POSTGRADO
ECONOMÍA Y NEGOCIOS
UNIVERSIDAD DE CHILE

"Determinantes del Nivel de Conocimiento Financiero de los Individuos en las Administradoras de Fondos de Pensiones"

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN FINANZAS

Alumno: María Ignacia Muñoz Carmona

Profesor Guía: David Díaz Solís Ph. D.

Santiago, Abril 2015

Contenido

Resumen

El objetivo de este trabajo es la descripción del nivel de conocimiento financiero del sistema de Administradoras de Fondos de Pensiones (AFP) utilizando variables demográficas y socioeconómicas: género, edad, nivel educacional, estado civil e ingreso promedio mensual; además de algunas variables específicas de la Encuesta de Protección Social (EPS) que se utilizaron para medir si los afiliados a las AFP estaban informados por Berstein y Ruiz (2005).

En este trabajo, se usa la metodología CRISP-DM, donde en un comienzo se analiza el problema, para luego comprender los datos y realizar una selección de las variables. Para la modelación, se ocupan árboles de decisión CHAID, dada la fácil comprensión de los patrones descritos con este modelo y su capacidad de generar árboles no binomiales que permiten generar árboles más pequeños y simples de entender. Finalmente, en la etapa de evaluación se analiza la capacidad de predicción, el rendimiento y los niveles de confianza de los árboles anteriormente mencionados.

Los resultados de la investigación muestran que se generaron dos árboles de decisión con un alto nivel de predicción (alrededor del 90%), de 5 niveles, por lo que resulta ser de fácil comprensión, y puede ser utilizado como base para medir el nivel de conocimiento financiero de las personas. En la matriz de coincidencias se observa el nivel de predicción, ya que en la diagonal se concentra el número de observaciones predichas y además, los niveles extremos están en cero. En cuanto al rendimiento, se tienen valores altos, lo que implica que el modelo no es aleatorio, sino que el algoritmo funciona como modelo de predicción. Por último, el modelo posee un 90% de confianza. Es destacable, que un modelo simple, fácil de comprender, tenga resultados como tales.

Teniendo una predicción acerca del nivel de conocimiento financiero del sistema de AFP de los individuos, se pueden tomar decisiones de políticas públicas como por ejemplo enseñar en los colegios finanzas básicas, para que desde joven se tenga noción acerca de todas las instituciones en las que estamos inmersos. Lo ideal es expandir esto hacia otras áreas, como salud, inversiones, entre otros.

1. Introducción

En Chile, a comienzo de la década de los ’80 el sistema de pensiones de reparto se encontraba colapsado (Superintendencia de AFP, 2002). Primero, tenía un alto costo tanto para los trabajadores como para los empleadores. Segundo, tenía una administración ineficiente, existía una gran cantidad de cajas de previsión, ya que cada sector económico solía tener la suya (Ver Anexo 1). Tercero, era un sistema discriminatorio, donde los requisitos para pensionarse no eran uniformes. Finalmente, y lo más relevante, es que los recursos se estaban volviendo insuficientes, ya que cada vez menos trabajadores costeaban a uno retirado (Superintendencia de AFP, 2002). Por todo esto, se cambió al sistema de capitalización individual, donde cada individuo ahorra durante su vida activa para financiar su vida pasiva, y estos fondos son administrados por una entidad privada, llamadas Administradoras de Fondos de Pensiones (AFP). Cabe mencionar, que se hizo obligatorio para las personas que comenzaron a trabajar desde 1983 (Superintendencia de AFP, 2010).

Las administradoras de fondos de pensiones fueron diseñadas para solucionar el problema de la no sustentabilidad ya que la población estaba envejeciendo. De esta forma, se esperaba que los individuos tuvieran un nivel de vida similar al de sus últimos años de trabajo. Las AFP tenían sustento en: garantía estatal a las pensiones, presunto bajo conocimiento e interés, y la regulación de las inversiones. Cabe mencionar que en ese momento existía un mercado de capitales poco desarrollado (Berstein y Ruiz, 2005).

Con esta reforma, comenzaron a convivir una gran cantidad de administradoras, pero poseían una alta regulación que disminuía la posibilidad de que compitieran más libremente. Dentro de éstas están las inversiones que las AFP realizan, la estructura de los precios y tarifas cobradas a los afiliados, la

información mínima a ser reportada y los servicios que se deben proveer. Esto produjo que las administradoras se enfocaran en contratar vendedores, lo que aumentó sus gastos y llevó al mercado de las AFP a una nueva organización industrial con menos actores. Además, existen economías de escala en la producción del servicio, lo que también contribuyó a la concentración del mercado.

En la actualidad, existe un problema en la manera que las AFPs compiten por nuevos afiliados. Dado que es más fácil identificar a las personas que ya se encuentran trabajando y por ende ya están cotizando, las AFPs focalizan sus esfuerzos en lograr que los afiliados se traspasen entre administradoras, dedicando bajos o nulos esfuerzos a buscar a personas que están por comenzar su vida laboral. En este proceso, se utilizan tanto las habilidades del agente como los premios que se les puedan ofrecer a los individuos, y se desvirtúa el traspaso (Berstein y Ruiz, 2005).

Otro problema central del actual sistema de AFPs es la obligatoriedad y complejidad que tiene. Dado esto, es común que personas menos informadas basen su decisión de elección de AFPs en criterios no técnicos. Esto produce que los cotizantes al recibir información de las administradoras le den baja importancia, al igual que la elección de éstas y se dejen llevar más por lo premios y atención recibida. El óptimo es que las personas al momento de escoger su AFP tengan la información necesaria de todas las opciones que les entrega el mercado, acerca de sus costos y rentabilidad, para compararlas entre sí. Pudiendo de esta forma realizar la elección que más se ajuste a sus necesidades y que más confianza les dé. Hoy en día, lo que ocurre es que los cotizantes manejan un bajo nivel de información y además un bajo nivel de involucramiento. No comprenden el sistema de pensiones, tampoco saben de qué forma deberían participar y no les interesa entender.

Este trabajo intenta explicar el nivel de conocimiento financiero (NCF) de los afiliados al sistema de pensiones AFP a través de variables demográficas y socioeconómicas (educación e ingreso), llamadas variables de caracterización, y de las generadas por la Encuesta de Protección Social (EPS), que corresponden a respuestas correctas en preguntas relevantes acerca del sistema previsional. Con esto se pretende contribuir al entendimiento de las características de los afiliados que explicarían su interés y nivel de conocimiento del sistema, información que puede ser de gran utilidad tanto para el mercado, como para la elaboración y seguimiento de políticas públicas.

En este sentido, en este trabajo se utilizará la metodología CRISP-DM para obtener un árbol de decisión que ayude a describir según las variables escogidas qué nivel de conocimiento posee una persona. Pudiendo ser utilizado posteriormente para analizar la alfabetización financiera que posee un individuo, y/o segmentos definidos de los mismos.

Los principales resultados obtenidos, gracias a la segmentación de la población en cuatro clústers, son que realmente existe un bajo nivel de conocimiento financiero, sobre todo en las personas de mayor edad. Lo que conlleva un riesgo bastante alto, ya que son personas con bajo nivel educacional, por lo que buscar una forma de enseñarles es bastante complejo. El grupo que presentó mayor NCF fue el adulto joven, con un promedio de edad de 40 años, con un nivel educacional alto. Una de las posibles razones es que están en la mitad de su vida, donde comienzan a pensar en el futuro y cómo lo van a sobrellevar, porque ya ven que sus padres están mayores y/o tienen hijos pequeños, entonces vislumbran lo que les depara el futuro por lo que deben comenzar a preocuparse tanto de la salud como del ahorro.

En cuanto al modelo descriptivo, se generaron cuatro árboles de decisión, en las próximas secciones se encuentra detallado el modelo y las razones de por qué se utilizó. Primero, se realizó el árbol con las variables de caracterización y

las vistas en la literatura de Berstein y Ruiz (2005), detalladas en la Sección III, para analizar qué tan buen modelo generaba y si el algoritmo utilizaba todas las variables. Como era de esperar, las seis preguntas de la literatura fueron ocupadas y el género, el ingreso promedio mensual y el estado civil como variable de caracterización, con un nivel de predicción del 90%. Y en segundo lugar, se generó el árbol con todas las variables relevantes de la EPS, es decir, las que aportaban información al modelo. Esto con el fin de averiguar si existían otras variables que ayuden a complementar las de literatura para analizar el conocimiento financiero de las personas. Como resultado, se obtuvieron nuevas variables, descritas en la Sección V, igualmente las escogidas por Berstein y Ruiz (2005) fueron utilizadas por el algoritmo, pero ninguna de caracterización. También con un nivel de predicción del 90%.

A continuación se realiza un breve análisis del contexto actual del mercado previsional en Chile, junto con una revisión de antecedentes acerca de la cultura financiera a nivel global, además de una breve descripción del uso de la minería de datos en distintos estudios. En la Sección III se detallan las variables y datos a utilizar, la construcción del Índice de Conocimiento Financiero, y estadística descriptiva de la muestra respecto a este. En la Sección IV se explicará la metodología a utilizar, basada CRISP-DM, para continuar en la Sección V exponiendo los resultados de los clústers formados y de los árboles de decisión. Finalmente se expondrán las conclusiones, junto con las principales implicancias que se derivan del estudio.

2. Revisión de literatura

2.1 Sistema de Pensiones

En nuestro país las cotizaciones tanto en fondos de pensiones como en el sistema previsional son obligatorias. Para el sistema de las AFP el monto corresponde al 10% del ingreso imponible, donde éste será invertido en un fondo dentro de cinco distintos dependiendo del riesgo que el individuo quiera asumir. Éstos van desde el fondo A al fondo E, donde el primero corresponde al más riesgoso y el último al que asume menos riesgo.

En 1980 cuando se implementó el sistema de capitalización individual existieron varios cambios dentro del mercado laboral, el mercado de capitales y el mercado del ahorro, por lo que es bastante difícil estimar el impacto de las AFP de manera aislada. En este contexto, un estudio de Corbo y Schmidt-Hebbel (2003), estima que la contribución de esta reforma al PIB del país entre los años 1981 y 2001, fue de un 0,5%, mediante el ahorro, inversión, los mercados de capitales y laborales.

Dentro del sistema se observa una baja sensibilidad a la demanda de AFP. Berstein y Micco (2002), plantean un marco teórico para el análisis del rol de los agentes de venta en la industria de las AFP, dado los regalos y ofertas que se les puede dar a los individuos para el traspaso entre administradoras. Berstein y Ruiz (2005) obtuvieron resultados consistentes con los obtenidos anteriormente de acuerdo a la sensibilidad de la demanda, y enfatizan la desinformación y bajo interés de los cotizantes a entender el sistema de pensiones. Marinovic y Valdes (2005), Cerda (2006), Berstein y Cabrita (2007), también obtuvieron resultados acordes. Lo que implica que los individuos no eligen su administradora basándose en la rentabilidad o en las comisiones cobradas, lo que conlleva un problema de competitividad y de concentración del mercado.

Otro problema observado de este sistema de pensiones, es la cobertura (Arrau y Valdés (2002), Valdés (2005), Holzman et al. (2005), Mitchell, Todd y Bravo (2007), Sistema de Pensiones, SAFP (2010)). En el año 2006, se observó una baja densidad de cotizaciones con respecto a lo esperado, sólo el 66% de los trabajadores cotizaban. Las razones fueron el trabajo independiente y la informalidad del mercado laboral, que dejaban una amplia ventana de tiempo a las personas sin cotizar, y el problema que se observó principalmente en las mujeres es que un porcentaje significativo de su vida activa se encontraban fuera de la fuerza de trabajo. Cabe agregar, que se observa que el sistema privado sólo atiende a una porción de la población, provocando que el estado complete el universo de individuos a través de las pensiones asistenciales (OIT (2006), Guardia et al. (2007), Bosch (2008)). Por todo lo anterior, el problema también pasa por el sistema de AFP en sí, la accesibilidad y cobertura es un tema pendiente aún. Los individuos con bajo nivel de conocimiento se ven afectados más aún, por ejemplo muchos no saben que si no cotizan mensualmente su jubilación disminuirá bastante en comparación a sí hacerlo.

Los resultados del mercado financiero chileno parecen replicarse a escala global. En el último tiempo, han existido varios autores (Skog (2006), Van Rooj, Lusardi y Alessie (2007), Mitchell, Todd y Bravo (2007), Lusardi (2008), Lusardi, Mitchell y Curto(2009), Lusardi, Keller y Keller (2009), Lusardi y Tufano (2009), Mitchell (2010), Fajnzylber y Reyes (2011)) que se han centrado en el tema del alfabetismo financiero a nivel mundial, llegando a la conclusión de que en general existe un nivel muy bajo de conocimiento financiero en todos los países. Por ejemplo, en Numeracy, Financial Literacy and Financial Decision Making, Lusardi (2012), se realiza un estudio acerca del conocimiento financiero en países que se espera que tengan un alto nivel, entre ellos Estados Unidos, Australia, y países de Europa, donde se obtuvo un resultado totalmente contrario. Lo que se observó fue que la alfabetización financiera se concentra principalmente en personas de género masculino, con un nivel alto de

educación y riqueza. Y otros estudios como Lusardi y Mitchell (2007), Lusardi y Tufano (2009), Dvorak y Hanley (2010) han obtenido resultados similares, cabe mencionar que el primero concluye lo mismo, pero además obtuvo que las personas blancas y las que están casadas poseen mayor nivel de conocimiento financiero.

Por el lado de nuestro país, existen los estudios de Skog (2006), Mitchell, Todd y Bravo (2007), y Fajnzylber y Reyes (2011). El primero, establece cuatro categorías para medir el conocimiento financiero, concluyendo que los cotizantes transversalmente manejan información escasa acerca de este sistema. Y Fajnzylber y Reyes (2011), concluye que el 40% de los individuos pertenecientes al sistema de pensiones no sabe cuál es el monto que tiene ahorrado.

Acerca de las características de cada individuo, Skog (2006) y Mitchell, Todd y Bravo (2007), presentan evidencia de que el género masculino, quienes poseen un mayor nivel de educación y quienes tienen ingresos más altos, muestran un mayor nivel de alfabetización financiera. Además, Skog (2008), concluye que trabajar con contrato laboral, pertenecer a grandes empresas o instituciones, afecta positivamente en el nivel de conocimiento financiero. También, Mitchell, Todd y Bravo (2007), plantean una relación positiva entre la experiencia en el sistema de pensiones y el nivel de conocimiento financiero.

Considerando ahora las implicancias que tiene el bajo alfabetismo financiero en los individuos, una arista que se genera es la subestimación de la esperanza de vida lo que a su vez trae como consecuencia una subestimación del riesgo de quedarse sin dinero en la vejez, Mitchell (2010). Según Lusardi y Mitchell (2007), el mayor conocimiento financiero aumenta la preocupación de los individuos por la planificación del retiro.

Por otra parte, Lusardi, Keller y Keller (2009) y Fajnzylber y Reyes (2011), exponen que un aumento en el nivel de conocimiento financiero cambia el

comportamiento de los individuos, impactando positivamente en el nivel del ahorro. Además, Lusardi (2012) muestra que aumenta el ahorro previsional voluntario (APV), así como también la inversión (Cole y Shastry (2008)) y la tenencia de acciones (Van Rooij, Lusardi y Alessie (2007)).

Dvorak y Hanley (2010) muestran que ni siquiera los cotizantes que poseen un conocimiento financiero básico, son capaces de entender y distinguir las distintas opciones de retiro ofrecidas por las administradoras. Por su lado, Mitchell, Mottola, Utkus y Yamaguchi (2009), analizaron los movimientos de los fondos de pensión en EE.UU, concluyendo que los movimientos son bastante bajos, no se realizan siquiera rebalanceos de portafolio antes shocks externos, lo que se le atribuye directamente a la deficiente información recibida por los participantes para la administración de sus fondos de pensión.

