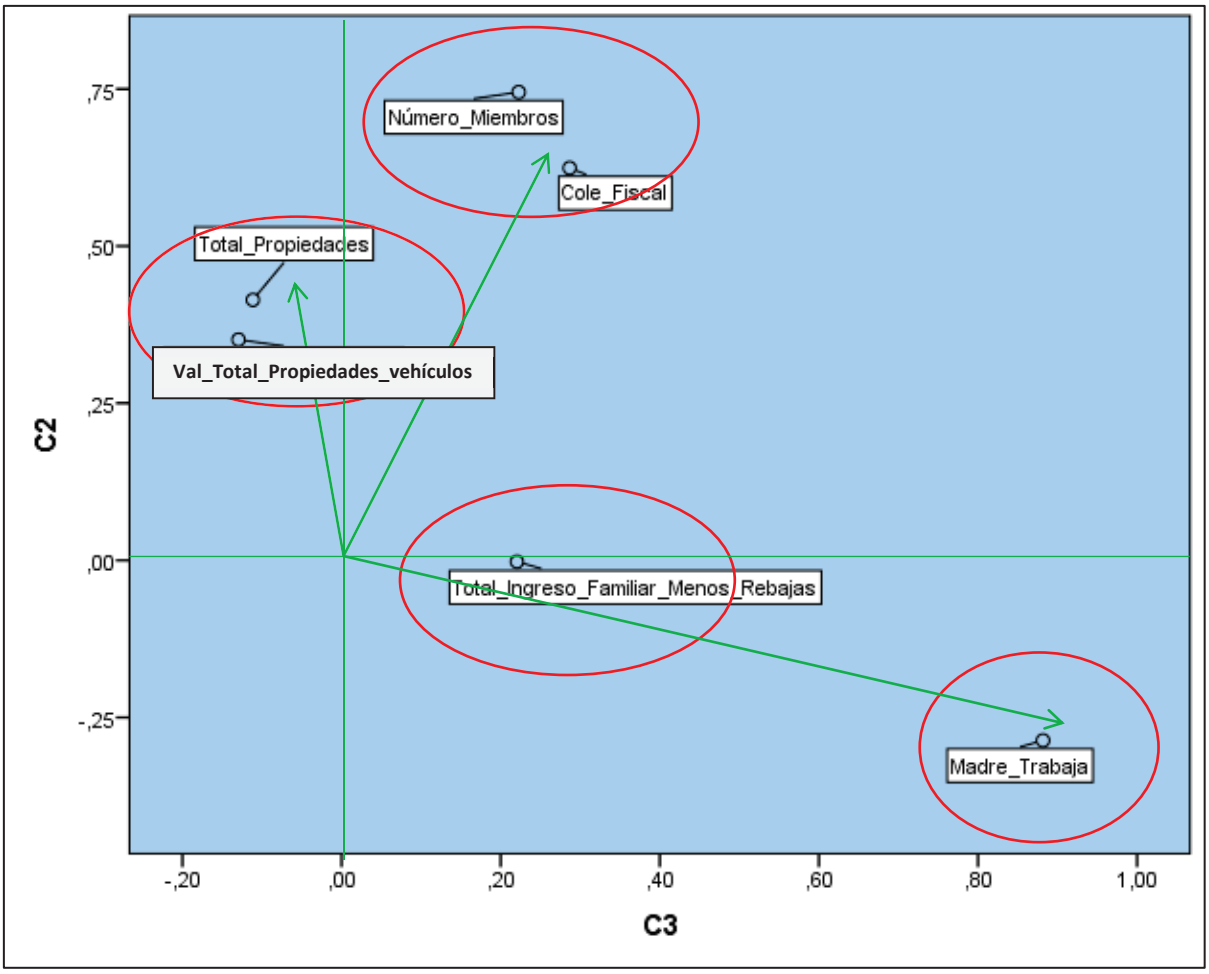
Elaborado por: El Autor

Figura 10. Gráfico de saturaciones Componente 1 vs Componente 3



Elaborado por: El Autor

Figura 11. Gráfico de saturaciones Componente 2 vs Componente 3



Elaborado por: El Autor

3.3.1 CÁLCULO DEL ÍNDICE DE NIVEL SOCIOECONÓMICO (INS)

Para el cálculo del índice es necesario tener las puntuaciones de los individuos en cada una de las componentes, para este fin se va a realizar la determinación de las ecuaciones de las componentes; esto es una combinación lineal de las variables y las puntuaciones consideradas por las saturaciones de las componentes. Las ecuaciones encontradas son:

$$C1 = -0{,}473 * Colegio\_Fiscal + 0{,}269 * Madre\_Trabaja - 0{,}317 * Número\_miembros + 0{,}732 * Total\_Ingreso\_Familiar\_Menos\_Rebajas + 0{,}806 * Valor\_Total\_Propiedades\_y\_vehículos + 0{,}790 * Total\_Propiedades\_y\_vehículos$$

$$C2 = 0{,}624 * Colegio\_Fiscal - 0{,}287 * Madre\_Trabaja - 0{,}745 * Número\_Miembros - 0{,}002 * Total\_Ingreso\_Familiar\_Menos\_Rebajas + 0{,}351 * Valor\_Total\_Propiedades\_y\_vehículos + 0{,}414 * Total\_Propiedades\_y\_vehículos$$

$$C3 = 0{,}287 * Colegio\_Fiscal + 0{,}882 * Madre\_Trabaja + 0{,}223 * Número\_Miembros + 0{,}002 * Total\_Ingreso\_Familiar\_Menos\_Rebajas - 0{,}129 * Valor\_Total\_Propiedades\_y\_vehículos - 0{,}111 * Total\_Propiedades\_y\_vehículos$$

Una vez calculadas las puntuaciones de los individuos en cada una de las componentes se procede a realizar el cálculo del Índice de Nivel Socioeconómico, para esto se va a utilizar la siguiente fórmula:

$$I_1 = \lambda_1 C_1 + \lambda_2 C_2 + \lambda_3 C_3$$

donde,

I_1: Es el índice de nivel socioeconómico sin normalizar

Para que el índice tenga una mejor interpretación, se realiza la normalización del mismo y poder obtener una escala entre 0 y 1, para esto se calcula de la siguiente forma:

$$I_{INS} = \frac{I_{1,MAX} - I_1}{I_{1,MAX} - I_{1,MIN}}$$

donde,

I_{INS}: Es el índice de nivel socioeconómico normalizado

I_1: Es el índice de nivel socioeconómico sin normalizar

$I_{1,MAX}$: Es el valor máximo del índice de nivel socioeconómico sin normalizar

$I_{1,MIN}$: Es el valor mínimo del índice de nivel socioeconómico sin normalizar

3.3.2 ESTRATOS DEL INS

Una vez que se ha calculado el INS, es necesario definir los estratos en los que se van a agrupar a los individuos analizados dentro de este estudio; para esto se realiza una clasificación de los individuos utilizando la técnica multivariante de agrupación denominada algoritmo K-Medias.

3.3.2.1 Cluster K – Medias

Con el INS definido en la subsección anterior, se procede a realizar un proceso de agrupamiento; para este fin, se aplica el algoritmo de agrupación K-Medias 5 grupos[2] con las mismas variables consideradas para la construcción del INS, exceptuando las variables categóricas (este procedimiento no procesa variables categóricas); los resultados se presentan a continuación:

Tabla 7. Conglomerados a partir del algoritmo K-Medias con 5 grupos

Grupo	Casos	%
1	8	0%
2	4.933	75%
3	1	0%
4	249	4%
5	1.375	21%
Total	6.566	100%

Elaborado por: El Autor

Como se puede observar en la tabla 7, la distribución de los casos en los grupos no es representativa para todos ellos. El 96% de los datos se conglomeran en dos grupos; por tanto se decide considerar más grupos para poder realizar la agrupación de los casos y la definición de los estratos del INS.

[2] El criterio de tomar 5 grupos se considera a partir de los estratos sociodemográficos definidos por el INEC en la Encuesta de Estratificación del Nivel Socioeconómico, 2011.

Por lo expuesto en el párrafo anterior, se realizó la agrupación con el algoritmo K-Medias con 6 grupos, pero los resultados no fueron buenos, lo mismo sucedió considerando 7 grupos. Luego, se intentó considerando 8 grupos y los resultados fueron mejores en términos de agrupación. Los resultados se presentan a continuación:

Tabla 8. Conglomerados a partir del algoritmo K-Medias con 8 grupos

Grupo	Recuento	%
1	4.048	61,7%
2	1	0,0%
3	11	0,2%
4	3	0,0%
5	197	3,0%
6	654	10,0%
7	67	1,0%
8	1.585	24,1%
	6.566	100,0%

Elaborado por: El Autor

En la tabla 8, se puede observar que la concentración en el grupo con mayor población (grupo 1) es ahora del 61,7%, el segundo grupo con mayor población (grupo 8) tiene el 24% de datos, el tercer grupo con mayor población (grupo 6) contiene el 10%, el cuarto grupo en tamaño (grupo 3) agrupa al 3% y finalmente existen 3 grupos con menos del 1% de los datos (grupo 2, 3 y 7). Para determinar cómo se está distribuyendo el INS en los 8 grupos se presenta la siguiente tabla:

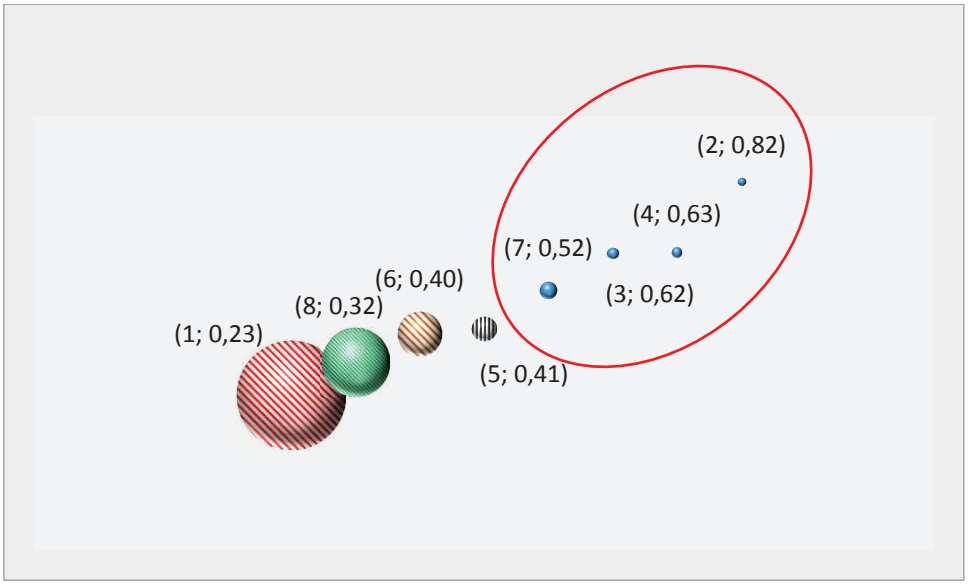
Tabla 9. Análisis descriptivo del INS por grupo