Cabe mencionar, que no sólo las personas se ven perjudicadas en el sistema de pensiones debido al bajo nivel de conocimiento financiero. Sino que también al momento de tomar créditos, ya que Courchane, Gailey y Zorn (2008) concluye que mientras exista un mayor nivel de conocimiento financiero, se evitará el sobre-endeudamiento y otros costos asociados a la deuda.

Dada la situación actual presentada, la solución que surge son los programas de educación financiera para la población, tanto en nuestro país como a nivel mundial. Se debe asegurar su eficacia, ya que si las personas lo encuentran complejo pueden surgir los mismos problemas que actualmente existen como el desinterés. Como plantea Cole y Shastry (2008), donde analiza el impacto de un año de educación en el nivel de conocimiento financiero, éste fue mayor que el programa de alfabetización financiera, por lo tanto, lo óptimo es que se comience desde la educación secundaria a enseñar cómo funciona el sistema financiero nacional actual.

2.2 Data Mining

Hoy en día, estamos en un planeta casi conectado completamente, donde existe una gran cantidad de datos que se van generando de forma continua. Dado esto, diversos campos de investigación quieren utilizarlos, desde la ciencia exacta hasta los negocios. Esto generó la necesidad de nuevas metodologías y herramientas, que permitan a los investigadores generar información a partir de estos datos y finalmente lograr el conocimiento.

En las finanzas esta tarea viene a ser especialmente compleja, dado que el conocimiento puede estar oculto en complejas relaciones no lineales que no son perceptibles a la simple observación o bien detectables con las herramientas comúnmente utilizadas para ello. Cabe mencionar, que sobre todo en el campo de los negocios, quienes están encargados de la toma de decisiones buscan cada vez más basar sus modelos en información más simplificada y de fácil interpretación.

Se entiende por Data Mining (DM), a la etapa de extracción de patrones de comportamiento no trivial de los datos, que forma parte del proceso global conocido como “Descubrimiento de Conocimiento en Base de Datos” (KDD: Knowledge Discovery in Databases) (Fayyad M.U, 1996; Weber R., 2000). El DM entrega diversas técnicas para encontrar patrones en grandes conjuntos de datos. Este enfoque multidisciplinario combina los resultados e intuiciones provenientes de varias ramas científicas, tales como la estadística, el aprendizaje de máquinas, tecnologías difusas y redes neuronales.

En la literatura existen muchas definiciones para caracterizar expresiones como KDD y DM, donde se realizan las siguientes aseveraciones: “El descubrimiento de conocimientos en base de datos es el proceso no trivial de identificar

patrones en datos que sean válidos, novedosos, potencialmente útiles y, por último, comprensibles", "Data Mining se refiere al acto de extraer patrones o modelos a partir de los datos" (Fayyad M.U. 1996).

Debido a la capacidad del DM y KDD de encontrar patrones en bases de datos y presentarlos de manera sencilla a los usuarios para que éstos puedan tomar decisiones informadas en base a ellos, esta tesis se centrará en su implementación. Esto como una herramienta alternativa a la econometría tradicional en la tarea de descubrir y explicar el nivel de conocimiento financiero de los afiliados al sistema previsional chileno de AFPs.

De acuerdo a la literatura, una representación frecuente de un proceso típico de KDD, contempla los siguientes pasos (Weber R., 2000):

1. **Desarrollar una comprensión del dominio de la aplicación:** Determinar, de manera global, los beneficios de un estudio de DM sobre los datos y las posibles alternativas de análisis, para delimitar el dominio del problema. Conocer el alcance, utilidad y significado de los datos.

2. **Crear un conjunto de datos objetivo:** Una vez, escogida la tarea global de DM sobre los datos, se creará un conjunto de datos objetivo.

3. **Limpieza y pre-procesamiento de datos:** Consiste en entender la estructura y formato de los datos de estudio, para poder determinar duplicidad de representación de datos, atributos que son relevantes para el estudio y comportamiento de estos.
    a. Se hacen en esta etapa estudios preliminares de los datos, pruebas de correlación, análisis de dependencia de variables, y

pruebas estadísticas que den una visión general del comportamiento de los datos.

4. **Reducción y transformación de los datos:** Se eliminan atributos irrelevantes para el estudio, campos nulos y atributos que se encuentren representados más de una vez.
    a. Finalizando esta etapa, se preparan los datos para la aplicación de los algoritmos de DM. Esta preparación puede incluir adaptación del formato de los datos, normalización, etc.

5. **Elegir la tarea de DM:** Se elige la tarea específica de DM para un análisis más detallado. Aquí se debe decidir si el propósito es, por ejemplo, la agrupación de objetos, la regresión o el modelaje de dependencia. Sobre la base de esta decisión, los más importantes algoritmos de DM deben ser seleccionados, (paso 6), los que se usan en la búsqueda real de patrones de datos (paso 7).

6. **Elegir los algoritmos de DM**.

7. **Data Mining:** Ejecución de los algoritmos de DM. Esto puede incluir un contraste de uno o varios algoritmos, evaluando su desempeño, resultados obtenidos, etc. Se utilizarán algoritmos de redes neuronales, agrupamiento difuso, clustering, redes de Kohonen, entre otros según sea el caso.

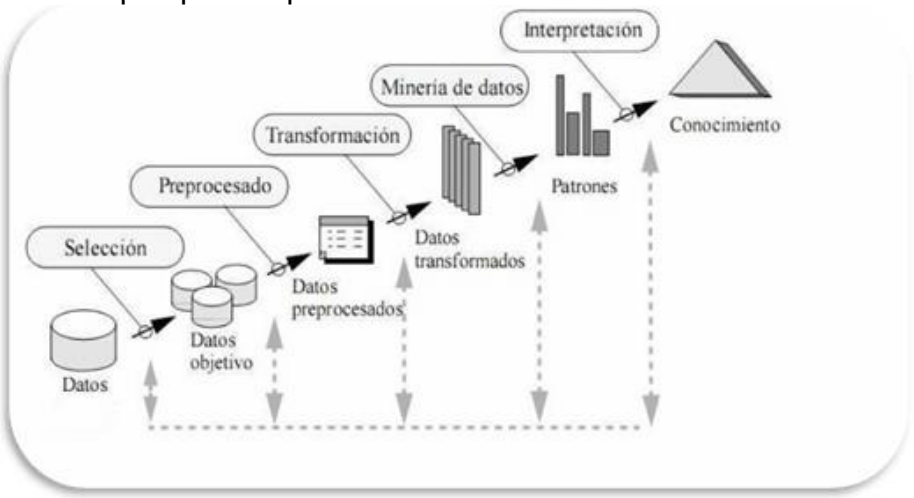
8. **Evaluar el resultado de DM:** Testear, en base al conocimiento experto de los datos de estudio de DM, la coherencia de los resultados arrojados por los algoritmos. Esta etapa corresponde a la validación del

conocimiento descubierto por el proceso, interpretación de las correlaciones obtenidas, o validación de resultados del algoritmo con conocimiento experto.

9. **Consolidar el conocimiento experto:** Sacar un aprovechamiento real de los resultados obtenidos en el estudio de DM, y su aprovechamiento en el contexto de la problemática de estudio.

A continuación, un esquema global del proceso KDD:

Figura 1
Etapas para el proceso de extracción de conocimiento



Fuente: Fayyad et al. (1996); Gómez Flechoso (1998).

Muchos investigadores se enfocan principalmente en el paso siete del esquema descrito, aunque en ocasiones una adecuada preparación de los datos puede aportar más que la misma elección del algoritmo, que comprende ajustar modelos o determinar patrones desde los datos disponibles. Los modelos ajustados juegan el rol de conocimiento inferido. Son dos los principales tipos de formalismos matemáticos usados en el proceso de ajuste de modelos: (1) estadísticas y (2) lógica.

Como es documentado en Fayyad, Piatetsky-Shapiro (1997), el enfoque estadístico permite efectos no determinísticos en el modelo, mientras que un modelo basado en lógica es puramente determinístico. El primer enfoque se utiliza usualmente dada la incertidumbre en los procesos de generación de datos.

La mayoría de los métodos de DM están basados en probados métodos de aprendizaje de máquinas, reconocimiento de patrones: clasificación, *clustering*, regresiones y más. Si bien, la gama de algoritmos descritos en la literatura puede ser abrumadora y desconcertante, debe enfatizarse que de los muchos métodos de DM de la literatura sólo subyacen algunas pocas técnicas fundamentales. La real representación que es usada por un modelo en particular, proviene generalmente de la composición de unos pocos y bien conocidos modelos o métodos como polinomios, *kernels, splines*, funciones de base etc. Estos algoritmos tienden a diferir en el criterio utilizado para medir la bondad de ajuste o el método de búsqueda para encontrar un buen ajuste.

La literatura relacionada señala que la predicción y la descripción corresponden, en la práctica, a los dos objetivos principales dentro de los métodos de DM, la predicción envuelve la utilización de algunas variables o campos de la base de datos para predecir valores desconocidos o futuros de otras variables de interés y la descripción se focaliza en encontrar patrones interpretables para describir los datos. La importancia relativa entre cada uno de estos conceptos puede variar significativamente dependiendo del área de investigación y de los tipos de datos a tratar.

Las tareas básicas dentro del DM son:

a. **Clasificación:** Consiste en la categorización de datos en grupos predefinidos o clases, también llamado como de aprendizaje supervisado, ya que las clases son determinadas después de examinar la data. Los algoritmos de clasificación, de acuerdo a Dunham (2003), requieren que las clases sean definidas de acuerdo a valores de atributos, éstas se describen por la observación de las características de los datos conocida su pertenencia a una determinada clase.

b. **Regresión:** Es usada para clasificar los datos en variables de predicción de valor real. Se asume que los datos objetivos son ajustables por medio de la utilización de algún tipo conocido de función y luego se determina la mejor función de este tipo que modela los datos disponibles. Finalmente se analizan los errores o residuos para determinar qué función ajusta mejor.

c. ***Clustering*:** Es similar a la clasificación, con la diferencia que los grupos no han sido predefinidos, si no que más bien definidos por los mismos datos. Es definido como un proceso no supervisado o de segmentación de los datos en grupos. El *clustering* es usualmente obtenido por medio de la determinación de la semejanza de la data sobre atributos predefinidos. Los datos más similares son agrupados en conglomerados o clústers. Debido a que éstos no han sido predefinidos, a menudo se requiere cierto conocimiento para interpretar el significado de los clústers creados.

d. **Reglas de Asociación:** Corresponde a la tarea de descubrir relaciones entre los datos. Un ejemplo común se refiere a las ventas del *retail* en que ciertos productos son comprados juntos. Los intérpretes de las reglas de asociación deben tener la precaución de que ellas no sean relaciones causales.

e. **Descubrimiento Secuencial:** Es utilizado para determinar patrones secuenciales en los datos. Son similares a las asociaciones en los datos o eventos que están relacionados, aunque la relación está basada en el factor tiempo. A diferencia del análisis de carros de *retail* mencionadas en el ítem anterior, el cual requiere que los ítems deben ser comprados al mismo tiempo, en el descubrimiento secuencial los productos son comprados sobre un periodo de tiempo en un determinado orden.

Con los antecedentes ya expuestos, es posible intuir que el DM y el KDD, corresponden a la evolución e integración de muchas disciplinas tales como: administración de base de datos, recuperación de información, reconocimiento de patrones, visualización de datos, estadística, algoritmos, computación de alto desempeño, aprendizaje de máquinas, inteligencia artificial, entre otros, agrupados bajo el objetivo general de conseguir extraer el mayor provecho de las bases de datos disponibles.

2.2.1 Aplicaciones en las finanzas

Los mercados financieros están constantemente generando un gran volumen de datos que puede resultar muy valioso en apoyar las decisiones financieras. La mayoría de la data financiera presenta características que la hacen particularmente compleja, tales como no linealidad, no estacionalidad y

quiebres estructurales, entre otros. Lo que la hace una tarea particularmente desafiante en el campo del DM.

En Zhan y Zhou (2004), se realiza un análisis donde se profundiza sobre aplicaciones potenciales de DM en problemáticas financieras, dentro de ellos destacan la predicción del mercado accionario, la administración de portafolios, predicción de quiebras, mercado cambiario y detección de fraudes.

La investigación relacionada ha demostrado que la predicción de retornos futuros se puede basar en la tasa de crecimiento de ciertos factores fundamentales, tales como: ingresos, ganancias por acción, inversiones de capital, deuda, participación de mercado, entre otras variables. Generalmente se han utilizado modelos de regresión para modelar los cambios, sin embargo, estos modelos tienen la limitante de predecir patrones lineales solamente.
Una de las técnicas de DM más utilizadas para enfrentar este problema corresponde a las redes neuronales: *back propagation neural networks* (BPNN), *recurrent neural networks* (RNN) *y probabilistic neural networks* (PNN). El supuesto básico de estas metodologías, es que inputs de series de tiempo similares deberían producir como output series de tiempo similares en la medida que se ignoren fluctuaciones de entrada. En comparación con los métodos de regresión, la literatura ha demostrado que las BPNN corresponden a una de las mejores predictores en anticipar los signos de los retornos accionarios. Los autores destacan que existen importantes temas relacionados con el diseño en la aplicación del enfoque de redes neuronales, para la predicción de los precios de acciones:

1. Determinar el largo óptimo de la serie de tiempo a ser considerada en el análisis de datos.
2. Seleccionar indicadores sensibles al tiempo como inputs de la red.
3. Decidir qué hacer con los rezagos de los datos.

4. La manera específica en que puede ser utilizada la misma serie de tiempo (precios, variaciones porcentuales, variaciones absolutas, promedios etc.).

Generalmente, el desempeño de las redes neuronales en los problemas de clasificación, como predicción de precios de acciones, es medido con indicadores de precisión de la predicción donde se considera el valor 1 cuando existe un acierto y 0 cuando no, la precisión consiste en la razón entre la suma de la cantidad de aciertos sobre el total de predicciones. Cabe mencionar, que algunos autores han señalado que la maximización de la precisión de la estimación podría no ser un objetivo adecuado para muchas tareas reales de clasificación. La métrica de precisión de la estimación asume conocida la clase de distribución y el mismo peso para errores tipo falso positivo y falso negativo. En aplicaciones financieras tales como la predicción del precio de acciones o la detección de fraudes en tarjetas de crédito, un error de clasificación podría tener costos asimétricos. Por ejemplo, el costo de predecir incorrectamente el precio de una acción u omitir un caso de fraude, podría ser mucho más caro que el costo de una falsa alarma, por lo tanto, algunas métricas alternativas que capturen este hecho han sido desarrolladas en la literatura. Otros métodos han sido aplicados a este tipo de problemas, tales como la regla de inducción, análisis estadístico, algoritmos genéticos, visualización de datos entre otros. Los autores comentan que, por ejemplo, el Sistema *Recon*, induce reglas de clasificación para modelar una determinada base de datos. Este sistema analiza una base de datos histórica y produce reglas que clasificarán las acciones actuales, como excepcionales o no excepcionales respecto a sus desempeños futuros. Cada regla tiene una determinada fortaleza y su predicción es ponderada por la cantidad de evidencia que apoya tal regla. El algoritmo de regla de inducción, parte distinguiendo todas las características numéricas, luego explora si existe espacio para posibles reglas, por lo que existe una

suerte de búsqueda por todo el espacio de reglas, que esencialmente consiste en un análisis de las posibles reglas estadísticas pertinentes de aplicar en el problema de clasificación. Se ha documentado que el sistema *Recon*, ha tenido un desempeño por sobre el *benchmark*, en términos de retornos totales aplicados a un periodo de cuatro años, por lo que las reglas de inducción pueden llegar a ser una herramienta valiosa en la selección de portafolios de acciones.

Modelos como el CAPM (*Capital Asset Pricing Model*) o APT (*Arbitrage Pricing Theory*), han sido útiles en explicar el precio de los activos en los mercados financieros, normalmente relacionando el desempeño, con el riesgo sistemático de los activos. A partir de estos modelos se han ido desarrollando modelos híbridos. Un ejemplo de esto, es la mezcla del APT con algoritmos genéticos para la selección de activos a ser parte del portafolio, o una red neuronal para la predicción de los retornos de los activos en el portafolio y un algoritmo genético para la determinación de los pesos óptimos de cada activo. Estas metodologías han sido complementadas por medio de Filtros de Kalman y redes neuronales multicapa con el fin de obtener un análisis de sensibilidad a las variables económicas. Otros modelos pertenecientes al DM que han sido empleados son los modelos markovianos, los cuales han reportado en investigaciones relacionadas que ésta herramienta puede producir un significativo mejor desempeño en relación con el *benchmark*. También existen otras herramientas como la visualización de datos han sido utilizadas para monitorear el cambio y tendencia de portafolios.

Un modelo clásico que marcó el progreso del desarrollo de la predicción de quiebras fue el modelo Z-Score de Altman (1968), el cual con sus 5 variables explicativas, por medio del uso de análisis discriminante múltiple, ha mostrado un importante poder predictivo. Además, Zhang y Zhou (2004), señalan que

adicional a esos esfuerzos, se han añadido otros métodos de DM, tales como regresiones logísticas, algoritmos genéricos, arboles de decisión clasificación y regresión (CART). Aunque las redes neuronales y modelos estadísticos han sido utilizados en la predicción de quiebras, ellos podrían encontrar el problema de frecuencias no uniformes respecto de los 2 estados de interés, lo que crea al menos dos obstáculos no menores, en la evaluación de la capacidad predictiva de una red. El primero de ellos se relaciona con el impacto de las frecuencias desiguales de los 2 estados quiebra/no quiebra, en el entrenamiento de una red neuronal o la estimación de los parámetros de un determinado modelo estadístico. Generando muestras aleatorias desde poblaciones desbalanceadas podría probablemente conducir a muestras que contiene una mayoría de eventos pertenecientes a un estado de interés. Consecuentemente, el desempeño de redes neuronales o modelos estadísticos podría resultar pobre si es probado en situaciones reales. Como una forma de abordar este problema, los investigadores han seleccionado un muestreo basado en elección, en el cual la probabilidad de una observación depende del valor de la variable dependiente. El segundo problema, corresponde a la evaluación de la precisión de varios modelos decisionales. El porcentaje de observaciones correctamente clasificadas puede ser potencialmente muy engañoso con muestras desbalanceadas. En general, el entrenamiento de una red neuronal con muestras balanceadas, en aplicaciones como predicción de quiebras, puede habilitar la red a familiarizarse por sí misma con el infrecuente estado de interés. Las redes neuronales entrenadas sobre muestras balanceadas proveen usualmente mejores resultados cuando son aplicadas bajo condiciones realistas.

El mercado cambiario corresponde a uno de los subyacentes más líquidos del mundo donde los operadores normalmente utilizan reglas técnicas sobre todo en horizontes de trading de corto plazo (en posiciones de mediano plazo o

estructurales se ponderan adicionalmente factores financieros y macroeconómicos, además de técnicos). El DM ha sido aplicado en la búsqueda de tales factores técnicos, en Walzack (2001) se utilizaron redes neuronales homogéneas, en la predicción del dólar contra distintas monedas del mundo entrenadas con datos de 21 años de historia y realización de predicciones de 1 día. En dicho trabajo, el autor documenta un 58% de precisión en la predicción de la libra esterlina y un 57% contra el marco alemán. El input para cada red utilizada consistió en uno o más rezagos de las paridades bajo estudio. Otro de los hallazgos reportados por este autor consiste en el hecho de que con menores tamaños de data, se obtienen mejores resultados, lo que se contradice con parte de la literatura previa. Otros autores han querido incorporar factores como titulares de noticias que pueden afectar el mercado cambiario, lo cual implica que no sólo se utilizan datos cuantitativos en las predicciones, sino que también las predicciones se basan en datos de texto describiendo el estado actual de los mercados financieros, aspectos políticos y noticias económicas en general (minería de textos). Estos textos no sólo contienen los efectos sino que también las posibles causas de los mismos.

La detección de fraudes con tarjetas de crédito tiene características especiales. Una de ellas, es el limitado intervalo de tiempo en el cual debe ser tomada la decisión de aceptación o rechazo. La segunda, corresponde al hecho de que los datos son altamente sesgados en el sentido que es mucho mayor la cantidad de transacciones legítimas que fraudulentas. La tercera característica particular de este problema radica en que, una gran cantidad de operaciones de tarjetas de crédito tienen que ser procesadas en un momento dado y muy pocas de ellas normalmente resultan ser fraudulentas. Una tarea mayor en la detección de fraudes es la construcción de algoritmos o modelos con la habilidad de reconocer variedades de patrones de fraude. La mayoría de los métodos de DM, a menudo descartan los *outliers*, sin embargo en la detección

de fraudes, los eventos de ocurrencias extrañas pueden ser más interesantes que los de ocurrencias regulares. Así el análisis de *Outlier* ha sido utilizado en la detección de patrones de fraude, los cuales son particularmente diferentes de las transacciones regulares de tarjetas de crédito. Los métodos utilizados normalmente para tratar este tipo de problemas son sistemas expertos, redes neuronales, métodos estadísticos, sistemas de clasificación mejorados, reglas de inducción de lógica difusa, CART, clasificadores de Bayes entre otros.

2.2.1.1 Árboles de decisión en Finanzas

En este trabajo se utilizará el Modelo de Árboles de Decisión porque permiten desarrollar sistemas de clasificación que predicen o clasifican observaciones futuras según un conjunto de reglas de decisión. Si se dispone de datos divididos en clases que son relevantes, se pueden usar los datos para generar reglas que puede usar para clasificar casos antiguos o recientes con la máxima precisión.

Este método posee varias ventajas:

a. El proceso de razonamiento detrás del modelo resulta claramente evidente cuando se examina el árbol. Esto contrasta con otras técnicas de modelado de “caja negra”, en las que la lógica interna puede resultar difícil de averiguar.

b. El proceso incluye automáticamente en su regla únicamente los atributos que realmente importan en la toma de decisiones. Los atributos que no contribuyan a la precisión del árbol se omiten. Esto puede proporcionar información de gran utilidad acerca de los datos y se puede usar para reducir los datos a campos relevantes antes de entrenar otra técnica de aprendizaje, como una red neuronal.

c. Los patrones encontrados usando el algoritmo de modelo de árbol de decisión se pueden convertir en una colección de reglas *if-then* (un conjunto de reglas), que en muchos casos muestra la información de forma más humanamente comprensible. La presentación del árbol de decisión resulta útil cuando se desea ver el modo en que los atributos de los datos pueden dividir o particionar la población en subconjuntos relevantes para el problema. La presentación del conjunto de reglas resulta de utilidad si se desea ver el modo en que determinados grupos de elementos se vinculan a una conclusión particular.

En este trabajo, para una mejor apreciación de los modelos, se mostrará como principal resultado el árbol de decisión como conjunto de reglas. En los Anexos se puede ver el árbol completo.

Algoritmos de generación de árboles de decisión

El programa utilizado para modelar, IBM SPSS Modeler 14.2, posee cuatro algoritmos disponibles para realizar un análisis de segmentación y clasificación. Todos estos algoritmos son básicamente similares: examinan todos los campos del conjunto de datos para detectar el que proporciona la mejor clasificación o predicción dividiendo los datos en subgrupos. El proceso se aplica de forma recursiva, dividiendo los subgrupos en unidades cada vez más pequeñas hasta completar el árbol (según defina determinados criterios de parada). Los campos objetivo y de entrada utilizados en la generación del árbol pueden ser continuos (rango numérico) o categóricos, dependiendo del algoritmo que se utilice. Si se usa un objetivo continuo, se genera un árbol de regresión; si se usa un objetivo categórico, se genera un árbol de clasificación. [35]

a. **C&RT:** El nodo de árbol de clasificación y regresión (C&R) genera un árbol de decisión que permite predecir o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera “puro” si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (sólo se crean dos subgrupos).

b. **CHAID:** El nodo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&RT y QUEST, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.

c. **QUEST:** El nodo QUEST proporciona un método de clasificación binario para generar árboles de decisión; está diseñado para reducir el tiempo de procesamiento necesario para realizar los análisis de C&RT y reducir la tendencia de los métodos de clasificación de árboles para favorecer a las entradas que permitan realizar más divisiones. Los campos de entrada pueden ser continuos (rango numérico), sin embargo el campo objetivo debe ser categórico. Todas las divisiones son binarias.

d. **C5.0:** El nodo C5.0 genera un árbol de decisión o un conjunto de reglas. El modelo divide la muestra basándose en el campo que ofrece la máxima ganancia de información en cada nivel. El campo objetivo debe ser categórico. Se permiten varias divisiones en más de dos subgrupos.

Para este estudio, se utilizó el método CHAID (*Chi-Squared Automatic Interaction Detection*), principalmente por que se ajusta mejor a los datos con los que contamos, es decir, este método admite todos los tipos de entradas y acepta tanto variables de frecuencia como ponderaciones de casos. Y lo más importante, es que puede generar árboles no binarios, lo que conlleva un modelamiento más ajustado y con una mejor predicción.

¿Cómo funciona CHAID? El método CHAID examina en primer lugar las tablas de tabulación cruzada entre los campos de entrada y los resultados para, a continuación, comprobar la significación mediante una comprobación de independencia de chi-cuadrado. Si varias de estas relaciones son estadísticamente importantes, CHAID seleccionará el campo de entrada de mayor relevancia (el *P-value* más pequeño). Si una entrada cuenta con más de dos categorías, se compararán estas categorías y se contraerán las que no presenten diferencias en los resultados. Para ello, se unirá el par de categorías que presenten menor diferencia, y así sucesivamente. Este proceso de fusión de categorías se detiene cuando todas las categorías restantes difieren entre sí en el nivel de comprobación especificado. En el caso de campos de entrada nominales, pueden fundirse todas las categorías. Sin embargo, en los conjuntos ordinales, únicamente podrán fundirse las categorías contiguas.

Usos generales del análisis basado en árboles de decisión

a. **Segmentación:** Identifica las personas que pueden ser miembros de una clase determinada.

b. **Estratificación:** Asigna los casos a una categoría de entre varias, por ejemplo, grupos de alto riesgo, bajo riesgo y riesgo intermedio.

c. **Predicción:** Crea reglas y las utiliza para predecir eventos futuros. Las predicciones también pueden significar intentos de relacionar atributos predictivos con valores de una variable continua.

d. **Reducción de datos y cribado de variables:** Selecciona un subconjunto útil de predictores a partir de un gran conjunto de variables para utilizarlo en la creación de un modelo paramétrico formal.

e. **Identificación de interacción:** Identifica las relaciones que pertenecen sólo a subgrupos específicos y las especifica en un modelo paramétrico formal.

f. **Fusión de categorías y unión de variables continuas:** Vuelve a codificar las variables continuas y las categorías de los predictores del grupo, con una pérdida mínima de información. [35]

En el campo de las finanzas, es compleja la utilización de árboles de decisión de la forma que se hace en este trabajo. Ya que principalmente se intentan resolver problemas de predicción que implica el uso de series de tiempo, para saber por ejemplo el retorno futuro del mercado accionario, lo que muchas veces generaría un árbol de decisiones enorme, con muchas ramas y difícil de entender.

En Diaz, Theodoulidis y Sampaio (2011), se utilizó el método CRISP-DM a través de los árboles de decisiones para describir patrones en el mercado accionario asociados a intentos de manipulación en los precios de las acciones. El objetivo del estudio es encontrar un nuevo modelo para la detección de fraudes. Los resultados del estudio corroboran patrones ya conocidos por los expertos en detección de fraude, por ejemplo, que los últimos trimestres y los fines de año, al igual que las horas cierre, son condiciones comunes para las manipulaciones, confirmando la existencia de una mayor liquidez, rentabilidad y volatilidad asociada a la muestra. Además, los resultados del estudio destacan y describen nuevos patrones: Cuando la liquidez y volatilidad están en rangos normales, una importante parte de las transacciones intradía presenta retornos anormales. Además, cuando los retornos están en los rangos normales, saltos aislados en la liquidez están asociados con transacciones sospechosas en más del 20% de los casos. En la misma línea, los saltos aislados en la volatilidad están asociados con bloques sospechosos en la misma proporción de casos.

2.2.2 Conclusiones de la Revisión Literaria en Data Mining

En general, dentro de las principales técnicas de DM, el método de redes neuronales es el más utilizado para modelar el comportamiento y proyectar valores futuros de variables financieras. En tanto, la inferencia estadística, algoritmos genéricos y métodos de regla de inducción, también tienen un rol relevante en la industria de servicios financieros. Las técnicas de visualización de data aún se emplean en aplicaciones financieras, dependiendo de la complejidad de la data y del problema a resolver. De acuerdo a lo abarcado en esta revisión literaria, hasta el momento, no existen trabajos de Data Mining para predecir algún nivel de conocimiento, ya sea en el sistema de pensiones como en otros campos.

2.3 Propuesta de investigación y objetivos

La principal motivación de este trabajo, es establecer las determinantes del nivel de conocimiento financiero en las AFP de los encuestados para la Encuesta de Protección Social, para después analizar las principales razones y generar un plan de políticas públicas para los ciudadanos chilenos.

Esto nace de la revisión en 3.1., ya que prácticamente todos los autores concluyen que en Chile las personas poseen un bajo nivel de conocimiento financiero, lo que en el contexto actual, donde empresas privadas manejan temas tan importantes como las pensiones y la salud, por lo que cada empresa tiene como objetivo siempre maximizar sus utilidades. Sucede que la información está disponible para las personas, pero ellas tienen el paradigma de que es un sistema complejo, difícil de entender, y por lo tanto no presentan mayor interés en averiguar y saber qué se hace con su dinero y cómo.

Respecto de la utilización del KDD y la DM como herramienta de investigación, la literatura financiera nos muestra su gran utilidad dada la amplia gama de problemas financieros en los que ya han sido utilizadas. Sin embargo, en específico, no fue posible encontrar trabajos de KDD o DM que se focalicen en la descripción del conocimiento financiero de los individuos, aportando esta tesis a ser uno de los primeros trabajos en lo que se apliquen dichas herramientas.

De lo anterior, el objetivo de este trabajo es responder las siguientes preguntas de investigación, utilizando herramientas de KDD y DM:

a. ¿Existen grupos homogéneos en la población que permitan determinar segmentos en base al nivel de conocimiento financiero de los individuos?, es decir, es posible determinar en cuántos grupos o clústers se divide la población según las siguientes variables: género, edad, nivel educacional, estado civil, ingreso promedio mensual, y la respuesta a

seis preguntas que están definidas en la siguiente sección, del trabajo de Berstein y Ruiz (2005).

b. ¿Cuál es el nivel de conocimiento financiero de AFP en la población?

c. Las variables definidas por Berstein y Ruiz (2005), ¿Son todas relevantes a la investigación? ¿Existen además otras variables en la EPS que complementen el estudio?

En las siguientes secciones se describe paso a paso lo realizado para poder contestar estas interrogantes.