Grupo	N	Promedio	Mediana	Desviación típica	Mínimo	Máximo
1	4.048	0,23	0,19	0,14	0,00	0,57
8	1.585	0,32	0,37	0,15	0,03	1,00
6	654	0,40	0,44	0,14	0,06	0,67
5	197	0,41	0,47	0,14	0,11	0,67
7	67	0,52	0,57	0,14	0,15	0,72
3	11	0,62	0,63	0,08	0,43	0,78
4	3	0,63	0,60	0,12	0,51	0,76
2	1	0,82	0,82		0,82	0,82

Elaborado por: El Autor

En la tabla 9, se muestra el orden de los grupos en orden ascendente del promedio del INS. Así, se puede observar que el grupo 1 (el de mayor cantidad de población) es el que tiene el promedio de INS más bajo; mientras que los cuatro grupos con mayor INS (superior a 0,5) contienen al 1,2% de la población de estudiantes de la EPN. La distribución de los grupos de manera gráfica es la siguiente:

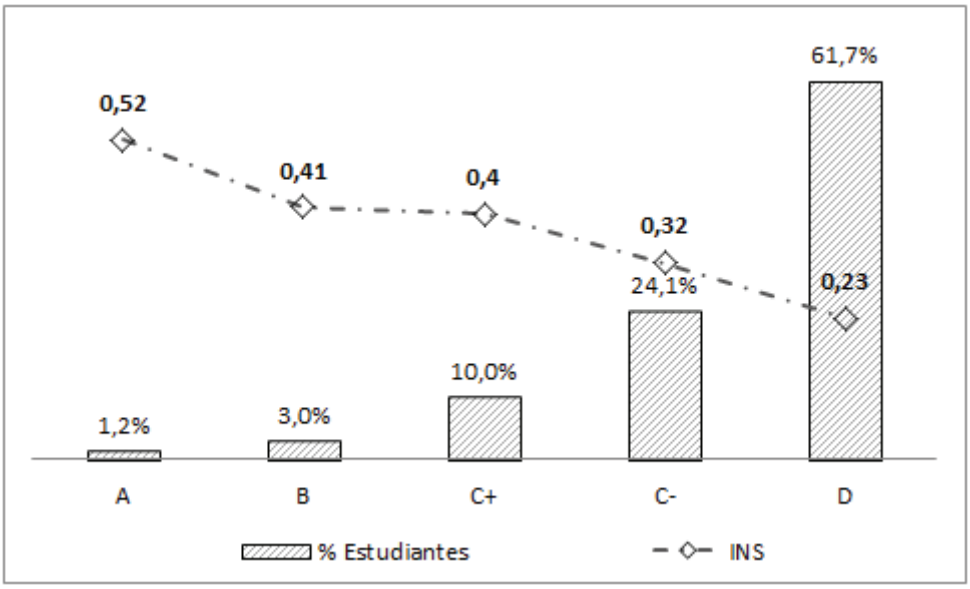
Figura 12. Distribución de la población en los grupos con el algoritmo K-Medias por el promedio del INS



Elaborado por: El Autor

En la figura 13, se muestra el grupo y el valor promedio del INS entre paréntesis. Los grupos 2, 3, 4 y 7 son los que tienen menor cantidad de registros, se procede a realizar una agregación de estos grupos en uno solo. Así, en este caso, los estratos quedan conformados por: Estrato A los casos que están en los grupos 2, 3, 4 y 7; Estrato B correspondiente al conglomerado 5 de la clasificación, Estrato C+ los correspondientes al conglomerado 6, Estrato C- es el conglomerado 8 y Estrato D aquellos que están en el conglomerado 1 de la clasificación.

Figura 13. Población estudiantil por Grupos e INS



Elaborado por: El Autor

CAPÍTULO 4
ASIGNACIÓN DE ESTRATOS SOCIOECONÓMICOS PARA ALUMNOS NUEVOS DE PREGRADO DE LA EPN

En el capítulo anterior se define la metodología para el cálculo del INS para los estudiantes de la EPN y se determinan los estratos en los que se dividirían los estudiantes; ahora, en este capítulo se muestran las metodologías que se utilizan para crear las reglas o propiedades que definen a cada estrato socioeconómico. En este caso, se consideran tres metodologías distintas: Árbol de decisión, Regresión Logística Multinomial y el SVM, la modelación se va a realizar con un total de 5.000 alumnos y se deja un grupo de 1.566 alumnos para poder aplicar las metodologías.

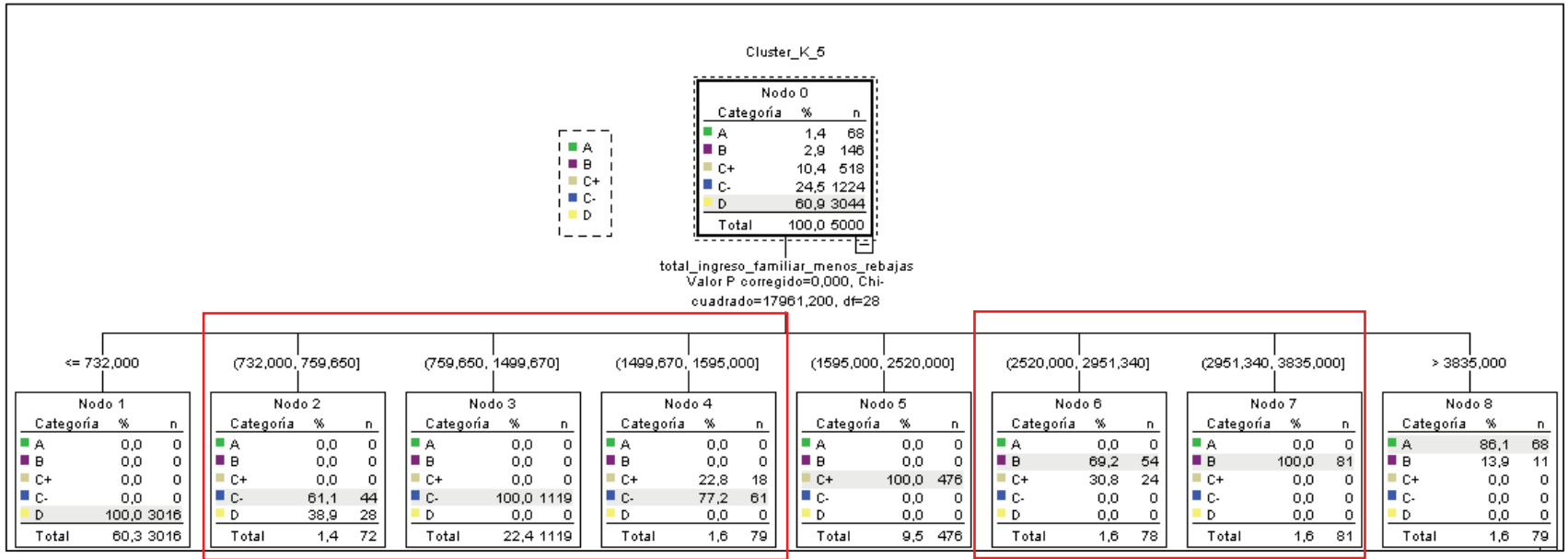
Los procesos de árbol de decisión y la regresión logística se realizan en el programa SPSS versión 20; mientras que el SVM se lo hace en R versión 3.3.0.

4.1 ÁRBOL DE DECISIÓN

La metodología del árbol de decisión es ampliamente utilizada para realizar una clasificación de individuos considerando un nodo objetivo (categorías previamente definidas); es por esto, que en este trabajo se lo va a utilizar para poder determinar las reglas de asignación de los estudiantes nuevos a uno de los estratos del INS.

En este caso, se aplica el árbol de decisión a la base de datos de 5.000 alumnos considerando las variables que se han definido dentro de este trabajo; sin embargo, solamente una variable es la que clasifica a los individuos, los resultados se muestran a continuación:

Figura 14. Árbol para clasificación de alumnos nuevos



Elaborado por: El Autor

El árbol de clasificación se determina solamente por la variable: *total de ingreso familiar menos rebajas* (ingreso neto de la familia); el árbol de clasificación propone las reglas para poder clasificar a un estudiante nuevo, las reglas son:

1. Estrato A: Estudiantes con ingreso familiar neto superior a USD 3.835 mensuales.
2. Estrato B: Estudiantes con ingreso familiar neto más de USD 2.520 y hasta USD 3.835 mensuales.
3. Estrato C+: Estudiantes con ingreso familiar neto más de USD 1.595 y hasta USD 2.520 mensuales.
4. Estrado C-: Estudiantes con ingreso familiar neto más de USD 732 y hasta USD 1.595 mensuales.
5. Estrato D: Estudiantes con ingreso familiar neto con hasta USD 732.

Una medida de la clasificación del procedimiento del árbol es la comparación entre los estratos definidos a través de la metodología utilizada para construir el INS y los casos clasificados por el árbol de clasificación; esto es:

Tabla 10. Tabla de confusión estratos INS vs Árbol de clasificación

		Árbol de clasificación					Total	% correcto
		A	B	C+	C-	D		
INS	A	9	5	0	0	0	14	64,3%
	B	3	48	0	0	0	51	94,1%
	C+	0	10	119	7	0	136	87,5%
	C-	0	0	0	361	0	361	100,0%
	D	0	0	0	12	992	1.004	98,8%
	Total	12	63	119	380	992	1.566	97,6%

Elaborado por: El Autor

El árbol de clasificación llega a un nivel de precisión del 97,6%; por tanto, se puede utilizar para la clasificación de nuevos estudiantes a la EPN, una vez que se haya definido el INS con la metodología presentada en este trabajo.

4.2 REGRESIÓN LOGÍSTICA MULTINOMIAL

Al igual que el árbol de clasificación, la regresión logística multinomial se puede utilizar para pronosticar la probabilidad de que un individuo pertenezca a un determinado grupo previamente definido. En este caso se utilizará para que a los nuevos estudiantes que ingresen a la EPN se les asigne un estrato del INS.