3. Datos

Para la realización de este estudio, se utilizará la Encuesta de Protección Social (EPS) 2009. En el ámbito de la información sobre el mercado laboral y la seguridad social, es la primera encuesta longitudinal levantada en Chile y la de mayor extensión. Su contribución radica principalmente en la riqueza de la información que proporciona al abarcar en un mismo cuestionario la historia laboral y previsional de los encuestados con información detallada en áreas como educación, salud, seguridad social, capacitación laboral, patrimonio y activos, historia familiar e información sobre el hogar.

Esta encuesta fue diseñada por un comité académico internacional, que cuenta con la participación de expertos en los diversos temas, entre los que se encuentran David Bravo, Olivia Mitchell y Petra Todd, académicos que poseen investigaciones en temas del alfabetismo financiero.

La EPS consta de cinco rondas, la primera realizada el 2002, y desde entonces se hace cada aproximadamente dos años. En esta ronda del estudio, la EPS 2009 se preocupó de seguir a los individuos pertenecientes a la muestra 2006 los cuales ya habían sido entrevistados al menos una vez en rondas anteriores, es decir, no se incorporó nueva muestra.

La EPS 2009 contempló una muestra objetivo de 19.512 individuos, de ellos, se logró entrevistar a 14.243 personas. Debido a las personas no disponibles y a los fallecidos.

Para el estudio, se toman a las personas que están afiliadas al sistema previsional, ya sea porque están cotizando, cotizaron en algún momento o ya se encuentran recibiendo la pensión de este sistema. Esto quiere decir, que en la encuesta, la variable ‘e1: ¿Se encuentra afiliado al sistema previsional?’ debe

estar contestada con un ‘sí’, lo cual dejó un total de 11.411 observaciones a estudiar.

Para el primer análisis del conocimiento financiero, se considerarán una serie de variables de caracterización: género como variable dicotómica, edad, educación según nivel, ingreso mensual y estado civil. Se especifica a continuación en el Cuadro 1.

Cuadro 1
Descripción de las variables de caracterización individual

Variable	Descripción	Valor
Género	Femenino	0
	Masculino	1
Edad		
Educación	Ninguna o Analfabeto	0
	Educación Básica	1
	Educación Media Científica-Humanista	2
	Media Técnica Profesional	2,5
	Superior en Centro de Formación Técnica	3
	Superior en Instituto Profesional	4
	Superior en Universidad	4,5
	Preparatoria (Sistema Antiguo)	1
	Humanidades (Sistema Antiguo)	2
	Técnica, comercial, normalista, industrial (Sistema Antiguo)	3
	Magíster o Postgrado	5
	Diferencial (Discapacitado)	1,5
Ingreso Mensual		
Estado Civil	Soltero (a)	0
	Anulado (a)	0
	Separado (a) unión de hecho	1
	Separado (a) unión legal	1
	Divorciado (a)	1,5
	Viudo (a)	1,5
	Conviviente	2
	Casado (a)	3

Fuente: Elaboración propia a partir de los datos de la EPS 2009

Para evaluar el conocimiento financiero, se utilizarán las respuestas a seis preguntas sobre el sistema de AFP, que se especifican en el Cuadro 2.

A partir de estas preguntas se construye un índice de conocimiento financiero, siguiendo la metodología planteada en Berstein y Ruiz (2005). Cabe mencionar,

que la pregunta 3 original se tuvo que reemplazar debido a la eliminación de las comisiones fijas a las que hacía referencia.

Cuadro 2
Variables de conocimiento financiero

Variables	Descripción	Valor
e5 ¿Sabe qué porcentaje le descuentan mensualmente para el sistema de pensiones?	Sí	1
	No; No sabe; No responde	0
e61 ¿Cómo era la información contenida en la última cartola de su AFP?	Suficientemente clara	3
	Medianamente clara	2
	Confusa o poco clara	1
	No lee (leyó) la cartola; No sabe; No responde	0
e63 ¿Sabe usted cuánto hay acumulado en su Cuenta Individual?	Sí	1
	No; No sabe; No responde	0
e64 ¿Sabe usted cuánto cobra su AFP de Comisión Variable, por administrar sus fondos?	Sí	2
	No	1
	No cobran; No sabe; No responde	0
e64a ¿Quién paga las Comisiones Variables?	El afiliado con su sueldo	1
	El afiliado con su fondo de pensiones	0
	El empleador; No sabe; No responde	0
e85 ¿Sabe Usted que cumpliendo con algunos requisitos, puede tomar la opción de pensionarse anticipadamente?	Sí	1
	No; No sabe; No responde	0

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 3, se tienen los estadísticos descriptivos de las seis preguntas escogidas: recuento, mínimo, máximo y media. Esto es relevante para tener una idea general de cómo se comportan estas variables en la población.

Cuadro 3
Estadísticos descriptivos

Variables	N	Mínimo	Máximo	Media
e5	11.410	0	1	0,29
e61	11.410	0	3	1,21
e63	11.410	0	1	0,37
e64	11.410	0	2	0,86
e64a	11.410	0	1	0,46
e85	11.410	0	1	0,32

Fuente: Elaboración propia a partir de los datos de la EPS 2009

Observando estos datos, a simple vista se intuye que existe un bajo nivel de conocimiento financiero, ya que la media de cada variable, está más cercana al valor mínimo que al valor máximo que indicaría un alto alfabetismo financiero.

Cuadro 4
Distribución de las variables de conocimiento financiero

Variables	0	1	2	3	Total
e5	71.2%	28.8%	-	-	100.0%
e61	38.7%	22.6%	17.8%	20.9%	100.0%
e63	62.8%	37.2%	-	-	100.0%
e64a	18.3%	77.4%	4.4%	-	100.0%
e64a	54.1%	45.9%	-	-	100.0%
e85	68.5%	31.5%	-	-	100.0%

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 4, se tiene la distribución que tiene cada variable acerca del índice del conocimiento financiero.

Sólo un 28.8% de la muestra de afiliados sabe qué porcentaje le descuenta su administradora de fondos. Mientras que sólo un 20.9% señala que la información de la cartola que recibe es suficientemente clara. También se observa que un 37.2% de los individuos sabe cuánto tiene acumulado en su fondo de pensiones. Además, se puede observar que sólo un 4.4% sabe cuál es la comisión cobrada por parte de la AFP. Casi un 46% de las personas sabe quién paga las comisiones, y por último un 31.5% sabe que existe la opción de

pensionarse anticipadamente. Dado lo anterior, se vislumbra que la población encuestada posee un bajo nivel de conocimiento financiero.

Cuadro 5
Distribución según nivel de conocimiento

Nivel de conocimiento	Respuestas correctas	Distribución	Distribución
Bajo	0	14,20%	27,80%
	1	13,60%	
Medio	2	24,50%	45,50%
	3	21,00%	
Alto	4	15,80%	26,70%
	5	8,20%	
	6	2,70%	

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 5, se tiene la distribución de la variable dependiente “Nivel de Conocimiento”:

- Bajo: El nivel de conocimiento se considera bajo, cuando el individuo responda ninguna o una pregunta correctamente.
- Medio: Se considera nivel medio, cuando el individuo responda dos o tres preguntas correctamente.
- Alto: Se considera nivel alto, cuando el individuo responda cuatro o más preguntas correctamente.

Las preguntas que se tomarán en cuenta para evaluar el nivel de conocimiento son las descritas en el Cuadro 2.

Se observa, que el nivel de conocimiento se concentra a nivel medio, asegurando saber tan sólo dos o tres preguntas del total de seis. Sin embargo, se observa que el nivel bajo es bastante mayor de lo que se quisiera, y por el contrario, el nivel alto es el menor.

En los siguientes cuadros, se realizará una vista simple de la distribución de la variable Nivel de Conocimiento según las variables: Género, Edad y Educación.

Cuadro 6
Distribución del nivel conocimiento según género

Índice de conocimiento	Mujeres	Hombres
Bajo	30,20%	25,73%
Medio	46,90%	44,37%
Alto	22,90%	29,90%

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 6, se observa la distribución del nivel de conocimiento según género. Se ve que los hombres poseen un mayor nivel de alfabetización financiera que las mujeres.

Cuadro 7
Distribución del nivel conocimiento según edad

Edad	Bajo	Medio	Alto
<= 25	24,7%	51,5%	23,8%
26 - 35	17,4%	52,0%	30,7%
36 - 45	16,0%	52,0%	32,0%
46 - 55	21,4%	47,5%	31,1%
55 - 65	33,3%	42,3%	24,4%
65+	71,7%	22,6%	5,7%
Total	27,8%	45,5%	26,7%

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 7, se observa que el rango que mayor nivel de conocimiento financiero acumula es el rango entre 36 y 45 años. Mientras que el que menos conocimiento demuestra, es el rango de individuos mayores de 65 años.

Una idea de lo que sucede, es que una persona de alrededor 40 años, está en el momento de su vida dónde probablemente tenga familia, un trabajo estable y ya tiene una visión más clara de su futuro. Por lo que comienza a interesarse por averiguar y aprender acerca de todos los sistemas en que estamos inmersos, ya sea de salud, previsional, entre muchos otros. Por lo que es esperable que posea un mayor nivel de conocimiento financiero.

Cuadro 8

Distribución del nivel de conocimiento según educación

Educación	Bajo	Medio	Alto
Ninguna o Analfabeto	58,4%	33,3%	8,2%
Básica	41,2%	45,5%	13,3%
Media	25,2%	48,5%	26,3%
Técnica	16,8%	45,7%	37,5%
Universitario	14,4%	39,8%	45,8%
Magister o Postgrado	12,0%	28,3%	59,8%
Total	27,8%	45,5%	26,7%

Fuente: Elaboración propia a partir de los datos de la EPS 2009

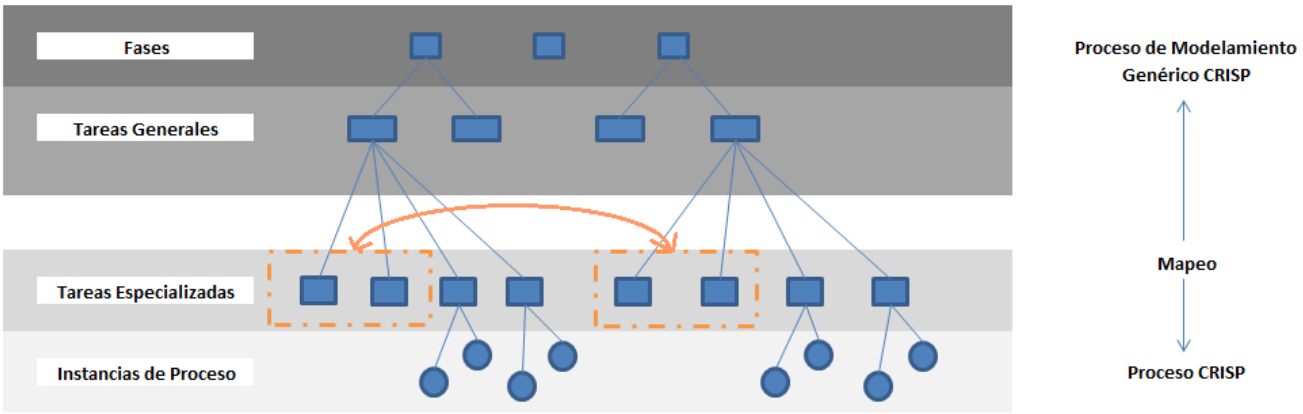
En el Cuadro 8, se observa que las personas con mayor conocimiento financiero son las que poseen altos niveles de educación, que es un resultado esperable. Es decir, mientras más educado sea el individuo, mayor será su nivel de alfabetización financiera.

4. Metodología

En este trabajo, se ocupará la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), creado a finales del año 1996 por tres líderes de la industria del Data Mining: DaimierBenz, SPSS y NCR.

La metodología se describe en términos de un proceso jerárquico, consistente en un grupo de tareas descritas en cuatro niveles de abstracción: fase, tarea genérica, tarea especializada e instancia de proceso.

Figura 2
Niveles de la Metodología CRISP-DM

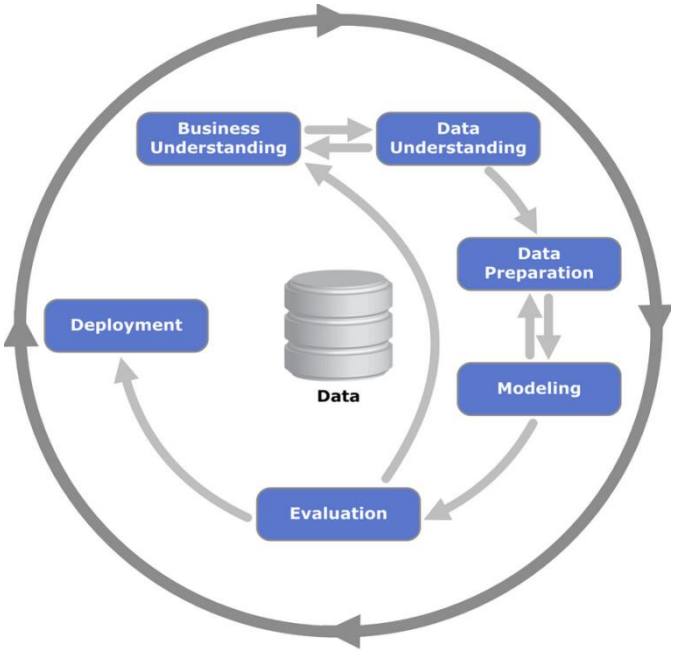


Fuente: Elaboración Propia en base a Modelo CRISP-DM, Chapman, et al (2000).

Observando la Figura 2, en el nivel superior se encuentran las Fases que se dividen en varias tareas genéricas. El segundo nivel es llamado genérico porque pretende ser lo bastante general como para cubrir todas las situaciones posibles. Ya en el tercer nivel, se describe cómo las tareas genéricas del segundo nivel han de ser tratadas en determinadas situaciones. En el último nivel, se encuentra el registro de las acciones, decisiones y resultados del proyecto representando lo que ha sucedido en un caso particular en lugar de lo que sucede a nivel general.

El modelo provee una representación completa del ciclo de vida de un proyecto de DM. Es necesario recalcar, que la secuencia de estas etapas no es estricta, y son frecuentes los movimientos hacia delante y hacia atrás. Estos dependen del resultado de cada nivel o cuál es la tarea siguiente que se ha de ejecutar.

Figura 3
Ciclo de Vida Metodología CRISP-DM



Fuente: IBM Knowledge Center (Abril, 2008)

Adicionalmente, la metodología CRISP-DM, establece una secuencia de trabajo que consta de 6 etapas: (1) Entendimiento del negocio, (2) Entendimiento de los datos, (3) Preparación de los datos, (4) Modelamiento, (5) Evaluación y (6) Implementación.

En la **Etapa de Entendimiento del Negocio** se debe definir el problema que será estudiado con técnicas de DM. En este caso particular, el problema de negocios escogido es la modelación del nivel de conocimiento financiero de los cotizantes del sistema de AFP con base en las respuestas que están

disponibles en la EPS (Encuesta de Previsión Social). Como se trata de intentar predecir la variable de nivel de conocimiento y entender las relaciones existentes entre las respuestas de los encuestados y su nivel de educación financiera, se utilizará un modelo de clasificación de árbol de decisión. Esto dada su característica de representar las relaciones existentes en los datos de manera tal, que sean de fácil manejo e interpretabilidad para los usuarios. Además, con el objeto de facilitar la descripción de las relaciones se optó por realizar un procedimiento previo, en el que utilizando algoritmos de clústers o conglomerados, se buscaron grupos o segmentos de encuestados que tuvieran características similares. Más detalles de la estrategia de modelamiento y preparación de los datos serán entregados en la descripción de las etapas respectivas.