Para el modelo se toma como categoría de referencia al Estrato A; las variables que resultaron ser significativas dentro del modelo, son: *Total ingreso familiar menos rebajas*, *madre trabaja*, *colegio fiscal* y *total de propiedades y vehículos*; los coeficientes se muestran a continuación:

Tabla 11. Coeficientes de la regresión logística

		ESTRATOS			
		B	C+	C-	D
VARIABLES MODELOS	Intersección	88,89	4,71	11,00	2.389,71
	total ingreso familiar menos rebajas	0,005	0,004	0,002	0,001
	Madre_Trabaja=0	1,19	0,65	0,72	0,83
	Madre_Trabaja=1	0,00	0,00	0,00	0,00
	Colegio_Fiscal=0	-0,43	-0,63	-0,45	-0,62
	Colegio_Fiscal=1	0,00	0,00	0,00	0,00
	Total_Propiedades=0	-95,27	-8,27	-11,83	-2.387,47
	Total_Propiedades=1	-96,88	-9,11	-12,40	-2.388,30
	Total_Propiedades=2	-97,92	-10,49	-13,59	-2.389,73
	Total_Propiedades=3	-97,18	-9,58	-12,65	-2.388,85
	Total_Propiedades=4	-95,87	-9,79	-13,20	-2.389,42
	Total_Propiedades=5	0,00	0,00	0,00	0,00

Elaborado por: El Autor

Todos los coeficientes presentados en la tabla 11 son significativos con el 95% de confianza. Los signos de las variables corresponden a la lógica del análisis; es decir, la familia con mayores ingresos tendrá una probabilidad de pertenecer al Estrato A. Lo mismo sucede con las variables categóricas considerando que la comparación con la categoría de referencia de cada variable. Así, por ejemplo en el caso del Total de propiedades tienen una menor probabilidad de pertenecer al Estrato A, que aquellos que tienen 5 propiedades. También es necesario realizar el análisis de la tabla de clasificación de la regresión multinomial con respecto a la clasificación del INS para poder determinar el nivel de asertividad que tiene el modelo. Los resultados se presentan a continuación:

Tabla 12. Tabla de confusión de la regresión logística multinomial

	Regresión logística multinomial					
INS	A	B	C+	C-	D	Porcentaje correcto
A			13	1		0,0%
B			49	2		0,0%
C+		1	130	5		95,6%
C-	83	1	24	159	94	44,0%
D	112				892	88,8%
Total	12,5%	0,6%	15,8%	10,0%	61,1%	75,4%

Elaborado por: El Autor

Como se puede observar, el modelo presentado solamente clasifica correctamente al 75,4% lo que implica que es un buen modelo. Sin embargo, no tiene el nivel de clasificación del árbol de decisión de alrededor del (97,6%), aunque utiliza más variables que el proceso anterior.

La regla de asignación en este caso es: Un individuo nuevo se asignará al estrato para el cual el valor de probabilidad calculado a partir de los coeficientes del modelo sea el más grande. Para calcular la probabilidad que un individuo pertenezca al Estrato A:

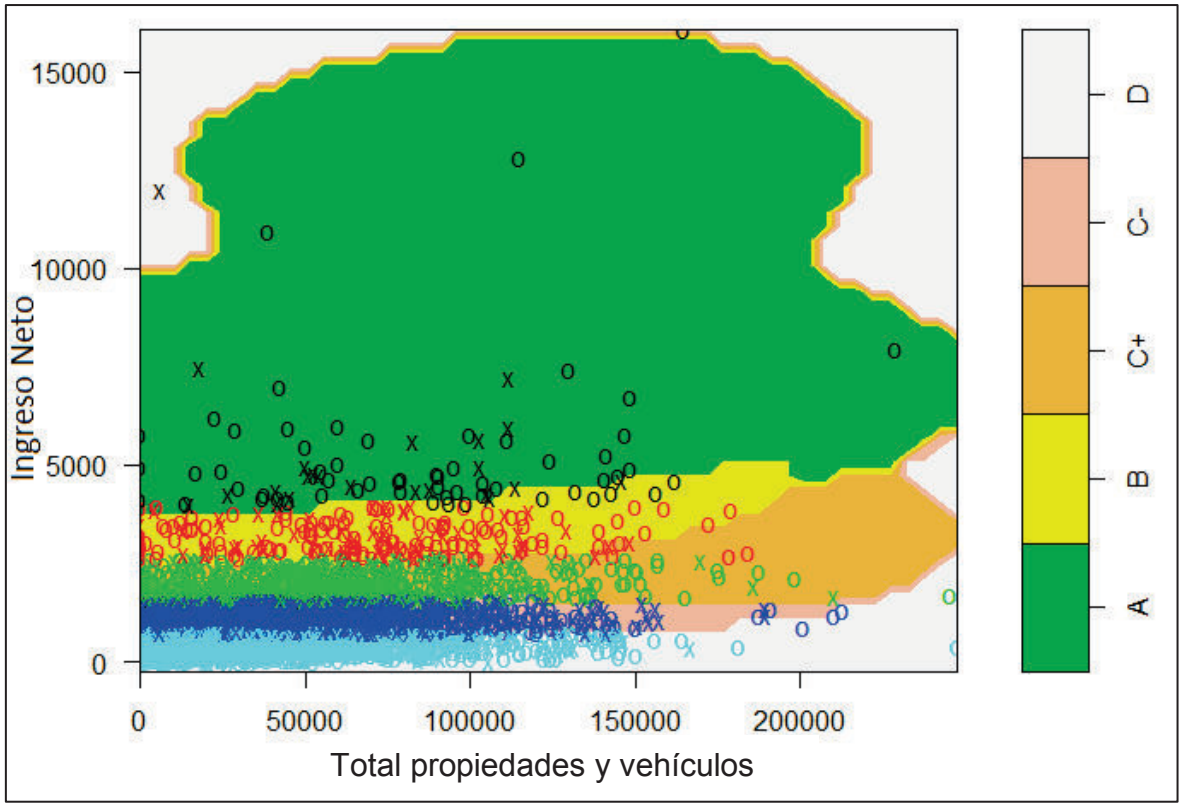
$$P(i \in A) = 1 - (P_B + P_{c+} + P_{c-} + P_D)$$

4.3 MÁQUINAS DE VECTORES DE SOPORTE

Como en los casos anteriores, las Maquinas de Vectores de Soporte (En inglés Support Vector Machines o SVM) se utilizará de igual manera para poder clasificar a los estudiantes nuevos que lleguen a la EPN y se les asigne un estrato del INS. Con este método se logró una efectividad del 93,36% con una función de ***kernel radial o gaussiano*** cuyo valor de gamma fue de 0,166. A partir de este análisis se identificó un total de 1.415 vectores de soporte para realizar la separación de los grupos. Dentro de este análisis se consideraron las variables: Número de miembros, Valor total de propiedades y vehículos, ingreso familiar menos rebajas, total propiedades y vehículos, Madre Trabaja, Colegio Fiscal.

A continuación se muestra el gráfico de la variable del ingreso familiar (Ingreso familiar menos rebajas) por el total de propiedades de la familia:

Figura 15. Gráfico de dispersión entre Ingreso neto familiar y total de propiedades por clasificación SVM



Elaborado por: El Autor

En la figura, la escala de colores indica el estrato al cual pertenecen los datos en la clasificación. Como se puede observar y concluir, el Estrato A es el más pequeño en número, pero a su vez en que tiene una mayor variabilidad en los datos; los otros estratos son más homogéneos con respecto a la variable *Ingreso Neto*, aunque más dispersos en la variable *Total de propiedades y Vehículos*.

Adicional, se presentan la tabla de clasificación entre el SVM y los estratos definidos para el INS:

Tabla 13. Tabla de clasificación del SVM vs INS

		SVM					Total	% Correcto
		A	B	C+	C-	D		
INS	A	16	2	0	0	0	18	89%
	B	1	44	7	0	0	52	85%
	C+	0	0	140	21	0	161	87%
	C-	0	0	1	352	34	387	91%
	D	0	0	0	14	934	948	99%
Total		17	46	148	387	968	1.566	95%

Elaborado por: El Autor

El nivel de asertividad del SVM es del 94,5% (95% aproximadamente) Esto indica que es un buen modelo para determinar la clasificación de los individuos en los estratos definidos por el INS.

Una de las debilidades del SVM es que no permite -de manera fácil- clasificar a un individuo, esto solamente se lo puede realizar con un software adecuado en el que se haya calculado el SVM. Es decir, para clasificar a un nuevo estudiante de la EPN se debe ingresar al *Software* especializado.

4.4 COMPARACIÓN DE LOS MÉTODOS DE CLASIFICACIÓN

En este proyecto se han desarrollado 3 métodos de clasificación automática de estudiantes, a continuación se muestra las fortalezas y debilidades de cada uno, y así decidir cuál de ellos es recomendable utilizar; esto es:

Tabla 14. Tabla de comparación de métodos de clasificación automática

	% Correcto de clasificación	Nivel de dificultad de interpretación	Nivel de dificultad de implementación	Número de variables
Árbol de clasificación	97,60%	Fácil	Fácil	1
Regresión logística multinomial	74,70%	Medio	Medio	5
SVM	94,56%	Difícil	Difícil	6

Elaborado por: El Autor

Como se puede ver, el método que mejor clasifica a los individuos es el ***Árbol de Clasificación***, además es un método de fácil interpretación e implementación, las reglas de clasificación derivadas del árbol se pueden agregar en cualquier *software* comercial, como el Excel o el SQL, y permite una clasificación inmediata de individuos. Sin embargo, en este caso, solamente utiliza una variable para realizar la clasificación (Ingreso familiar neto o Ingreso total del hogar menos rebajas); esto puede ser una debilidad al momento de no contemplar otras variables que pueden ayudar dentro de la clasificación.

El SVM es el método más complicado de interpretar y por ende de implementar, no permite determinar las reglas de clasificación de un nuevo estudiante ni tampoco se puede hacer en cualquier *software*, su nivel de clasificación es del 94,56%. Este sería la última opción entre los modelos desarrollados.

Por otro lado, la regresión logística multinomial es la que tiene el nivel más bajo de asertividad en la clasificación (75,4%); su nivel de interpretación y de implementación es de dificultad media ya que involucra cálculos adicionales tanto en la interpretación como en la implementación. Sin embargo, este modelo no categoriza de manera adecuada a los estratos A y B.

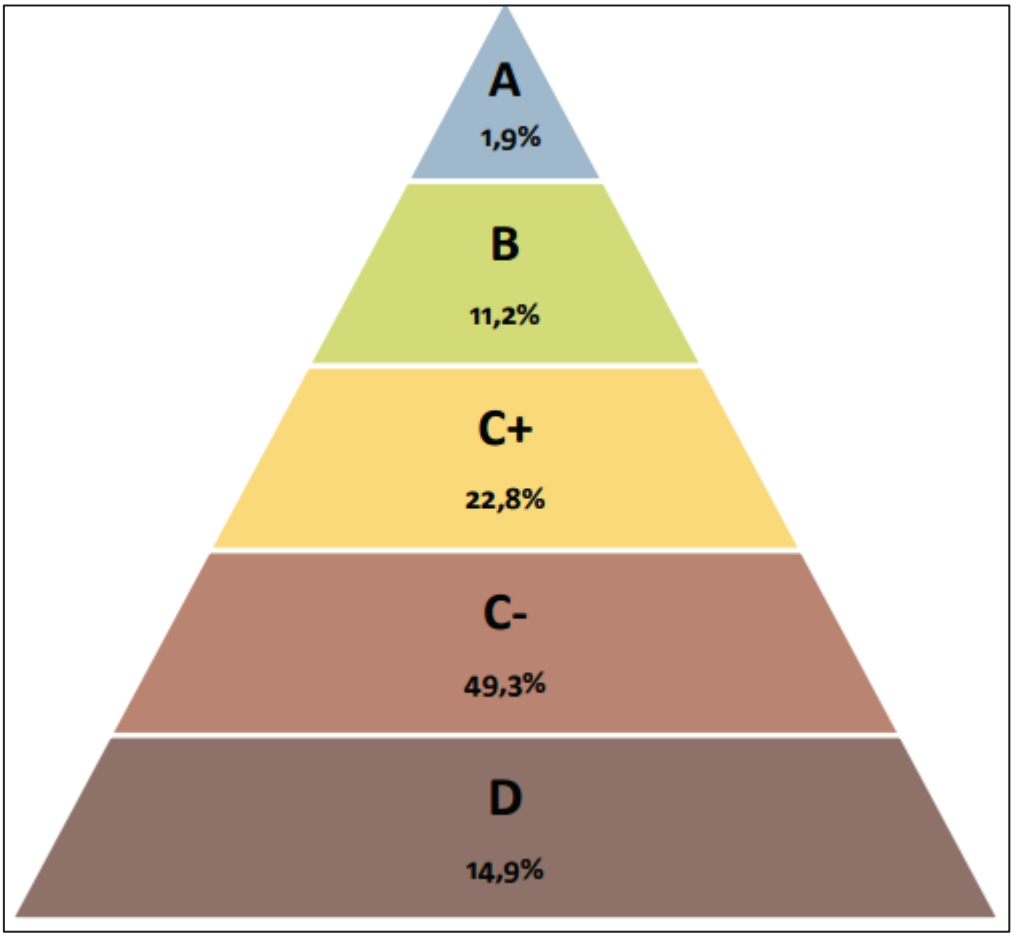
A manera de conclusión, en este proyecto, se considera que el árbol de clasificación es el mejor método para asignar a los nuevos estudiantes de la EPN a uno de los estratos definidos en el INS.

CAPÍTULO 5

INCIDENCIA DE LOS ESTRATOS SOCIO ECONÓMICOS EN LA ASIGNACIÓN DE BECAS Y COBRO DE ARANCELES

Para iniciar con este capítulo, se presentan los estratos socioeconómicos definidos por el INEC a través de la ***Encuesta de Estratificación del Nivel Socioeconómico 2011*****,** en la que se determinan los estratos de la siguiente manera:

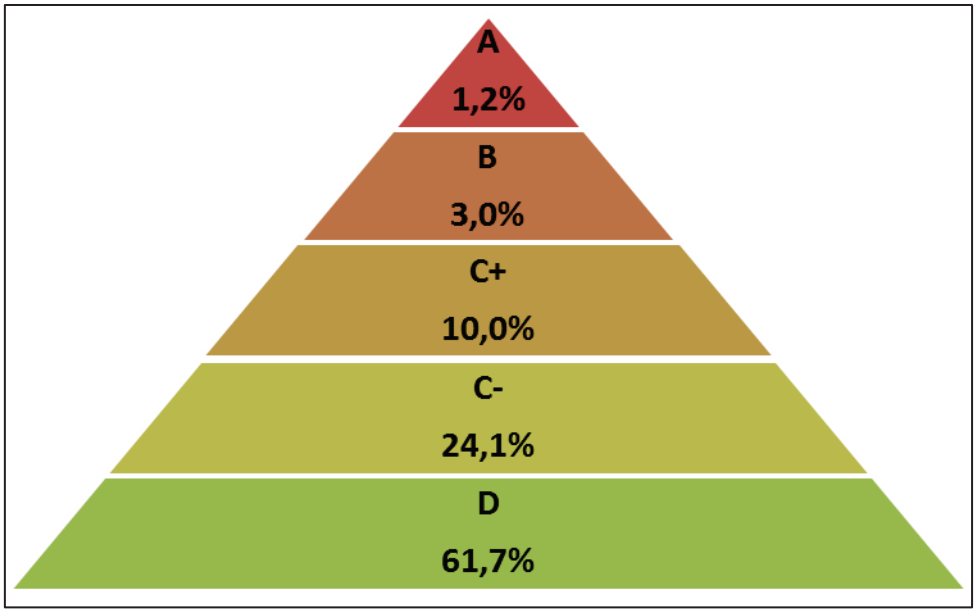
Figura 16. Estratos socioeconómicos definidos por el INEC



Fuente: INEC
Elaborado por: INEC

Para el caso de este estudio se tiene que la distribución de los estratos se encuentra de la siguiente manera:

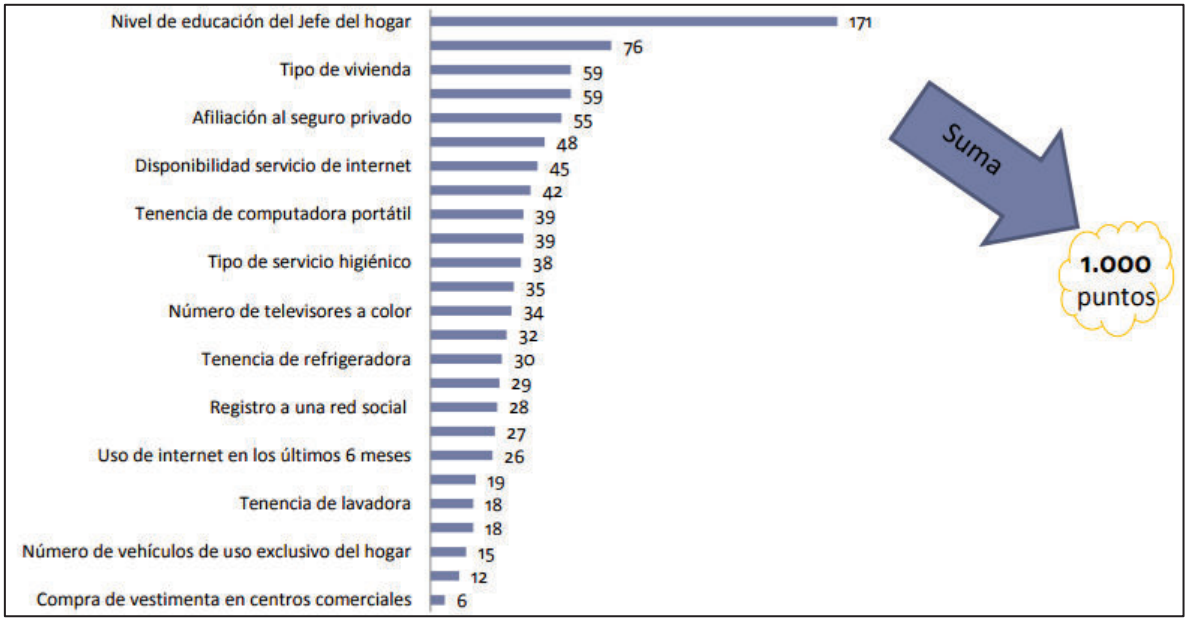
Figura 17. Estratos socioeconómicos definidos por el INS



Elaborado por: El Autor

Como se puede observar en las figuras anteriores, la distribución se realiza de manera similar; esto es una pequeña cantidad de población en los estratos altos (A y B) y la mayor concentración en los estratos bajos; esto ratifica la validez de los estratos determinados en este proyecto. Sin embargo, cabe recalcar que el estrato más bajo definido en este proyecto es más grande (en términos de porcentaje de concentración) que el estrato C-; mientras que el INEC tiene las proporciones al revés. Esto se puede explicar ya que la variable que mejor discrimina a los grupos en la estratificación del INEC es el nivel de educación del jefe de hogar (ver figura 17) y algunas otras variables propias de la encuesta; es por esta razón que se pueden mejorar los estratos definidos en este proyecto.

Figura 18. Importancia de las variables en la estratificación del INEC



Fuente: INEC
Elaborado por: INEC

Una vez que se ha expuesto la validez de los estratos definidos en este proyecto, es importante analizar la regulación sobre las becas en las universidades del Ecuador, es obligación de cada universidad pública tener por lo menos el 10% de estudiantes becados (Art. 77 de la Ley Orgánica de Educación Superior); en este caso, con este proyecto se provee una regla de asignación de las becas para los estudiantes de la EPN. Se debe asignar becas dando prioridad a aquellos estudiantes con los valores del INS más bajos.

Ahora, es necesario comparar la distribución de la actual estratificación que utiliza la EPN con respecto a la clasificación que se propone en este proyecto; así, se tiene:

Tabla 15. Comparación de estratificaciones: Quintiles EPN – INS (número de casos)

		QUINTIL EPN						TOTAL
		5	4	3	2	1	S/A*	
INS	A	82	-	-	-	-	-	**82**
	B	194	2	-	-	-	1	**197**
	C+	572	78	2	-	-	2	**654**
	C-	328	868	324	55	4	6	**1.585**
	D	81	362	826	1.291	1.475	13	**4.048**
TOTAL		**1.257**	**1.310**	**1.152**	**1.346**	**1.479**	**22**	**6.566**

*Son estudiantes sin datos

Elaboración: Autor

Tabla 16. Comparación de estratificaciones: Quintiles EPN – INS (Porcentaje)

		QUINTIL EPN						TOTAL
		5	4	3	2	1	S/A*	
INS	A	1,2%	0,0%	0,0%	0,0%	0,0%	0,0%	**1,2%**
	B	3,0%	0,0%	0,0%	0,0%	0,0%	0,0%	**3,0%**
	C+	8,7%	1,2%	0,0%	0,0%	0,0%	0,0%	**10,0%**
	C-	5,0%	13,2%	4,9%	0,8%	0,1%	0,1%	**24,1%**
	D	1,2%	5,5%	12,6%	19,7%	22,5%	0,2%	**61,7%**
TOTAL		**19,1%**	**20,0%**	**17,5%**	**20,5%**	**22,5%**	**0,3%**	**100,0%**

*Son estudiantes sin datos

Elaboración: Autor

Los quintiles de la EPN tienen un orden descendente con respecto al ingreso per cápita familiar; es decir, en el quintil 5 están los estudiantes con ingreso per cápita más alto y en el quintil 1 los estudiantes con el ingreso per cápita más bajo. Así, en este caso, 1.616 estudiantes (de la diagonal de la tabla) coinciden en el orden establecido por ambos criterios de estratificación, esto representa el 24,6% de la base analizada.