En la **Etapa de Entendimiento de los Datos**, lo primero que se hizo fue analizar las variables disponibles en la EPS y comprender qué era cada una. Dentro de las aproximadamente 900 variables que trae esta encuesta, sólo se utilizarán las del Módulo A: Información General del Entrevistado; Módulo D: Activos y Patrimonio; Módulo E: Protección Social; Módulo J: Otros; Módulo K: Conocimiento Financiero y Habilidades No Cognitivas; Módulo I: Historia Individual; y el ingreso que percibe cada individuo, variable que fue formada mediante otras que conforman el Módulo C: Ingresos Familiares y Módulo D: Activos y Patrimonio. Para ellos se estudió con detención la estadística descriptiva, para que fueran variables que poseen desviación estándar mayor a cero, es decir, que aporte información al modelo, y el manual de datos disponible donde se especifica con detalle la metodología utilizada para realizar la encuesta, y la codificación que fue seguida en la tabulación de las respuestas. Más información acerca de las variables estudiadas y la metodología de la EPS se encuentra disponible en la sección III.

En la **Etapa de Procesamiento de Datos**, se procedió a realizar diferentes procesos tendientes a estructurar los datos de manera que se facilite su modelación utilizando el algoritmo de árbol de decisión. Como se explicó en la sección anterior, primero se creó una variable dependiente la que representa el número de respuestas correctas a las preguntas especificadas en el Cuadro 2. A partir de estas preguntas se construye un índice de conocimiento financiero, siguiendo la metodología planteada en Berstein y Ruiz (2005).

Después, se seleccionaron los individuos que estén afiliados al sistema de pensiones. Además, se recodificaron las variables como categóricas y se realizó una clasificación de la población encuestada en distintos grupos, esto con el fin de agrupar individuos con características similares, como género, edad, educación y conocimiento financiero. Luego se hizo otra clasificación dentro de estos clúster para segmentar aún más a las personas. El propósito de la agrupación, es que cuando ingrese otro individuo a la muestra, según algunas características y respuestas, ingrese a un clúster y se pueda predecir su nivel de conocimiento. En el análisis de clúster, se utilizó el procedimiento de X-Means o Bietápico el cual sugiere o detecta el número de grupos que es posible encontrar en los datos.

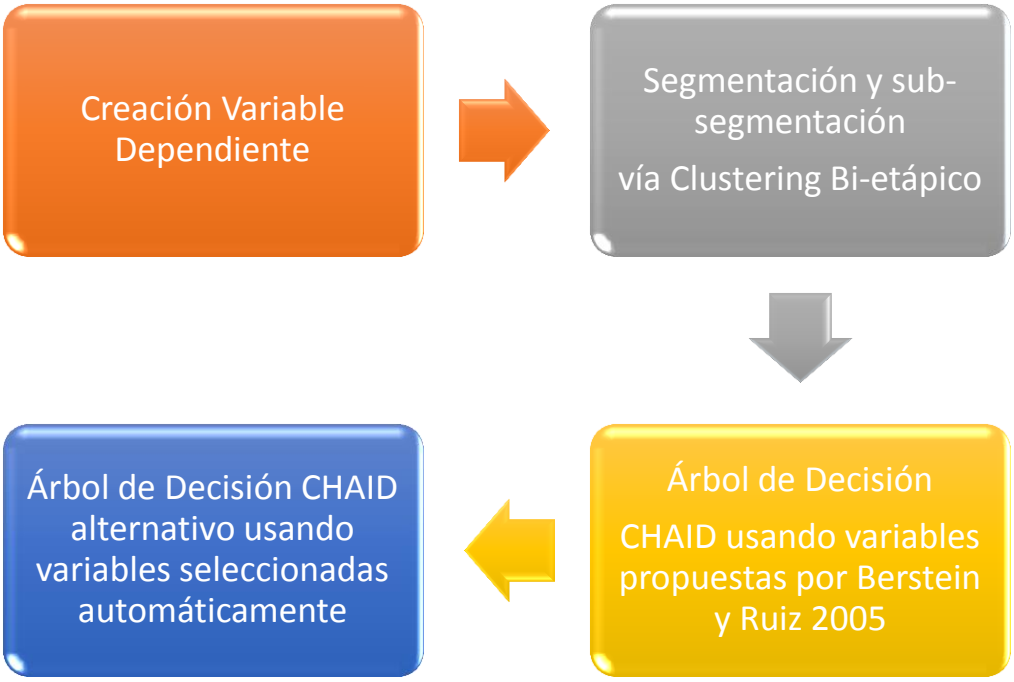
Una vez definido cada clúster y sub-clúster, en la Etapa de Modelación se procedió a generar un árbol de decisión que sea capaz de discriminar entre los diferentes niveles de conocimiento financiero que se detectan en cada uno de ellos. Dado que la cantidad de encuestados en cada uno de los niveles de conocimiento no es homogénea (existen muchos encuestados con bajo nivel de conocimiento y pocos con alto nivel de conocimiento) se optó por realizar un procedimiento de balanceo con sub-muestreo de la clase mayoritaria. Para esto, se seleccionan 3.047 casos al azar de cada clúster, ya que el más pequeño posee esta cantidad de observaciones, y luego se unió cada sub-muestra. Cabe mencionar, que del total de 9.141 observaciones, se seleccionó

el 50% para realizar la Muestra de Entrenamiento y con el otro 50% se realizó la Muestra de Prueba.

Con esto se genera un árbol de decisión de 5 niveles (Ver Sección 6) con las variables escogidas en un principio, que corresponde al género, a la edad, al nivel de educación, estado civil, ingreso y la respuesta a las seis preguntas vistas anteriormente. Se realizó con método CHAID con validación cruzada en IBM SPSS Modeler 14.2. El algoritmo seleccionó las variables que describen las seis preguntas de conocimiento, el género, el ingreso y el estado civil.

Luego de esto, la pregunta que surge es ¿Serán las únicas variables que puedan predecir el nivel de conocimiento? ¿Qué otras variables también podrían ser consideradas? Es por esto, que se realiza el mismo procedimiento de árbol de decisión de 5 niveles, CHAID con validación cruzada en IBM SPSS Modeler 14.2 generando un nuevo árbol. Esto para determinar qué variables de todas las que contiene la Encuesta de Protección Social, ayudan a predecir el nivel de conocimiento financiero de los individuos. Para esto, el único cambio que se realiza es recodificar algunas variables, así como también eliminar otras que podrían ensuciar la muestra al no ser relevantes y no aportar nueva información. La Figura 4 resume las etapas del proceso y modelación de datos.

Figura 4
Sub-etapas del Proceso de Modelación de Datos



Fuente: Elaboración Propia

En la **Etapa de Evaluación** se procedió a verificar tanto la calidad de los clústers o segmentos creados, como la capacidad de los árboles de decisión de clasificar y describir los patrones existentes entre las variables de la encuesta y el nivel de conocimiento financiero. Para estudiar la calidad de los clústers se utilizó el método de interpretación y Validación de Silueta, el cual es una representación gráfica de cuan bien cada dato se ajusta en los grupos o clústers encontrados refiriéndose a su cohesión y separación.

La medida de clasificación predeterminada, Silueta, tiene un valor predeterminado de 0 porque un valor inferior a 0 (es decir, negativo) indica que la distancia media entre un caso y los puntos de su clúster asignado es mayor que la distancia media mínima hasta los puntos de otro clúster. Por lo tanto, los modelos con una Silueta negativa pueden descartarse de manera segura.

La medida de clasificación es de hecho un coeficiente de silueta modificado, que combina los conceptos de cohesión de clústeres (favoreciendo a los modelos que contengan clústeres cohesivos) y separación de clústeres (favoreciendo a los modelos que contengan clústeres altamente separados). El coeficiente de Silueta medio es simplemente la media de todos los casos del siguiente cálculo por cada caso individual:

$$\frac{(B-A)}{\max(A,B)} \qquad (1)$$

donde A es la distancia desde el caso hasta el centroide del clúster al que pertenece el caso; y *B* es la distancia mínima desde el caso hasta el centroide de cada uno de los otros clústeres.

El coeficiente de Silueta (y su media) van desde -1 (lo que indica un modelo muy pobre) hasta 1 (lo que indica un modelo excelente). La media puede realizarse a nivel de casos totales (lo cual produce una silueta total) o a nivel de clústeres (lo cual produce una silueta de clústeres). Las distancias pueden calcularse utilizando distancias euclídeas.

Para la evaluación de la predicción del árbol de decisión, se utilizará el Nodo de Análisis que está disponible en los Nodos de Resultados del IBM SPSS Modeler 14.2. Este arroja cuatro tablas, la primera es la Comparación de los valores predichos con respecto a los valores reales, y arroja el porcentaje de predicciones correctas y erróneas. La segunda en la Matriz de Coincidencias, la cual muestra el patrón de coincidencias entre cada campo generado (predicho) y su campo objetivo para objetivos categóricos (marca, nominal u ordinal). Se muestra una tabla con filas definidas por valores reales y columnas definidas por valores predichos, con el número de registros que tienen ese patrón en cada casilla. Esto es útil para identificar errores sistemáticos en las predicciones. Si existe más de un campo generado relacionado con el mismo

campo de salida pero generado por modelos distintos, los casos en los que estos campos concuerdan y no concuerdan se cuentan y se muestran los totales. En los casos en los que concuerdan, se muestra otro conjunto de estadísticos correcto/incorrecto. La tercera es la Evaluación del Rendimiento, esta muestra estadísticos de evaluación del rendimiento para modelos con resultados categóricos. Este estadístico, mostrado para cada categoría de los campos de salida, es una medida del contenido de información medio (en bits) del modelo para predecir registros pertenecientes a dicha categoría. Se tiene en cuenta la dificultad del problema de clasificación, de forma que las predicciones precisas para categorías inusuales obtendrán un índice de evaluación del rendimiento mayor que las predicciones precisas para categorías comunes. Si el modelo no hace más que adivinar una categoría, el índice de evaluación del rendimiento para esa categoría será 0. Y la cuarta es el Informe de valores de confianza, los siguientes estadísticos se muestran para los valores de confianza modelo:

a. **Rango:** Muestra el rango (los valores máximos y mínimos) de valores de confianza para registros en los datos de la ruta.

b. **Media para correctos:** Muestra la confianza media para los registros que se han clasificado correctamente.

c. **Media para incorrectos:** Muestra la confianza media para los registros que se han clasificado de forma incorrecta.

d. **Siempre correctos por encima de:** Muestra el umbral de confianza por encima del cual las predicciones son siempre correctas y el porcentaje de casos que cumplen este criterio.

e. **Siempre incorrectos por debajo de:** Muestra el umbral de confianza por debajo del cual las predicciones son siempre incorrectas y muestra el porcentaje de casos que cumplen este criterio.

f. **X% Precisión por encima de:** Muestra el nivel de confianza en el que la precisión es *X%*. *X* es aproximadamente el valor especificado para Umbral para en las opciones de Análisis. Para algunos modelos y conjuntos de datos, no se puede elegir un valor de confianza que ofrezca el umbral exacto especificado en las opciones (normalmente debido a los clústeres de casos similares con el mismo valor confianza cerca del umbral). El umbral que se muestra es el valor más cercano al criterio de precisión especificado que se puede obtener con un solo umbral de valor de confianza.

g. **X Veces correctas por encima de:** Muestra el valor de confianza en el cual la precisión es *X* veces mejor de lo que es para el conjunto de datos global. *X* es aproximadamente el valor especificado para Mejora en la precisión en las opciones de Análisis. [36]

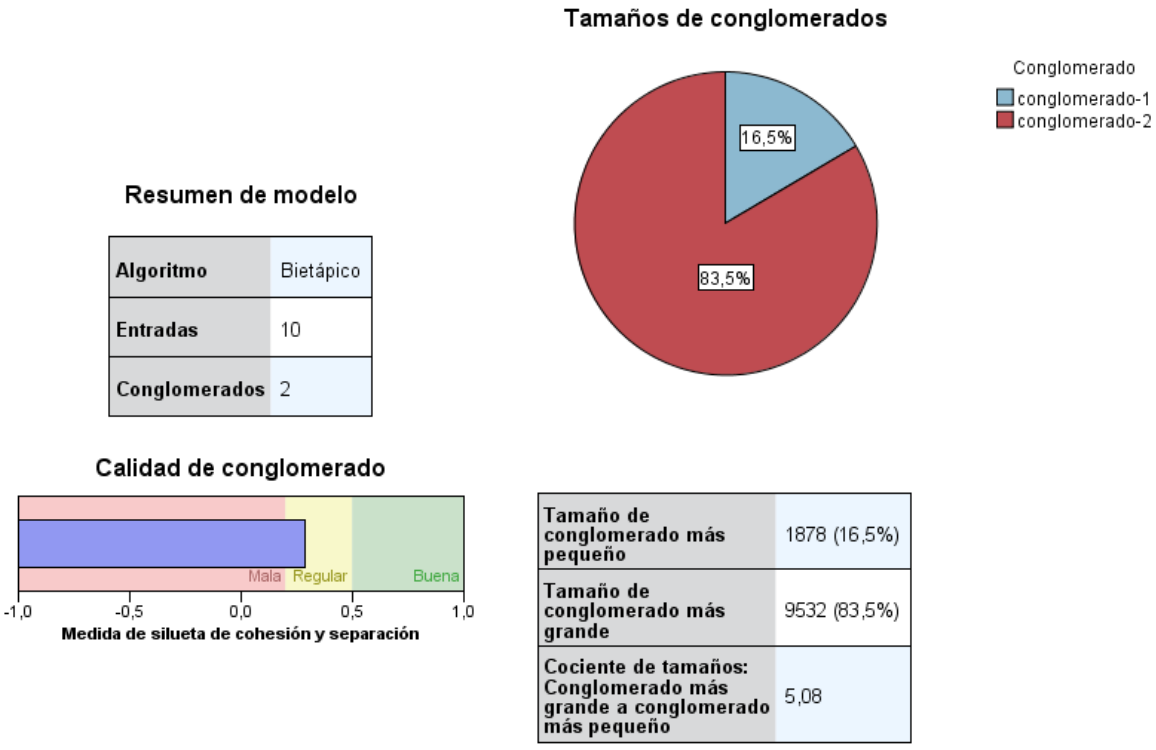
5. Resultados

En esta sección se analizarán los resultados del pre-procesamiento y generación de clúster y la implementación de los diferentes árboles de decisión y sus resultados respectivos.