Si además se consideran las diagonales secundarias inferior y superior para comparar las clasificaciones (debido al error que se comete al utilizar las técnicas estadísticas), se tiene que 3.507 estudiantes (53,41%) coinciden en las dos metodologías de estratificación. Esto puede deberse a que la metodología de estratificación de la Politécnica se basa en discriminar grupos

de igual tamaño; por tanto, si dos estudiantes tienen el mismo nivel de ingreso per cápita familiar puede estar en grupos diferentes dado que pueden estar ordenados de tal forma que los puntos de corte del quintil los separe.

También puede suceder que dos estudiantes con ingreso familiar per cápita similar tengan un estrato diferente por las propiedades que posee cada uno.

La zona azul de la tabla suma 1.639 estudiantes que están en los quintiles superiores de la EPN, y también en los grupos más vulnerables de acuerdo al INS, esto equivale a que el 25% de los estudiantes están mal clasificados con respecto al INS.
Comparado las becas por situación económica otorgadas hasta el semestre 2016 – A (Marzo 2016) con respecto a los quintiles y a los estratos del INS, se tiene lo siguiente:

Tabla 17. Becas asignadas en la EPN por INS

INS	SIN BECA	CON BECA	TOTAL	% BECAS
A	81	1	82	0,13%
B	196	1	197	0,13%
C+	652	2	654	0,27%
C-	1.547	38	1.585	5,06%
D	3.339	709	4.048	94,41%
	5.815	751	6.566	

Elaborado por: El Autor

Se puede observar que con respecto al INS el 99,47% de las becas están asignadas en los dos estratos con puntuaciones más bajas.

Tabla 18. Becas asignadas en la EPN por QUINTIL

QUINTIL	SIN BECA	CON BECA	TOTAL	% BECAS
5	1.244	13	1.257	1,73%
4	1.259	51	1.310	6,79%
3	1.044	108	1.152	14,38%
2	1.122	224	1.346	29,83%
1	1.128	351	1.479	46,74%
0	18	4	22	0,53%
	5.815	751	6.566	

Elaborado por: El Autor

Para la clasificación por Quintiles, se observa que se otorga becas en todos los grupos y en los dos más bajos apenas llega el 44,21%.

Ahora, comparando de manera simultánea la metodología de quintiles de la EPN con el INS en cuanto a la asignación de becas por situación económica, se tiene:

Tabla 19. Becas económicas por Quintiles EPN – INS

		QUINTIL EPN						TOTAL
		5	**4**	**3**	**2**	**1**	**S/A***	
INS	**A**	1	0	0	0	0	0	**1**
	B	1	0	0	0	0	0	**1**
	C+	2	0	0	0	0	0	**2**
	C-	4	15	11	5	2	1	**38**
	D	5	36	97	219	349	3	**709**
TOTAL		**13**	**51**	**108**	**224**	**351**	**4**	**751**

Como se puede observar, no existe una correlación directa entre ambas metodologías, sin embargo, es importante recalcar que la mayor cantidad de becas está concentrada en el nivel D del INS y en los quintiles 1 y 2 de la clasificación de la EPN.

Esto muestra que el INS está bien construido y determina de manera adecuada a los estudiantes a los que se les debe otorgar una beca por situación socioeconómica baja.

Como una referencia se analiza la distribución de becas por un motivo diferente a la situación económica, en este caso, se tiene que hay un total de 1.167 becas otorgadas, 751 por situación económica y 416 por otros motivos, los datos se muestran en la siguiente tabla:

Tabla 20. Comparación de asignación de otras becas: Quintiles EPN - INS

		QUINTILES					TOTAL
		1	2	3	4	5	
INS	A					12	12
	B					32	32
	C+				6	50	56
	C-		3	26	70	32	131
	D	61	58	36	26	4	185
	TOTAL	**61**	**61**	**62**	**102**	**130**	416

Elaborado por: El Autor

Se puede notar con respecto a los Quintiles EPN, que las becas que no son por situación económica están en un 55,77% en los quintiles 4 y 5, y el 44,23% en los quintiles de 1 al 3. Con respecto al INS, el 10,6% de las becas que no son por situación económica se encuentran en los grupos de mejor INS, mientras que el 89,4% se ubican en los grupos de menor INS, concentrando el 44,5% de estas becas el estrato D.

Por otro lado, se tiene que según la Ley Orgánica de Educación Superior, un alumno pierde parcialmente la gratuidad de la educación cuando pierde una o varias materias, en cuyo caso deberá pagar los aranceles correspondientes al número de créditos que deba repetir el estudiante, considerando un análisis socioeconómico del hogar del estudiante. En este caso, se toman como referencia los pagos realizados por los estudiantes durante el período 2016-A y se realiza el cálculo de algunos indicadores que servirán de guía para determinar el ingreso de la EPN a través de los estratos definidos en este proyecto:

a. Se calcula el índice de repetición de los estudiantes por estrato, que consiste en el cociente entre los estudiantes que repitieron alguna materia en el período anterior sobre el total de estudiantes en el estrato.
b. El costo promedio del arancel por cada crédito perdido en el período anterior, esto se calcula como el pago total realizado por el estudiante para el número de créditos

perdidos en el período anterior. Esta aproximación se la realiza ya que no se pudo obtener acceso al costo por crédito y de matrícula de cada estudiante.

Con esto se tienen los siguientes resultados:

Tabla 21. Pago de aranceles promedio en la EPN por INS

INS	Número de estudiantes	Índice de repitencia	Promedio de créditos perdidos	Estudiantes que pierden al menos 1 crédito	Arancel promedio por crédito (USD)	Ingreso estimado por aranceles a la EPN (USD)
A	82	22%	9	18	27	4.631
B	197	19%	9	38	23	8.188
C+	654	22%	10	143	21	29.694
C-	1.585	23%	11	366	19	75.553
D	4.048	30%	11	1.219	10	136.530
TOTAL	6.566			1.784		254.596

Elaborado por: El Autor

Como se puede observar, el pago promedio por crédito que realizan los estudiantes que han repetido una matrícula en la EPN corresponde a la realidad de los hogares de los estudiantes; es decir, estudiantes categorizados en el Estrato A pagan un valor más alto que aquellos estudiantes del estrato D. Es importante recalcar el hecho que los estudiantes del estrato D, tienen un índice de repitencia más alto que los estudiantes de los otros estratos.

Finalmente, con esta estimación de aranceles y los estratos definidos, se esperaría que la EPN reciba USD 254.596 por concepto de estudiantes que no se acogen al beneficio de la ley de gratuidad de la educación.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

En este trabajo se ha logrado construir un índice de clasificación socioeconómica para los estudiantes, matriculados en el periodo 2016-A, de la Escuela Politécnica Nacional. La metodología desarrollada permite establecer grupos homogéneos de estudiantes (estratos) con los cuales se puede trabajar para establecer políticas y programas tales como: Becas, Descuentos y Cobro de Rubros. Además, los estratos definidos en este trabajo, son consistentes con los resultados entregados por el Instituto Nacional de Estadística y Censos (INEC) en cuanto a la composición de los estratos definidos.

Las principales variables identificadas para la construcción del INS están en las dimensiones de patrimonio familiar, características del hogar y estado laboral de la madre, relacionadas todas ellas con la capacidad económica de la población estudiantil analizada; esto muestra que el INS puede ser usado como una medida adecuada para la estratificación de los estudiantes por nivel socioeconómico.

De las metodologías utilizadas para determinar las reglas de asignación de los estudiantes a cada estrato se puede concluir que el Árbol de Clasificación es el mejor método para asignar a los estudiantes nuevos a uno de los estrados definidos por el INS. Una de las ventajas de este método es su fácil interpretación e implementación y tiene un nivel de asertividad del 97%, aproximadamente. Cabe recalcar que el Árbol considera únicamente una variable para realizar la clasificación; mientras que el ACPC considera todas las variables en el análisis y por lo tanto se constituye como un método más robusto.

Se puede observar que aproximadamente el 11% de los estudiantes poseen beca otorgada por la EPN, de estos el 99,47% están el estrato con puntuación más baja del INS y; por tanto, los estratos construidos poseen consistencia con lo dispuesto en la LOES y el reglamento interno de la EPN en los cuales se señala que al menos el 10% de estudiantes matriculados deben estar

incluidos dentro de programas de crédito educativo, becas y ayudas económicas considerando el nivel socioeconómico del hogar del estudiante.

RECOMENDACIONES

La información obtenida para el desarrollo de este trabajo fue proporcionada por la Dirección de Gestión de Información de la EPN la cual es suministrada por los estudiantes como parte de los procesos que se deben seguir para ser matriculados. En este punto, es recomendable definir y mejorar los sistemas de validación de datos incluyendo, en la medida de lo posible, información pública disponible en la Dirección Nacional de Registro de Datos Públicos (DINARDAP), Instituto Ecuatoriano de Seguridad Social IESS, Buró de crédito, Sistema de Rentas Internas SRI, y otros, lo que permitirá mejorar la propuesta del INS y los estratos definidos.

Por otro lado, se recomienda verificar la posibilidad de agregar a las bases de datos variables con un mayor nivel de desagregación. Una de estas variables relevantes es la escolaridad del jefe de hogar (padre o madre) ya que es la variable con mayor importancia definida por el INEC en la estratificación de la población nacional.

Se recomienda, también, que el monitoreo del índice propuesto en este proyecto se lo realice cada año con el fin de encontrar cambios en el comportamiento de la población estudiantil y de ello ocurrir se deberá actualizar el INS, sin embargo, si cambian las condiciones de la población o las variables medidas cambian de manera abrupta se debe considerar la actualización en un período más corto. Para realizar el monitoreo de manera rápida y estandarizada se recomienda su automatización.

Se recomienda también, realizar un análisis más profundo para determinar los estratos de los estudiantes con las variables que se logren levantar cuando un estudiante llega a la EPN, esto permitirá tener estimaciones más cercanas a la realidad de los estudiantes y así tener una mejor visión de las políticas de becas, ayudas económicas y aranceles cobrados por la EPN.

BIBLIOGRAFÍA

Álvarez, R. (2015). Estratificación socioeconómica de la población urbana de la provincia de Santa Elena. *Revista Científica y Tecnológica UPSE*, *2*(2), 1-5.

Andrew, A. M. (2000). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods by Nello Christianini and John Shawe-Taylor, Cambridge University Press, Cambridge, 2000, xiii+ 189 pp., ISBN 0-521-78019-5).

Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *Analyst*, *135*(2), 230-267.

Bull, S. B., & Donner, A. (1987). The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. *Journal of the American Statistical Association*, *82*(400), 1118-1122.

Cayuela, L. (2011). Análisis multivariante. *Área de Biodiversidad y Conservación, Universidad Rey Juan Carlos, Departamental*.

Cawley, G. C., Talbot, N. L., & Girolami, M. (2007). Sparse multinomial logistic regression via bayesian l1 regularisation. *Advances in neural information processing systems*, *19*, 209.

Corso, C. L. (2009). Aplicación de algoritmos de clasificación supervisada usando Weka. *Córdoba: Universidad Tecnológica Nacional, Facultad Regional Córdoba*.

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines.

Cuadras, C. M. (2007). *Nuevos métodos de análisis multivariante*. CMC Editions.

Delhy, C., & Fernando, L. (2014). *Construcción de un modelo de clasificación socioeconómica para el descuento de las colegiaturas de los estudiantes de la Flacso* (Master's thesis, Quito, Ecuador: Flacso Ecuador).

Espinoza Villagómez, A. C., & Guevara Escobar, P. K. (2013). *Modelo de segmentación socioeconómica usando análisis de componentes principales categóricos con base en el censo de población y vivienda 2010* (Bachelor's thesis).

FactoClass, P., Pardoa, C. E., & Del Campob, P. C. (2007). Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass. *Revista colombiana de estadística*, *30*, 231-245.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, *16*(10), 906-914.

Fung, G. M., & Mangasarian, O. L. (2005). Multicategory proximal support vector machine classifiers. *Machine learning*, *59*(1-2), 77-97.

Galbiati, J. Anàlisis de conglomerados. *Recuperado de http://ecosdelaeconomia. files. wordpress. com/2011/09/conglomerados.pdf*.

Greene, W. H. (2007). Econometric analysis 6th edition. *International edition, New Jersey: Prentice Hall*.

Erazo, H. (2013). EL GASTO EQUIVALENTE DE LOS HOGARES ECUATORIANOS EN FUNCIÓN DE SU CONSUMO ALIMENTARIO, COMPOSICIÓN Y TAMAÑO, SEGÚN LA ENCUESTA DE CONDICIONES DE VIDA 2006 (Tesis de pregrado). Escuela Politécnica Nacional, Quito, Ecuador.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

IBM Corporation, (2011). Manual del usuario del sistema básico de IBM SPSS Statistics 20 recuperado de: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Core_System_Users_Guide.pdf

Instituto Nacional de Estadísticas y Censos del Ecuador. Resumen Metodológico y Principales Resultados de la Encuesta Nacional de Ingresos y Gastos de Hogares Urbanos y Rurales (ENIGHUR). 2012(1):5-6.

Jobson, J. (2012). *Applied multivariate data analysis: volume II: Categorical and Multivariate Methods*. Springer Science & Business Media.

Krishnapuram, B., Carin, L., Figueiredo, M. A., & Hartemink, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, *27*(6), 957-968.

Ley Orgánica de Eduación Superior., *Registro Oficial N 298., Quito – Ecuador., Octubre 2010*.

Luts, J., Ojeda, F., Van de Plas, R., De Moor, B., Van Huffel, S., & Suykens, J. A. (2010). A tutorial on support vector machine-based methods for classification problems in chemometrics. *Analytica Chimica Acta*, *665*(2), 129-145.

Meyer, D. Support Vector Machines—the Interface to libsvm in package e1071.(2014). *Recuperado de: http://cran. r-project. org/web/packages/e1071/vignettes/svmdoc. pdf*.

Mitchell, T. (1997). *Machine Learning*. Ithaca, NY: McGraw-Hill.

Mulay, S. A., Devale, P. R., & Garje, G. V. (2010). Intrusion detection system using support vector machine and decision tree. *International Journal of Computer Applications*, *3*(3), 40-43.

Peña, D. (2002). *Análisis de datos multivariantes*. Madrid: McGraw-Hill.

Rojo, J. *Arboles de Clasificación y Regresión*. Recuperado de http://humanidades.cchs.csic.es/cchs/web_UAE/tutoriales/PDF/AnswerTree.pdf

ANEXOS

ANEXO A.

CONCEPTOS DE COMPONENTES PRINCIPALES CATEGÓRICOS

Este anexo es una adaptación de Peña (2002).

MATRICES COMPATIBLES CON MÉTRICAS EUCLÍDEAS

Para poder reconstruir la matriz Y a partir de la matriz T y esta a partir de la matriz D. Es necesario que los valores propios de la matriz T, que se construye a partir de la matriz D original, sean no negativos.

Sea una matriz D, se define que esta matriz es compatible con una métrica euclídea si la matriz de similitud T, que se construye a partir de ella, como sigue, es semidefinida positiva:

$$T = -\frac{1}{2}PDP, \text{ donde: } P = \left(I - \frac{1}{n}11'\right)$$

Esta condición es necesaria y suficiente; es decir, si la matriz D se ha construido a partir de una métrica euclídea, entonces T es semidefinida positiva, y si la matriz T tiene esta propiedad es posible encontrar una métrica euclídea que reproduzca la matriz D, Peña (2002).

CONSTRUCCIÓN DE LAS COORDENADAS PRINCIPALES

En general, no siempre la matriz D es compatible con una métrica euclídea; sin embargo, si la matriz de similitud obtenida a partir de D, tiene los p valores propios más grandes positivos, y si los restantes $n - p$ valores propios no nulos son mucho menores que los demás, se puede obtener una representación de los puntos con los p vectores propios correspondientes.

El procedimiento para obtener las *coordenadas principales* es:

1. Suponga que se tiene una matriz de distancias al cuadrado D.

2. Se construye la matriz T.

3. Calcular los valores propios de la matriz T. Tomar los ℓ primeros valores propios; tal que, los $n - \ell$ valores propios restantes sean aproximadamente 0.
   Dado que $P1 = 0$, donde 1 es un vector propio de "unos", la matriz Q tiene rango máximo $n - 1$ y siempre tendrá el vector propio 1 unido al valor propio cero, Peña (2002).

4. Se calcula:

$$v_i\sqrt{\lambda_i}$$

   donde,

   λ_i: valor propio de T

   v_i: es el vector propio unitario asociado a λ_i

   Entonces, T se puede aproximar por:

$$T \approx \left(V_\ell \Lambda_\ell^{\frac{1}{2}}\right)\left(\Lambda_\ell^{\frac{1}{2}} V_\ell'\right)$$

   donde,

   V: es una matriz de orden $n\,x\,p$ que contiene los vectores propios correspondientes a valores propios no nulos de T.

   Λ: es una matriz de orden $p\,x\,p$ que contiene los valores propios.

Finalmente, se debe tomar como coordenadas de los puntos de las variables:

$$Y_\ell = V_\ell \Lambda_\ell^{\frac{1}{2}}$$

La precisión de la aproximación de T a partir de los p valores propios positivos de la matriz de similitud, se da a través del siguiente coeficiente:

$$a_{1,p} = \frac{\sum_1^p \lambda_i}{\sum_1^p |\lambda_i|} * 100$$

ANEXO B.
ALGORITMO DE PARTICIÓN: K-MEDIAS

Este anexo es una adaptación de Peña (2002).

Algoritmo K – MEDIAS

El criterio de aglomeración óptimo que se utiliza en este algoritmo corresponde a la minimización de la suma de cuadrados de las distancias dentro de los grupos para todas las variables, se describe por:

$$\min SC = min \sum_{k=1}^{K} \sum_{j=1}^{p} \sum_{i=1}^{n_k} \left(x_{ijk} - \bar{x}_{jk}\right)^2$$

$$= min \sum_{k=1}^{K} \sum_{j=1}^{p} n_k \left(x_{jk} - \bar{x}_{jk}\right)^2$$

$$= min \sum_{k=1}^{K} \sum_{j=1}^{p} n_k s_{jk}^2$$

donde,

x_{ijk}: es el valor de la variable j en el elemento i del grupo k

$\bar{x}_{jk}$: es la media de la variable j en el grupo k

n_k: es el número de elementos del grupo k

s_{jk}^2: es la varianza de la variable j en el grupo k

Las varianzas de los grupos son una medida de la dispersión de la clasificación y al minimizarlas se obtienen grupos más homogéneos.

Un método alternativo para lograr homogeneidad en los grupos es minimizar las distancias al cuadrado entre los puntos y los centros de grupo; si se mide la distancia de manera euclídea, el criterio se escribe como sigue:

$$min \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)' (x_{ik} - \bar{x}_k) = min \sum_{k=1}^{K} \sum_{i=1}^{n_k} d^2(i,k)$$

donde,

$d^2(i,k)$: es el cuadrado de la distancia euclídea entre el elemento i del grupo k y la media del grupo

El resultado anterior se puede escribir en función de la traza, de la siguiente manera:

$$min \sum_{k=1}^{K} \sum_{i=1}^{n_k} tr\left(d^2(i,k)\right) = \min tr \left[\sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k) (x_{ik} - \bar{x}_k)' \right]$$

por tanto,

$$\min tr\left(d^2(i,k)\right) = \min SC$$

Este criterio, se denomina el *criterio de la traza* (Ward 1963) y coincide con el de la suma de cuadrados de las distancias internas.

En la práctica, dado que minimizar la varianza para todas las particiones posibles de los datos es casi imposible (a menos que *n* sea pequeño), es preferible la utilización del criterio de la traza; por tanto, el algoritmo de k – medias funciona de la siguiente manera:

1. Tomar una partición inicial
2. Comprobar si moviendo alguno de los elementos se reduce la $tr\left(d^2(i,k)\right)$
3. Si es posible reducir la $tr\left(d^2(i,k)\right)$, se recalcula las medias de los dos grupos afectados por el cambio y volver al paso 2; caso contrario, terminar.

Este procedimiento requiere que se fije el número de grupos iniciales. No existe una justificación teórica que sugiera el número de grupos a seleccionarse; sin embrago, se puede considerar el siguiente cociente:

$$F = \frac{SC(K) - SC(K+1)}{SC(K+1)/(n-K-1)} \quad con\ K = 0, 1, 2, \ldots$$

Hartigan (1975), sugiere que se agregue un grupo adicional si este cociente es mayor a 10.

ANEXO C.
CONCEPTOS DE ÁRBOLES DE CLASIFICACIÓN

Esta sección corresponde a una adaptación del IBM Knowledge Center en el desarrollo de su herramienta Answer Tree 2015 y también se ha adaptado del documento Arboles de Clasificación y Regresión por José Manuel Rojo (2006).