5.1 Generación de Clústers

Primero, se realizó una segmentación general, mediante el procedimiento Bietápico, lo que generó simplemente dos clúster:

Figura 5
Resumen Primera Segmentación Bietápica



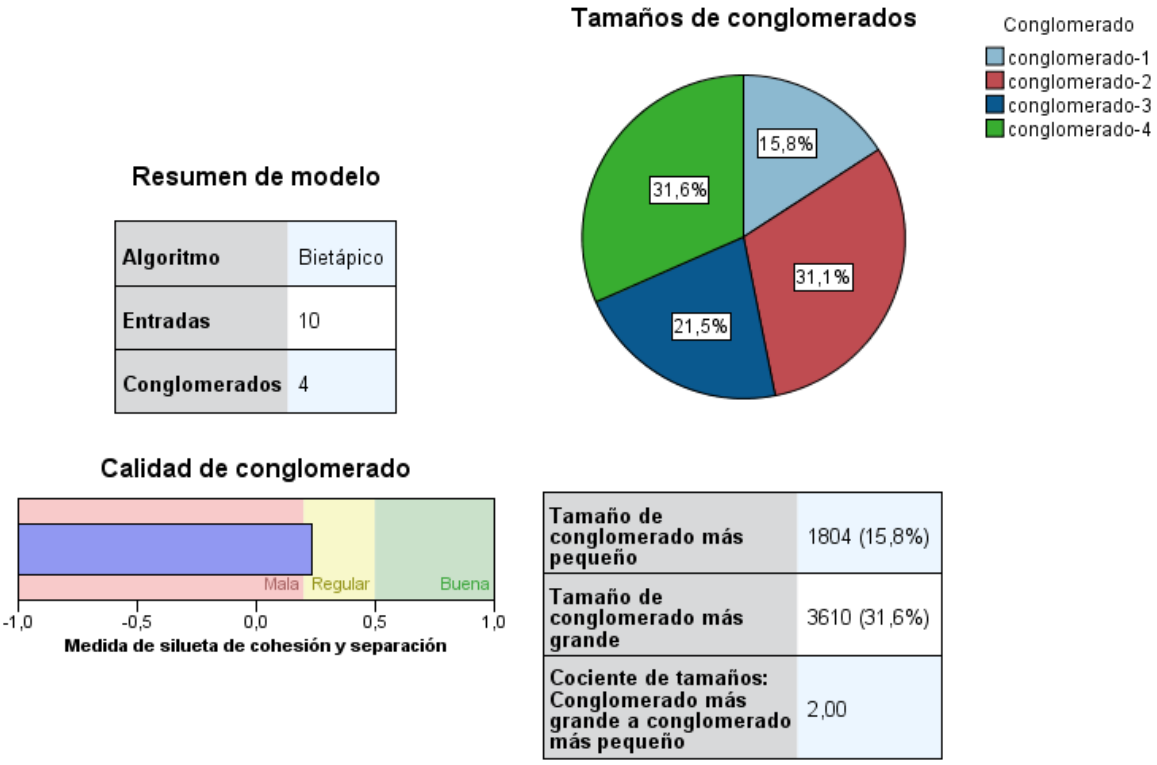
Fuente: Elaboración propia a partir de los datos de la EPS 2009

Según la Medida de silueta, es una buena segmentación, pero para fines del estudio no es lo recomendable ya que quedan bastante desbalanceados los grupos.

Por lo tanto, ahora realizando el mismo procedimiento, se le asignó al programa que creara como mínimo 4 clústers, lo que tiene como resultado:

Figura 6

Resumen Segunda Segmentación Bietápica con 4 Clústers mínimo



Fuente: Elaboración propia a partir de los datos de la EPS 2009

De esta forma, según la Medida de Silueta se mantiene una calidad aceptable, y además se tiene una segmentación más balanceada de la población.

A continuación, en el Cuadro 9, se muestra la descripción de cada uno de estos 4 clústers generados, mostrando la distribución de las variables: Género, Nivel Educacional y Estado Civil, y además la media de la variable Edad en cada grupo.

Cuadro 9

Descripción del Clúster General

Variable	Valores	Clúster 1	Clúster 2	Clúster 3	Clúster 4
Tamaño		1.804	3.545	2.451	3.610
Género	Mujeres	49,33%	47,08%	42,92%	44,65%
	Hombres	50,67%	52,92%	57,08%	55,35%
Edad		65,22	46,21	42,47	45,31
Nivel Educacional	Ninguna/Analfabeto	7,20%	2,40%	0,40%	1,80%
	Educación Básica/Preparatoria	51,00%	34,50%	6,20%	32,10%
	Diferencial (Discapacitado)	0,00%	0,00%	0,00%	0,10%
	Educación Media/Humanidades	27,20%	37,10%	28,60%	37,60%
	Educación Media Técnica	4,00%	11,70%	17,40%	11,50%
	CFT/Industrial	4,40%	3,50%	8,60%	3,50%
	IP	1,00%	4,30%	11,20%	4,70%
	Educación Universitaria	4,90%	6,00%	25,20%	8,40%
	Postgrado	0,40%	0,50%	2,30%	0,30%
Estado Civil	Soltero(a)/Anulado(a)	15,70%	22,70%	28,30%	22,30%
	Separado(a) unión de hecho o legal	7,20%	10,10%	9,10%	9,30%
	Divorciado(a)/Viudo(a)	18,50%	4,00%	1,60%	3,90%
	Conviviente	6,80%	13,00%	9,30%	13,60%
	Casado(a)	51,80%	50,20%	51,70%	50,90%

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 9, se tiene que en todas las agrupaciones hay más hombres que mujeres, la mayor diferencia está en el Clúster 3 con un 30% más de hombres. De acuerdo a la edad, el Clúster 1 es el mayor con una media de 65 años, y el más joven es el Clúster 3 con una media de 42 años.

En el Clúster 1, priman los individuos con sólo Educación Básica/Preparatoria. En el Clúster 2 y 4, se observa un nivel educacional similar, con alrededor de un 30% con Educación Básica/Preparatoria y un 35% con Educación Media. Por último, el Clúster 3 es el que posee un más alto nivel de educación con un 25% de individuos Universitarios.

Finalmente, no se observa una distribución similar en los Clúster 2, 3 y 4, en el Clúster 1 está el nivel más bajo de Solteros y un porcentaje relevante en Divorciados/Viudos.

A continuación, se muestra el Cuadro 10 con la distribución que tienen las preguntas asociadas al Nivel de Conocimiento Financiero (NCF), y en la última fila se tiene la variable como tal que muestra el NCF que posee cada Clúster.

Cuadro 10

Descripción NCF del Clúster General

Variable	Valores	Clúster 1	Clúster 2	Clúster 3	Clúster 4
¿Sabe qué porcentaje le descuentan mensualmente para el sistema de pensiones?	No; No sabe; No responde	87,6%	96,2%	7,5%	81,6%
	Sí	12,4%	3,8%	92,5%	18,4%
¿Cómo era la información contenida en la última cartola de su AFP?	No lee la cartola; No sabe; No responde	98,2%	37,2%	6,2%	32,4%
	Confusa o poco clara	1,6%	27,7%	21,0%	29,2%
	Medianamente clara	0,1%	18,4%	29,1%	18,3%
	Suficientemente clara	0,1%	16,6%	43,7%	20,1%
¿Sabe usted cuánto hay acumulado en su Cuenta Individual?	No; No sabe; No responde	99,8%	66,8%	29,1%	63,4%
	Sí	0,2%	33,2%	70,9%	36,6%
¿Sabe usted cuánto cobra su AFP de Comisión Variable, por administrar sus fondos?	No cobran; No responde	100,0%	0,0%	0,7%	7,3%
	No	0,0%	99,4%	82,7%	90,7%
	Sí	0,0%	0,6%	16,6%	1,9%
¿Quién paga las Comisiones Variables?	El afiliado con su fondo de pensiones; el empleador; No sabe; No responde	100,0%	0,0%	30,8%	100,0%
	El afiliado con su sueldo	0,0%	100,0%	69,2%	0,0%
¿Sabe Usted que cumpliendo con algunos requisitos, puede tomar la opción de pensionarse anticipadamente?	No; No sabe; No responde	98,9%	74,8%	36,4%	68,8%
	Sí	1,1%	25,2%	63,6%	31,2%
Nivel de Conocimiento Financiero	Bajo	99,9%	0,9%	0,0%	37,0%
	Medio	0,1%	79,2%	15,9%	55,3%
	Alto	0,0%	19,9%	84,1%	7,7%

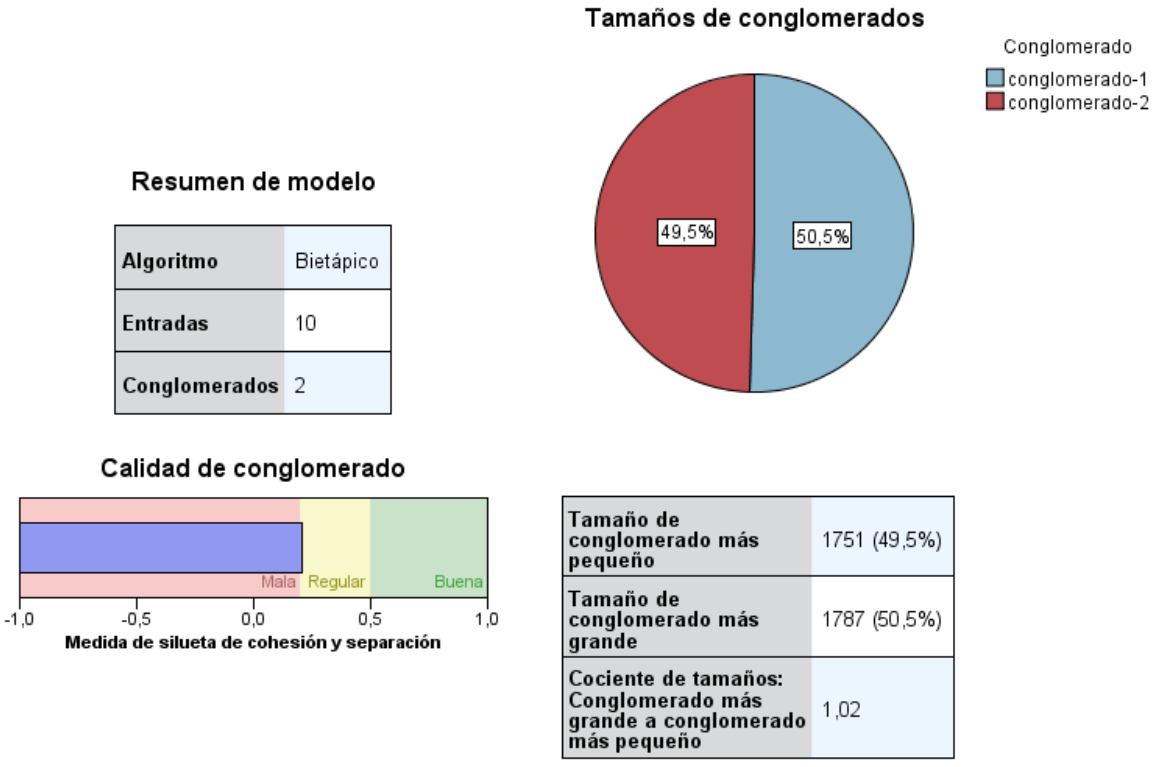
Fuente: Elaboración propia a partir de los datos de la EPS 2009

Considerando las respuestas correctas de cada individuo perteneciente a los diferentes clústers, se tiene que el Clúster 1 es que el que peor NCF, el Clúster 4 lo sigue, luego el Clúster 2 y finalmente el Clúster 3 es el que posee el mayor NCF.

Con el fin de modelar de mejor forma el nivel de conocimiento financiero de los individuos en nuestro país, se intenta buscar ciertos patrones dentro de los clústers ya formados, encontrando sub-clústers para lograr definir más detalladamente las características de las personas. Para esto, se realizó el mismo procedimiento Bietápico anterior, que sugiere la cantidad que estima óptima de clústers, para cada Clúster.

Para el Clúster 1, al momento de realizar este procedimiento, no se pudo realizar ya que prácticamente no hay patrones distintos para la formación de sub-grupos. Por lo que se continuó con el Clúster 2:

Figura 7
Resumen Segmentación Bietápica Sub-Clústers (Clúster 2)



Fuente: Elaboración propia a partir de los datos de la EPS 2009

En la Figura 7, según la Medida de Silueta, esta división provoca que esté al límite de lo aceptable en lo que respecta a calidad de los sub-clústers.

Cuadro 9
Descripción de los Sub Clústers (Clúster 2)

Variable	Valores	Clúster 1	Clúster 2
Tamaño		1.787	1.751
Género	Mujeres	49,92%	44,20%
	Hombres	50,08%	55,80%
Edad		47,03	45,41
Ingreso Promedio Mensual		144.395	296.615
Nivel Educacional	Ninguna/Analfabeto	3,8%	1,0%
	Educación Básica/Preparatoria	43,1%	25,6%
	Diferencial (Discapacitado)	0,1%	0,0%
	Educación Media/Humanidades	36,2%	37,9%
	Educación Media Técnica	9,1%	14,3%
	CFT/Industrial	2,0%	5,1%
	IP	2,5%	6,2%
	Educación Universitaria	3,2%	8,9%
	Postgrado	0,0%	1,0%
Estado Civil	Soltero(a)/Anulado(a)	24,23%	21,13%
	Separado(a) unión de hecho o legal	10,30%	9,88%
	Divorciado(a)/Viudo(a)	5,26%	2,80%
	Conviviente	13,54%	12,51%
	Casado(a)	46,67%	53,68%

Fuente: Elaboración propia a partir de los datos de la EPS 2009

De acuerdo al Cuadro 11, las principales diferencias entre los Sub-Clústers formados, es el Ingreso Promedio Mensual y el Nivel Educacional. Dónde el Clúster 2, muestra un nivel mayor de ambas variables.

A continuación, se muestra el Cuadro 12 con la distribución que tienen las preguntas asociadas al Nivel de Conocimiento Financiero (NFC), y en la última fila se tiene la variable como tal que muestra el NFC que posee cada Sub-Clúster.

Cuadro 12

Descripción NCF de los Sub-Clústers (Clúster 2)

Variable	Valores	Clúster 1	Clúster 2
¿Sabe qué porcentaje le descuentan mensualmente para el sistema de pensiones?	No; No sabe; No responde	93,56%	98,91%
	Sí	6,44%	1,09%
¿Cómo era la información contenida en la última cartola de su AFP?	No lee la cartola; No sabe; No responde	63,63%	10,28%
	Confusa o poco clara	27,98%	27,53%
	Medianamente clara	8,00%	29,07%
	Suficientemente clara	0,39%	33,12%
¿Sabe usted cuánto hay acumulado en su Cuenta Individual?	No; No sabe; No responde	100,00%	32,84%
	Sí	0,00%	67,16%
¿Sabe usted cuánto cobra su AFP de Comisión Variable, por administrar sus fondos?	No cobran; No responde	0,00%	0,00%
	No	99,16%	99,71%
	Sí	0,84%	0,29%
¿Quién paga las Comisiones Variables?	El afiliado con su fondo de pensiones; el empleador; No sabe; No responde	0,00%	0,00%
	El afiliado con su sueldo	100,00%	100,00%
¿Sabe Usted que cumpliendo con algunos requisitos, puede tomar la opción de pensionarse anticipadamente?	No; No sabe; No responde	85,84%	63,56%
	Sí	14,16%	36,44%
Nivel de Conocimiento Financiero	Bajo	1,85%	0,00%
	Medio	97,15%	60,82%
	Alto	1,01%	39,18%

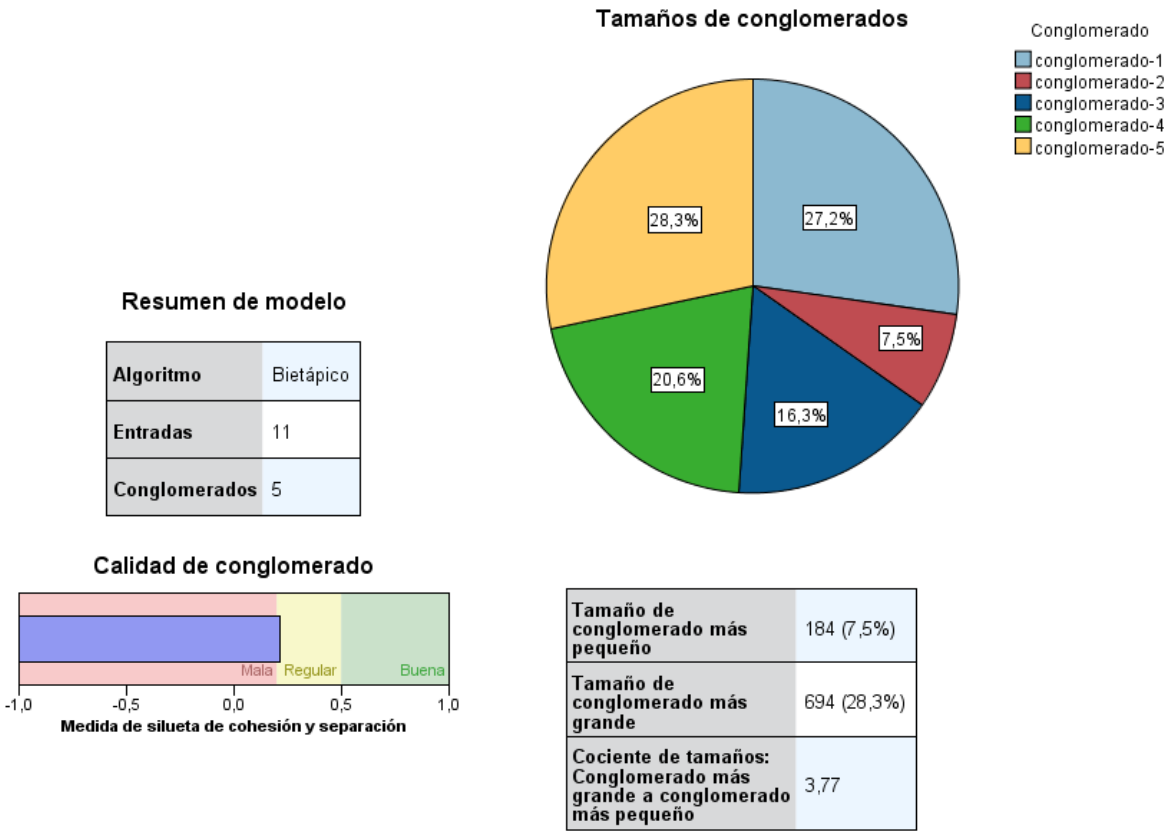
Fuente: Elaboración propia a partir de los datos de la EPS 2009

En este Cuadro se observa claramente una diferencia en el NCF, dónde el primer grupo se concentra en el nivel medio, mientras que el segundo se distribuye entre un NCF medio con un 60% y un 40% NCF alto.

Por lo tanto, se logra ver que ambos sub-clústers sí son distintos, y se corrobora la idea de que dentro de cada clúster hay patrones que se pueden seguir analizando para lograr un mejor modelamiento.

A continuación, en la Figura 8 se muestra la segmentación del Clúster 3:

Figura 8
Resumen Segmentación Bietápica Sub-Clústers (Clúster 3)



Fuente: Elaboración propia a partir de los datos de la EPS 2009

Esta vez el programa escogió 5 Sub-Clústers. Según la Medida de Silueta, está dentro de lo aceptable en lo que respecta a calidad.

En el Cuadro 13, se tiene la distribución de las variables ya mencionadas anteriormente:

Cuadro 13

Descripción de los Sub Clústers (Clúster 3)

Variable	Valores	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5
Tamaño		667	184	399	505	694
Género	Mujeres	0,0%	51,6%	47,6%	100,0%	37,8%
	Hombres	100,0%	48,4%	52,4%	0,0%	62,2%
Edad		44,3	42,5	41,9	42,5	41,0
Ingreso Promedio Mensual		577.136	1.786.301	526.365	393.022	382.971
Nivel Educacional	Ninguna/Analfabeto	0,1%	0,0%	0,5%	1,2%	0,3%
	Educación Básica/Preparatoria	6,0%	0,0%	2,3%	3,2%	12,7%
	Diferencial (Discapacitado)	0,0%	0,0%	0,0%	0,0%	0,0%
	Educación Media/Humanidades	35,8%	2,7%	20,3%	26,3%	35,2%
	Educación Media Técnica	16,8%	29,9%	15,5%	14,3%	18,0%
	CFT/Industrial	6,0%	13,6%	9,5%	13,7%	5,5%
	IP	11,2%	12,5%	11,5%	11,9%	10,1%
	Educación Universitaria	21,7%	40,2%	36,3%	26,7%	17,0%
	Postgrado	2,2%	1,1%	4,0%	2,8%	1,3%
Estado Civil	Soltero(a)/Anulado(a)	16,9%	33,2%	33,6%	33,3%	31,3%
	Separado(a) unión de hecho o legal	6,0%	10,3%	11,3%	14,1%	6,9%
	Divorciado(a)/Viudo(a)	1,3%	1,1%	1,5%	3,2%	1,0%
	Conviviente	9,3%	4,3%	8,5%	7,5%	12,1%
	Casado(a)	66,4%	51,1%	45,1%	42,0%	48,7%

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En este Cuadro se pueden observar varias diferencias que definen al Clúster. Cabe recordar, que éste es el Clúster que tiene más alto nivel de conocimiento financiero.

En la variable de género, el sub-clúster 1 tiene 100% hombres, mientras que el sub-clúster 4 tiene 100% de mujeres. En el sub-clúster 5, también se observa una gran diferencia entre sexos, con un 62% de hombres. La edad promedio, se mantiene similar, donde el sub-clúster 1 es el mayor con 44 años. Mientras que el sub-clúster 5 es el menor con 41 años. En el ingreso promedio mensual, existen diferencias relevantes. El sub-clúster 2, que es el más pequeño, tiene

un ingreso promedio de 1.786.301. Mientras que el del sub-clúster 1 y 2, se mantiene alrededor de los 550.000. Y el sub-clúster 4, con 100% de mujeres, tiene el menor ingreso promedio de 393.022.

En la variable de nivel educacional, el sub-clúster 2 resalta con un 40% de universitarios. Después lo sigue el sub-clúster 3, luego el sub-clúster 4, el sub-clúster 1 y finalmente el sub-clúster 5.

El sub-clúster 1, tiene el mayor porcentaje de individuos casados y el menor de solteros o anulados. Mientras que los demás sub-clústers tienen una distribución similar del estado civil.

Cuadro 14

Descripción NCF de los Sub-Clústers (Clúster 3)

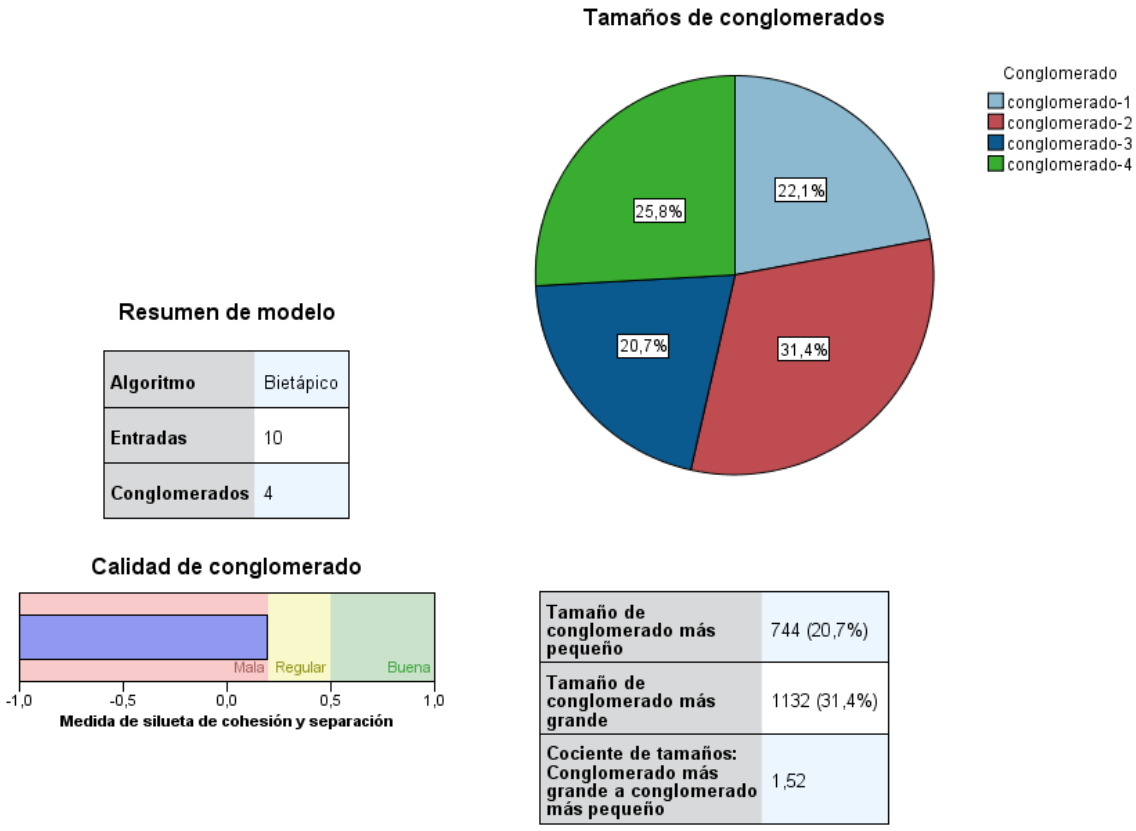
Variable	Valores	Clúster 1	Clúster 2	Clúster 3	Clúster 4	Clúster 5
¿Sabe qué porcentaje le descuentan mensualmente para el sistema de pensiones?	No; No sabe; No responde	2,8%	83,7%	0,0%	1,4%	0,4%
	Sí	97,2%	16,3%	100,0%	98,6%	99,6%
¿Cómo era la información contenida en la última cartola de su AFP?	No lee la cartola; No sabe; No responde	3,0%	1,1%	13,0%	4,6%	8,1%
	Confusa o poco clara	13,9%	8,2%	16,8%	17,6%	36,0%
	Medianamente clara	26,2%	32,6%	30,8%	30,3%	29,1%
	Suficientemente clara	56,8%	58,2%	39,3%	47,5%	26,8%
¿Sabe usted cuánto hay acumulado en su Cuenta Individual?	No; No sabe; No responde	0,4%	7,6%	100,0%	0,0%	42,8%
	Sí	99,6%	92,4%	0,0%	100,0%	57,2%
¿Sabe usted cuánto cobra su AFP de Comisión Variable, por administrar sus fondos?	No cobran; No responde	0,6%	0,0%	0,0%	2,4%	0,0%
	No	70,8%	84,2%	88,0%	79,8%	92,9%
	Sí	28,6%	15,8%	12,0%	17,8%	7,1%
¿Quién paga las Comisiones Variables?	El afiliado con su fondo de pensiones; el empleador; No sabe; No responde	45,9%	3,3%	40,6%	49,3%	4,2%
	El afiliado con su sueldo	54,1%	96,7%	59,4%	50,7%	95,8%
¿Sabe Usted que cumpliendo con algunos requisitos, puede tomar la opción de pensionarse anticipadamente?	No; No sabe; No responde	14,8%	3,3%	8,0%	23,4%	91,5%
	Sí	85,2%	96,7%	92,0%	76,6%	8,5%
Nivel de Conocimiento Financiero	Bajo	0,0%	0,0%	0,0%	0,0%	0,0%
	Medio	3,6%	2,7%	26,6%	9,3%	29,7%
	Alto	96,4%	97,3%	73,4%	90,7%	70,3%

Fuente: Elaboración propia a partir de los datos de la EPS 2009

Recordando que este es el clúster con más alto NCF. Todos los sub-clústers tienen un elevado NCF, con el nivel alto mayor al 90%, excepto el sub-clúster 3 y 5, que está cerca del 70%.

A continuación, en la Figura 9 se muestra la segmentación del Clúster 4:

Figura 9
Resumen Segmentación Bietápica Sub-Clústers (Clúster 4)



Fuente: Elaboración propia a partir de los datos de la EPS 2009

En este caso, se segmentó el Clúster 4 en 4 Sub-Clústers. Según la Medida de Silueta, la calidad del conglomerado está en el límite del nivel aceptable. Por lo tanto, se considerará la división para este Clúster.

A continuación, en el Cuadro 15, se tiene la distribución de las variables dentro de los Sub-Clústers.

Cuadro 15

Descripción de los Sub Clústers (Clúster 4)

Variable	Valores	Clúster 1	Clúster 2	Clúster 3	Clúster 4
Tamaño		795	1132	744	931
Género	Mujeres	48,9%	0,0%	100,0%	50,9%
	Hombres	51,1%	100,0%	0,0%	49,1%
Edad		46,25	44,83	42,74	47,12
Ingreso Promedio Mensual		160.290	271.188	156.470	417.428
Nivel Educacional	Ninguna/Analfabeto	4,4%	0,9%	0,9%	1,4%
	Educación Básica/Preparatoria	41,8%	34,9%	23,4%	27,3%
	Diferencial (Discapacitado)	0,0%	0,2%	0,0%	0,0%
	Educación Media/Humanidades	34,7%	38,9%	42,9%	34,2%
	Educación Media Técnica	8,2%	12,6%	13,0%	11,7%
	CFT/Industrial	2,4%	3,0%	3,8%	4,9%
	IP	3,6%	3,1%	6,5%	6,2%
	Educación Universitaria	4,5%	6,3%	9,3%	13,7%
	Postgrado	0,4%	0,2%	0,3%	0,5%
Estado Civil	Soltero(a)/Anulado(a)	24,7%	20,0%	26,1%	20,3%
	Separado(a) unión de hecho o legal	10,3%	5,1%	13,0%	10,5%
	Divorciado(a)/Viudo(a)	5,9%	1,1%	4,8%	4,7%
	Conviviente	12,5%	15,2%	13,2%	12,9%
	Casado(a)	46,7%	58,6%	42,9%	51,6%

Fuente: Elaboración propia a partir de los datos de la EPS 2009

Al igual que en la sub-segmentación anterior, se tienen dos grupos con un 100% de personas del mismo género. El Sub-Clúster 2 con un 100% de hombres y el Sub-Clúster 3 con un 100% de mujeres, mientras que los otros tienen una distribución similar del 50% aproximadamente.

La edad promedio del Sub-Clúster 3 es la menor con 42 años, mientras que la mayor es la del Sub-Clúster 4 con 47 años.

En el ingreso promedio mensual también hay diferencias. Los ingresos más bajos están en el Sub-Clúster 1 y Sub-Clúster 3 con 160.000 pesos, después los sigue el Sub-Clúster 2 con un ingreso promedio de 270.000 y el Sub-Clúster 4 es el que posee el mayor nivel con 417.000 pesos.

El nivel educacional es similar en todos los Sub-Clúster, caracterizándose por tener la mayor cantidad de individuos con Educación Media.

Con el estado civil se observa semejanza en la distribución, con alrededor del 50% de los individuos casados.

Cuadro 16

Descripción NCF de los Sub-Clústers (Clúster 4)

Variable	Valores	Clúster 1	Clúster 2	Clúster 3	Clúster 4
¿Sabe qué porcentaje le descuentan mensualmente para el sistema de pensiones?	No; No sabe; No responde	91,2%	67,8%	87,4%	85,5%
	Sí	8,8%	32,2%	12,6%	14,5%
¿Cómo era la información contenida en la última cartola de su AFP?	No lee la cartola; No sabe; No responde	99,9%	6,5%	5,8%	27,6%
	Confusa o poco clara	0,1%	46,0%	46,0%	19,8%
	Medianamente clara	0,0%	24,6%	23,3%	22,3%
	Suficientemente clara	0,0%	22,9%	25,0%	30,3%
¿Sabe usted cuánto hay acumulado en su Cuenta Individual?	No; No sabe; No responde	100,0%	46,7%	54,3%	59,5%
	Sí	0,0%	53,3%	45,7%	40,5%
¿Sabe usted cuánto cobra su AFP de Comisión Variable, por administrar sus fondos?	No cobran; No responde	0,0%	11,2%	10,5%	6,3%
	No	99,9%	87,3%	88,2%	89,2%
	Sí	0,1%	1,5%	1,3%	4,5%
¿Quién paga las Comisiones Variables?	El afiliado con su fondo de pensiones; el empleador; No sabe; No responde	100,0%	100,0%	100,0%	100,0%
	El afiliado con su sueldo	0,0%	0,0%	0,0%	0,0%
¿Sabe Usted que cumpliendo con algunos requisitos, puede tomar la opción de pensionarse anticipadamente?	No; No sabe; No responde	100,0%	82,2%	100,0%	1,0%
	Sí	0,0%	17,8%	0,0%	99,0%
Nivel de Conocimiento Financiero	Bajo	91,3%	25,7%	39,2%	2,4%
	Medio	8,7%	66,0%	60,6%	77,9%
	Alto	0,0%	8,3%	0,1%	19,8%

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En la segmentación, se nota una diferencia relevante acerca del NCF que presentan los individuos. El Sub-Clúster 1 posee más del 90% en el Nivel Bajo, tanto el Sub-Clúster 2 como el Sub-Clúster 3, tienen un 60% en el Nivel Medio y más del 25% en el Nivel Bajo. En cambio, el Sub-Clúster 4 posee un 77% en Nivel Medio y casi el 20% en Nivel Alto.