Los árboles de clasificación son técnicas de explotación de datos que consisten en estudiar grandes cantidades de datos con el fin de encontrar patrones no triviales. Representan una serie de pautas basadas en ciertas variables explicativas que se muestran según se recogen en el árbol.

Los árboles se construyen a través de un algoritmo que va dividiendo los registros de la base de datos en nodos de forma recursiva, de manera que con cada subdivisión las frecuencias relativas de las categorías de la variable dependiente vayan tendiendo a 0 o a 1.

Es necesario también, considerar algunas definiciones que sirven para entender de mejor manera el método, esta sección se ha adaptado del documento Arboles de Clasificación y Regresión por José Manuel Rojo (2006):

- **Grafo.** Un grafo G es un par de conjuntos (N, A), donde N representa un conjunto cualquiera y A es un subconjunto de pares de N. Al conjunto N se le suele denominar el conjunto de *Nodos* y al conjunto A se le denomina como el conjunto de *Aristas*.

- **Camino.** Se denomina camino a una sucesión de nodos unidos por aristas de forma que no se repita ninguna arista.

- **Grafo conexo.** Se dice que un grafo $G = (N, A)$ es conexo si para cualquier par de nodos pertenecientes a un grafo existe un camino que los une.

- **Circuito.** Un circuito es un camino que inicia y termina en el mismo nodo sin repetir ninguna arista.

- **Grado de incidencia.** Es el número de aristas que llegan a un nodo.

- **Árbol.** Se dice que el grafo $G = (N, A)$ es un árbol si se verifica que:

  a. Es conexo
  b. No tiene circuitos

  Un árbol tiene un nodo inicial que se denomina *Nodo Raíz* y a partir de él, los demás nodos se denominan *Nodos Terminales*.

- **Árbol binario.** Se dice que el Árbol $F = (N, A)$ es un árbol binario si existe un nodo que está conectado por dos aristas y el resto de nodos están conectados por uno o tres aristas.

Medidas para encontrar la mejor clasificación de un árbol

Tenemos el Índice de Gini y el Índice Binario

Índice de Gini

El índice de Gini en el nodo t se define como:

$$g(t) = 1 - \sum_{i=1}^{k} p(i/t)^2$$

donde,

i: representa las distintas categorías de la clase criterio.

Cuando todos los casos del nodo t pertenecen a la misma categoría, el índice de Gini toma el valor cero, se dice entonces que el nodo se vuelve puro. Este índice mide la impureza en la clasificación; a media que clasificación es correcta, el índice de Gini va tomando valores cercanos a 0.

Como criterio de mejora en una clasificación, debido a la división de los datos en dos grupos, se utiliza lo siguiente:

$$\Phi(s,t) = g(t) - p_{iz} * g(t_{iz}) - p_{de} * g(t_{de})$$

donde,

$g(t)$: Es el valor del índice de Gini en el nodo t.

p_{iz}: es la proporción de casos enviados al nodo izquierdo.

p_{de}: es la proporción de los casos enviados al nodo derecho.

$g(t_{iz})$: es el valor del índice de Gini en el nodo izquierdo

$g(t_{de})$: es el valor del índice de Gini en el nodo derecho

s: es la división propuesta.

Mientras más altos sean los valores de esta función (o más cercanas a g(t)) mejor será la clasificación.

Índice Binario

El índice binario, al igual que el índice de Gini, se basa en encontrar la división s que maximice la función de clasificación, esta función se define de la siguiente manera:

$$\phi(s,t) = p_{de} * p_{iz} \left[\sum_{i=1}^{k} |p(i/t_{iz}) - p(i/t_{de})| \right]$$

donde,

p_{iz}: es la proporción de casos enviados al nodo izquierdo.

p_{de}: es la proporción de los casos enviados al nodo derecho.

$p(i/t_{iz})$: es la probabilidad de que la categoría i de la clase criterio pertenezca al nodo izquierdo.

$p(i/t_{de})$: es la probabilidad de que la categoría i de la clase criterio pertenezca al nodo derecho.

s: es la división propuesta.

Entropía y Ganancia de Información

Esta sección se ha adaptado del libro Machine Learning de Tom Mitchell

Entropía: La entropía se relaciona con el desorden o caos, a mas entropía más desorden. En el contexto de los árboles de clasificación, la entropía es una medida de incertidumbre usada para determinar que atributo mejora la clasificación ya que un atributo o variable que mejor discrimina reduce la entropía. Este índice varía entre 0 y 1.

Sea S una variable cualitativa, su entropía se calcula con la siguiente fórmula:

$$Entropia(S) = \sum_{i=1}^{n} -p_i \, log_2 p_i$$

Donde :

S = Variable cualitativa con n categorías.

p_i = Probabilidad de los posibles valores en el nodo.

Ganancia de Información

La ganancia de información es una medida de discriminación que permite continuar con el proceso de clasificación al discriminar el atributo seleccionado de entre los atributos aún no clasificados.

Sea S una conjunto, donde A corresponde a los atributos de los objetos y V(A) es el conjunto de valores que A puede tomar.

$$Ganancia\ de\ informacion(S, A) = Entropia(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

ANEXO D.
MODELOS ACUMULADOS PARA DATOS ORDINALES

Esta sección es una adaptación de Econometric Analysis de Greene, 2007.

Cuando las respuestas de Y son ordinales se puede utilizar los denominados modelos logit acumulados, los cuales consisten en lo siguiente:

La probabilidad de una variable Y es la probabilidad de que Y sea menor o igual que un determinado valor k. Así, para una categoría dada k se define la *probabilidad acumulada* como

$$P(Y \leq k) = \pi_1 + \cdots + \pi_k, \;\; con\, k = 1,2, \ldots, K$$

Las probabilidades acumuladas reflejan el orden entre las categorías:

$$P(Y \leq 1) \leq \cdots \leq P(Y \leq K) = 1$$

Los modelos logits de las probabilidades acumuladas son:

$$\log\big(P(Y \leq k)\big) = \log\left[\frac{P(Y \leq k)}{1 - P(Y \leq k)}\right] = \log\left[\frac{\sum_{i=1}^{k} \pi_i}{\sum_{i=k+1}^{K} \pi_i}\right], \quad con\, k = 1, \ldots, K-1$$

Cada variable independiente tiene un solo coeficientes que no depende del valor k. La dependencia con respecto al valor k se verifica solamente con el coeficiente α_k. Considerando que si $k' < k$, entonces:

$$\log\big(P(Y \leq k'/x)\big) \leq \log\big(P(Y \leq k/x)\big)$$

ahora, se debe verificar que $\alpha_j \leq \alpha_{j'}$; es decir, que los valores de α_j aumentan con los valores de j.

Si se toman dos variables independientes X_1 y X_2, entonces:

$$\log\big(P(Y \leq k/X_1)\big) - \log\big(P(Y \leq k/X_2)\big) = \sum_{i=1}^{n} \beta_i (x_{1i} - x_{2i})$$

Por otro lado,

$$\log\big(P(Y \leq k/X_1)\big) - \log\big(P(Y \leq k/X_2)\big) = \log\left[\frac{\frac{P(Y \leq k|X1)}{P(Y > k|X1)}}{\frac{P(Y \leq k|X2)}{P(Y > k|X2)}}\right] = \sum_{i=1}^{n} \beta_i (x_{1i} - x_{2i})$$

La parte intermedia de la ecuación se denomina cociente de *odds* acumulado.

De este modo, los *odds* para un valor de respuesta menor o igual que k para $X = X_1$ es igual a:

$$\exp\left(\sum_{i=1}^{n} \beta_i (x_{1i} - x_{2i})\right) \; veces \; los \; odds \; para \; X = X_2$$

esta proporcionalidad no depende del valor de k. Es por esto que este modelo se denomina, también, modelo de *odds proporcionales.*

Si ambos individuos coinciden en todos los valores salvo en una componente, y si se supone además que difieren en una unidad en la $i - ésima$ componentente; se tiene, bajo estas hipótesis, la siguiente ecuación:

$$\frac{\frac{P(Y \leq k|X1)}{P(Y > k|X1)}}{\frac{P(Y \leq k|X2)}{P(Y > k|X2)}} = \exp\big(\beta_i (x_{1i} - x_{2i})\big) = e^{\beta_i}$$

De esta manera, si se mantienen las demás variables constantes, el cambio de la función de distribución de Y condicionada a X_1 y a X_2 es igual a e^{β_i}.

ANEXO E.

TEOREMA KHUN – TUCKER Y NÚCLEOS

Esta sección ha sido tomada y adaptada del trabajo hecho por Betancourt (2005), publicado en la Scientia et Technica Año XI, No 27, Abril 2005. UTP. ISSN 0122-1701.

Teorema de Khun – Tucker. Sea $\bar{\alpha}_\iota$ una solución del problema dual de maximización lagrangiana; entonces, satisface:

$$\bar{\alpha}_\iota\left[y_i\left(\bar{p}.z_i+\bar{b}\right)-1+\bar{\varepsilon}_\iota\right]=0,\quad i=1,\dots,\ell$$

$$(C-\bar{\alpha}_\iota)\varepsilon_i=0,\quad i=1,\dots,\ell$$

De estas igualdades se tiene que los únicos valores para los cuales $\bar{\alpha}_\iota \neq 0$ son los que para las constantes en las inecuaciones se satisfacen con el signo de igualdad. El punto x_i correspondiente a $\bar{\alpha}_\iota > 0$ se denomina *vector de soporte*.

Se pueden distinguir dos tipos de vectores de soporte:

Si $0 < \bar{\alpha}_\iota < C$, entonces x_i satisface $y_i\left(\bar{p}.z_i+\bar{b}\right)=1\ y\ \varepsilon_i=0$.

$\bar{\alpha}_\iota = C, \varepsilon \neq 0$ y x_i no satisface:

$$\begin{cases} p.z+b\geq 1, & y_i=1 \\ p.z+b\leq -1, & y_i=-1 \end{cases} \qquad i=1,\dots,\ell$$

En este caso, estos vectores de soporte se definen como errores y se alejan del margen de decisión.

Para desarrollar el hiperplano óptimo se utiliza la siguiente ecuación:

$$\bar{p}=\sum_{i=i}^{\ell}\bar{\alpha}_\iota y_i z_i$$

entonces, b se determina a partir del teorema de Kuhn – Tucker.

La función de decisión general (a partir del desarrollo precedente) se puede expresar como:

$$f(x)=sign(p.z+b)=sign\left(\sum_{i=i}^{\ell}\alpha_i y_i z_i.z+b\right)$$

Ejemplos de núcleos (Kernel)

- **Kernel estándar gaussiano (radio basal)**

$$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$$

- **Kernel polinomial de grado d**

$$K(x_1, x_2) = (\langle x_1, x_2 \rangle + 1)^d$$

- **Red neuronal**

$$K(x_1, X_2) = \tanh(k_1 \langle x_1, x_2 \rangle + k_2)$$

Si se toma como referencia la función de kernel polinomial de grado d para construir un clasificador SVM, se tendría que resolver el siguiente problema de maximización:

$$Max \;\; P(\alpha) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$s.a \sum_{i=1}^{\ell} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, \ell$$

y tiene una función de decisión de la siguiente forma:

$$f(x) = sign(p.z + b) = sign\left(\sum_{i=1}^{\ell} \alpha_i y_i K(x_i, x_j) + b\right)$$

ANEXO F.
CÓDIGOS PARA GENERACIÓN DE MODELOS APLICADOS

F.1 CÓDIGO PARA GENERACIÓN DEL ACPC EN SPSS

```
CATPCA VARIABLES=COLE_FISCAL madre_trabaja  NroMiem
     total_ingreso_familiar_menos_rebajas total_propiedades
    valor_total_propiedades
    /ANALYSIS=COLE_FISCAL(WEIGHT=1,LEVEL=NOMI)
madre_trabaja(WEIGHT=1,LEVEL=NOMI)
   NroMiem(WEIGHT=1,LEVEL=NUME)
    total_ingreso_familiar_menos_rebajas(WEIGHT=1,LEVEL=NUME)
    total_propiedades(WEIGHT=1,LEVEL=NUME)
valor_total_propiedades(WEIGHT=1,LEVEL=NUME)
  /DISCRETIZATION=COLE_FISCAL(GROUPING,NCAT=7,DISTR=NORMAL)
    madre_trabaja(GROUPING,NCAT=7,DISTR=NORMAL)
    NroMiem(GROUPING,NCAT=7,DISTR=NORMAL)
    total_ingreso_familiar_menos_rebajas(GROUPING,NCAT=7,DISTR=NORMAL)
   total_propiedades(GROUPING,NCAT=7,DISTR=NORMAL)
    valor_total_propiedades(GROUPING,NCAT=7,DISTR=NORMAL)
  /MISSING=COLE_FISCAL(PASSIVE,MODEIMPU) madre_trabaja(PASSIVE,MODEIMPU)
    total_ingreso_familiar_menos_rebajas(PASSIVE,MODEIMPU)
total_propiedades(PASSIVE,MODEIMPU)
valor_total_propiedades(PASSIVE,MODEIMPU)
  /DIMENSION=3
  /NORMALIZATION=VPRINCIPAL
  /MAXITER=100
  /CRITITER=.00001
  /PRINT=CORR LOADING
  /PLOT=BIPLOT(LOADING) (20) OBJECT (20) TRIPLOT(20) CATEGORY(NroMiem
    total_ingreso_familiar_menos_rebajas total_propiedades
valor_total_propiedades
    COLE_FISCAL
    madre_trabaja )
    (20) LOADING(20)
  /SAVE=OBJECT.
```

F.2. CÓDIGO PARA GENERACIÓN DEL K – MEDIAS EN SPSS

```
QUICK CLUSTER ISE_normal total_propiedades
total_ingreso_familiar_menos_rebajas
```

```
  /MISSING=LISTWISE
  /CRITERIA=CLUSTER(8) MXITER(10)
CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /SAVE CLUSTER
  /PRINT  ANOVA.
```

F.3. CÓDIGO PARA GENERACIÓN DEL ÁRBOL DE CLASIFICACIÓN EN SPSS

```
TREE Cluster_K_5 [n] BY total_ingreso_familiar_menos_rebajas [s]
familia_tiene_propiedad_o_vehiculo[n]
 valor_total_propiedades [s] cole_fiscal[n] madre_trabaja[n]
Pichincha[n]
  /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES
NODEDEFS=YES SCALE=AUTO
  /DEPCATEGORIES USEVALUES=[VALID]
  /PRINT MODELSUMMARY CLASSIFICATION RISK
  /RULES NODES=TERMINAL SYNTAX=INTERNAL TYPE=SCORING
  /SAVE PREDVAL
  /METHOD TYPE=EXHAUSTIVECHAID
  /GROWTHLIMIT MAXDEPTH=AUTO MINPARENTSIZE=5 MINCHILDSIZE=1
  /VALIDATION TYPE=NONE OUTPUT=BOTHSAMPLES
  /CHAID ALPHASPLIT=0.05 ALPHAMERGE=0.05 SPLITMERGED=NO
CHISQUARE=PEARSON CONVERGE=0.001
    MAXITERATIONS=1000 ADJUST=BONFERRONI INTERVALS=63
  /COSTS EQUAL
  /MISSING NOMINALMISSING=MISSING.
```

F.4. CÓDIGO PARA GENERACIÓN DE REGRESIÓN LOGÍSTICA MULTINOMIAL EN SPSS

```
NOMREG Cluster_K_5 (BASE=FIRST ORDER=ASCENDING) BY madre_trabaja
COLE_FISCAL total_propiedades WITH
    total_ingreso_familiar_menos_rebajas
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20)
LCONVERGE(0) PCONVERGE(0.000001)
    SINGULAR(0.00000001)
  /MODEL
```

```
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
  /INTERCEPT=INCLUDE
  /PRINT=CLASSTABLE PARAMETER SUMMARY CPS MFI.
```

F.5. CÓDIGO PARA GENERACIÓN DEL SVM EN R

```
library(e1071)
library(rpart)
library(readxl)
library(dplyr)

#traer datos y definir grupos
datos <- read_excel("Datos_2_Regresion_ACP_6500_regs_para_SVM.xlsx", sheet =
"Hoja2")
datos$Cluster_K_5 <- factor(datos$Cluster_K_5,levels = c(1,2,3,4,5),labels =
c("A","B","C+","C-","D"))

#Separar muestra para entrenamiento y validacion
trainset <-subset(datos, datos_para_muestra==1)
testset <-subset(datos, datos_para_muestra==0)

#Entrenar modelo
svm.model <- svm(Cluster_K_5 ~ ., data = trainset[,c(2,3,7:11)])

#Evaluar modelo
svm.pred <- predict(svm.model, testset[,c(2,3,7:11)],decision.values = TRUE)
t1<-table(pred = svm.pred, true = testset$Cluster_K_5)
classAgreement(t1)
t1

#Ver modelo y caracteristicas
print(svm.model)
summary(svm.model)


#ver los vectores de soporte
svs<-svm.model$SV

#atributos modelo
attributes(svm.model)

#valores ajustados
svm.model$fitted
```

```
#valores para decidir
svm.model$decision.values
```

ANEXO G.

COMPARACIÓN DE QUINTILES : INGRESO PER CAPITA Y EQUIVALENTE

En la tesis de la Carrera de Ciencias Económicas y Financieras de la EPN, EL GASTO EQUIVALENTE DE LOS HOGARES ECUATORIANOS EN FUNCIÓN DE SU CONSUMO ALIMENTARIO, COMPOSICIÓN Y TAMAÑO, SEGÚN LA ENCUESTA DE CONDICIONES DE VIDA 2006 realizada por Hamilton Erazo (2013), se describe que el gasto de una familia tiene una relación directa con el número de miembros de la misma, sin embargo a medida que aumenta el tamaño de la misma el gasto no crece de la misma manera, así se tiene la siguiente tabla que indica en un hogar el número real de miembros versus el número equivalente de miembros, por ejemplo un hogar con cuatro miembros en gasto equivale a un hogar con 3,18 miembros:

Tabla 22. Miembros de un hogar y su equivalencia en gasto

Número de miembros	
Real	Equivalente
NroMiem	NroMiemEq
1	1,00
2	2,00
3	2,64
4	3,18
5	3,62
6	4,00
7	4,32
8	4,58
9	4,78
10	4,92
11	5,00

Fuente: Erazo, 2013

Elaborado por: El Autor

Considerando esto se ha construido el Índice Per Cápita Familiar Equivalente (IPFEq), el cual resulta de dividir el Ingreso entre el número de miembros equivalente, así tenemos la siguiente distribución:

Tabla 23. Quintiles usando los miembros equivalentes:

Quintil Eq	N
1	1.313
2	1.305
3	1.322
4	1.247
5	1.379

Elaborado por: El Autor

Donde los estudiantes del quintil 1 tienen menor IPFEq que los del quintil 2, y así sucesivamente.

Tabla 24. Quintil EPN vs Quintil Equivalente

		QUINTIL EPN						TOTAL
		5	4	3	2	1	S/A*	
Quintil Equivalente	5	608	336	168	149	117	1	**1.379**
	4	485	556	132	49	21	4	**1.247**
	3	66	232	517	408	93	6	**1.322**
	2	55	99	197	481	469	4	**1.305**
	1	43	87	138	259	779	7	**1.313**
TOTAL		**1.257**	**1.310**	**1.152**	**1.346**	**1.479**	**22**	**6.566**

Elaborado por: El Autor

- Registros sin clasificación.

Se aprecia que la diagonal principal coincide con el 44,79%, y la triple diagonal coincide con el 83,14%, por lo que podemos decir que los dos quintiles (EPN y Equivalente) expresan la misma clasificación.

Tabla 25. INS vs Quintil Equivalente

		QUINTIL EQUIVALENTE					TOTAL
		5	4	3	2	1	
INS	A	70	10	1		1	**82**
	B	156	27	5	2	7	**197**
	C+	278	317	39	15	5	**654**
	C-	418	677	312	117	61	**1.585**
	D	457	216	965	1.171	1.239	**4.048**
TOTAL		**1.379**	**1.247**	**1.322**	**1.305**	**1.313**	**6.566**

Elaborado por: El Autor

La coincidencia entre el INS y el Quintil Equivalente es poca, en la diagonal principal coincide el 22,72% y en la triple diagonal el 53,9 %.

Tabla 26. Becas por Quintil Equivalente

QUINTIL EQUIVALENTE	SIN BECA	CON BECA	TOTAL	% BECAS
5	1.344	35	1.379	2,54%
4	1.220	27	1.247	2,17%
3	1.213	109	1.322	8,25%
2	1.091	214	1.305	16,40%
1	947	366	1.313	27,88%
	5.815	751	6.566	

Elaborado por: El Autor

Se aprecia que el Quintil Equivalente asigna becas en todos los grupos, incluso en aquellos de mayor IPFEq, y en los grupos de menor IPFEq apenas el 44,27%.

Se puede concluir que las clasificaciones estudiantiles por Quintil Equivalente y por Quintil EPN producen resultados similares, por tanto son una clasificación similar.