5.2 Modelación con Árboles de Decisión

5.2.1 Utilizando sólo las variables utilizadas anteriormente en la literatura

La distribución del Nivel de Conocimiento Financiero en la población es bastante desbalanceada, por lo que se debe realizar un sobre-muestreo o un sub-muestreo. Como fue mencionado anteriormente, se realizó un sub-muestreo, seleccionando 3.045 individuos de cada NCF, ya que el NCF Alto tiene este número de personas y es el más pequeño, logrando así tener una muestra balanceada para trabajar.

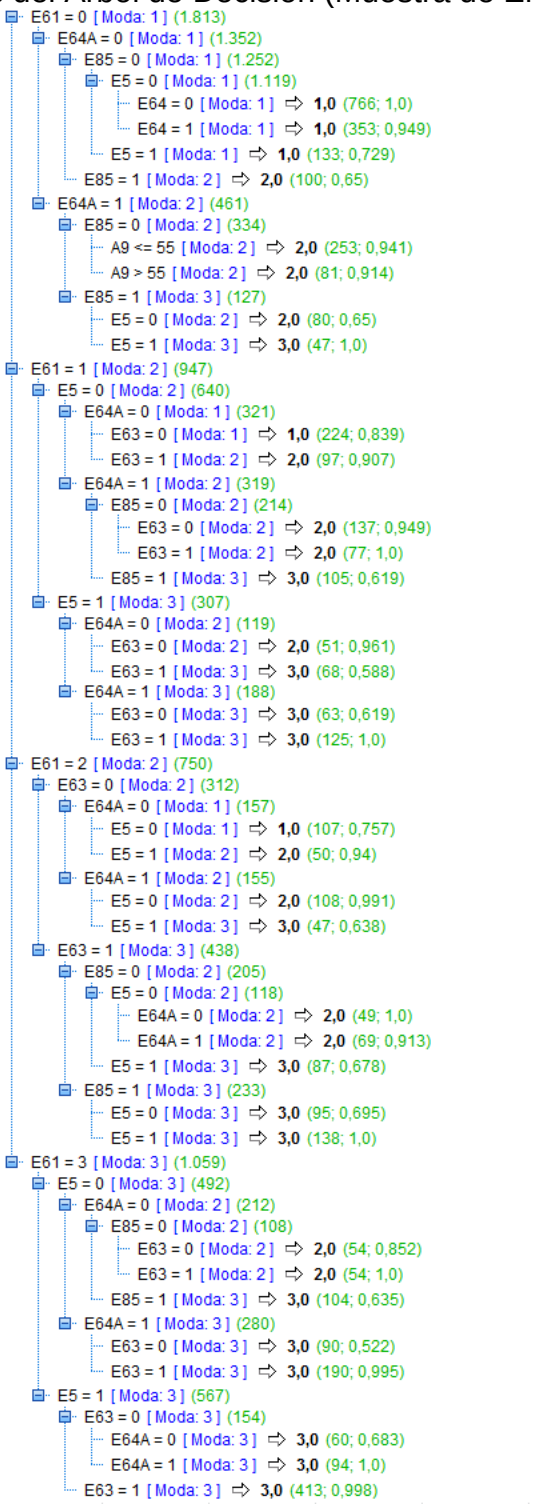
Se tienen 9.134 observaciones, por lo que se realiza una división al azar de la muestra en partes iguales, dejando 4.567 observaciones para una Muestra de Entrenamiento y 4.567 para la Muestra de Prueba.

Primero, se genera un Árbol de Decisión de 5 niveles donde se utilizaron las variables de la literatura, descritas en el Cuadro 1 y Cuadro 2, para la modelación. Esto se hizo para observar la importancia de cada variable escogida. El Árbol de formó con el método CHAID con validación cruzada en IBM SPSS Modeler 14.2.

En la Figura 10, se tienen las reglas del árbol de decisión creado con la Muestra de Entrenamiento, donde en azul se muestra la moda, es decir, la respuesta que más se repite en esa cadena. En verde se muestra la cantidad de individuos que pertenecen a esa cadena, y los que están al final muestran el número de observaciones que llegaron ahí y el nivel de confianza. Los números en negrita, representan el fin de una cadena mostrando el valor que tomó la variable.

Figura 10

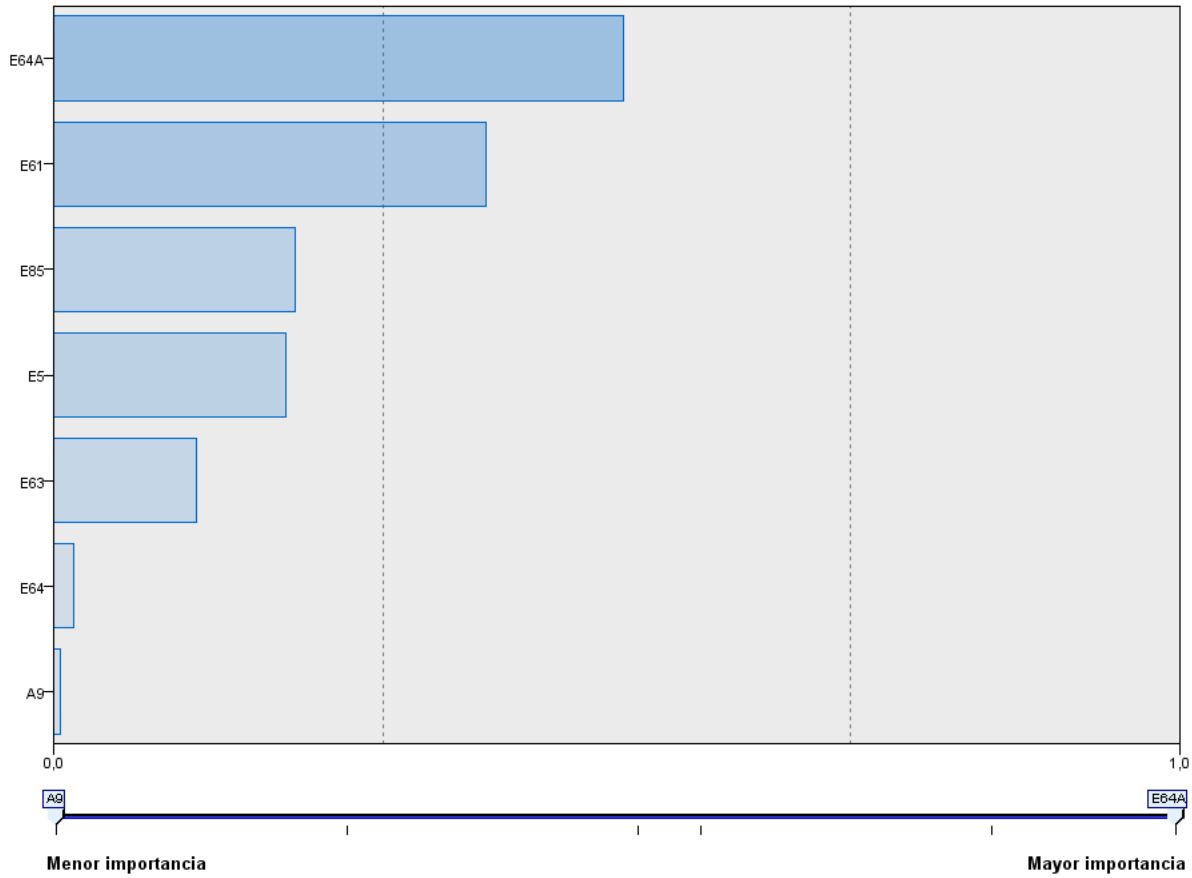
Reglas del Árbol de Decisión (Muestra de Entrenamiento)



Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Árbol de Decisión se pueden observar las variables escogidas por el algoritmo, pero queda más claro en la Figura 11, a continuación:

Figura 11
Importancia del Predictor (Muestra de Entrenamiento)



Fuente: Elaboración propia a partir de los datos de la EPS 2009

Las seis preguntas escogidas fueron relevantes para la formación del Árbol de Decisión, pero de las variables de caracterización, ninguna fue considerada para el árbol. Y en la Figura 11, alcanza a tener una pequeña importancia la Edad.

A continuación están los cuadros de análisis de resultados para el árbol generado.

Cuadro 17
Comparando $R-Y con Y
(Muestra de Entrenamiento)

Correctos	4.064	88,95%
Erróneos	505	11,05%
Total	4.569	

Fuente: Elaboración propia a partir de los datos de la EPS 2009

Se define Y como la variable dependiente, que en este caso corresponde al Nivel de Conocimiento Financiero.

En el Cuadro 17 se tiene la cantidad de observaciones que fueron predichas de forma correcta o errónea. El Árbol de Decisión predijo correctamente el 88,95%, el cual corresponde a un alto porcentaje.

Cuadro 18
Matriz de coincidencias para $R-Y
(Muestra de Entrenamiento)

	Bajo	**Medio**	**Alto**
Bajo	1.467	56	0
Medio	116	1.139	268
Alto	0	65	1.458

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 18 se tiene la matriz de coincidencias, donde las filas muestran las reales. Es decir, de las que son NCF Bajo, 1.467 fueron predichas como NCF Bajo, 56 NCF Medio y 0 NCF Alto. De las que son NCF Medio, 116 fueron predichas como NCF Bajo, 1.139 NCF Medio y 268 NCF Alto. Y de las que son NCF Alto, 0 fueron predichas como NCF Bajo, 65 NCF Medio y 1.458 NCF Alto.

Cuadro 19
Evaluación del Rendimiento
(Muestra de Entrenamiento)

Bajo	1,023
Medio	0,998
Alto	0,930

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 19 se muestra la evaluación del rendimiento. Las predicciones precisas para categorías inusuales obtendrán un índice de evaluación del rendimiento mayor que las predicciones precisas para categorías comunes. Si el modelo no hace más que adivinar una categoría, el índice de evaluación del rendimiento para esa categoría será 0. En este caso se puede ver que están elevados, por lo que tiene un buen nivel predictivo y no utiliza el azar, sino que cierto algoritmo.

Cuadro 20
Informe de valores de confianza para $RC-Y
(Muestra de Entrenamiento)

Rango	0,516-0,997
Media para correctos	0,902
Media para incorrectos	0,692
Siempre correctos por encima de	0,993 (16,77% de casos)
Siempre incorrectos por encima de	0,516 (0% de casos)
90,16% precisión por encima de	0,577
2,0 veces correctas por encima de	0,946 (66,66% de casos)

Fuente: Elaboración propia a partir de los datos de la EPS 2009

El Cuadro 20 muestra un informe de los valores de confianza. Primero se muestra el rango de confianza, luego la confianza media para los registros que se han clasificado correctamente, que es de un 90%. En tercer lugar se muestra la confianza media para los registros que se han clasificado de forma incorrecta, que es de un 69%. En cuarto y quinto lugar, se muestran los umbrales desde que las predicciones son siempre correctas e incorrectas. Después, el umbral

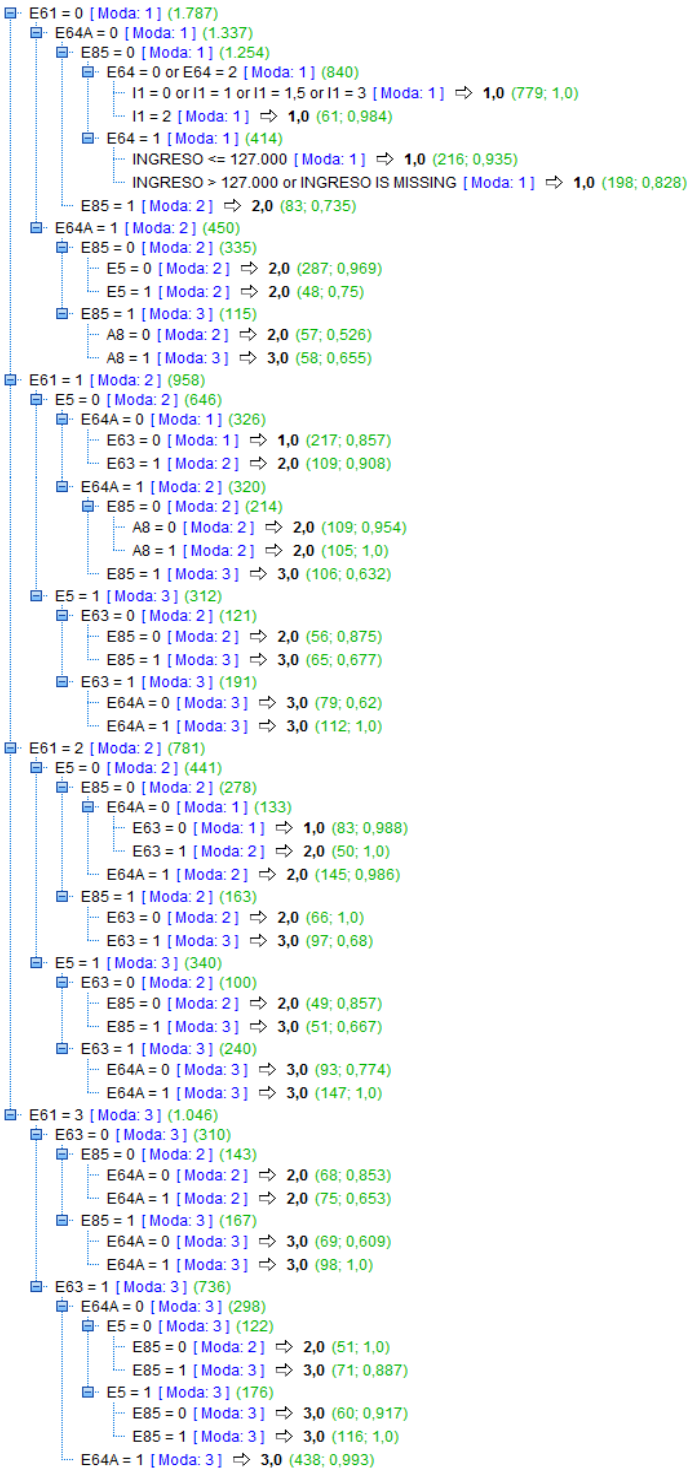
de 57% de confianza, posee una precisión de 90,16%. Y finalmente, 94% es el umbral de confianza, donde la precisión es 2 veces mejor de lo que es para el conjunto de datos global.

Luego de realizar el Proceso de Entrenamiento, se realiza la Prueba para observar el error real del modelo.

Para esto se procede a generar un Árbol de Decisión de 5 niveles, manteniendo los mismos parámetros que anteriormente.

A continuación, en la Figura 12, se tienen las reglas del Árbol de Decisión con la Muestra de Prueba.

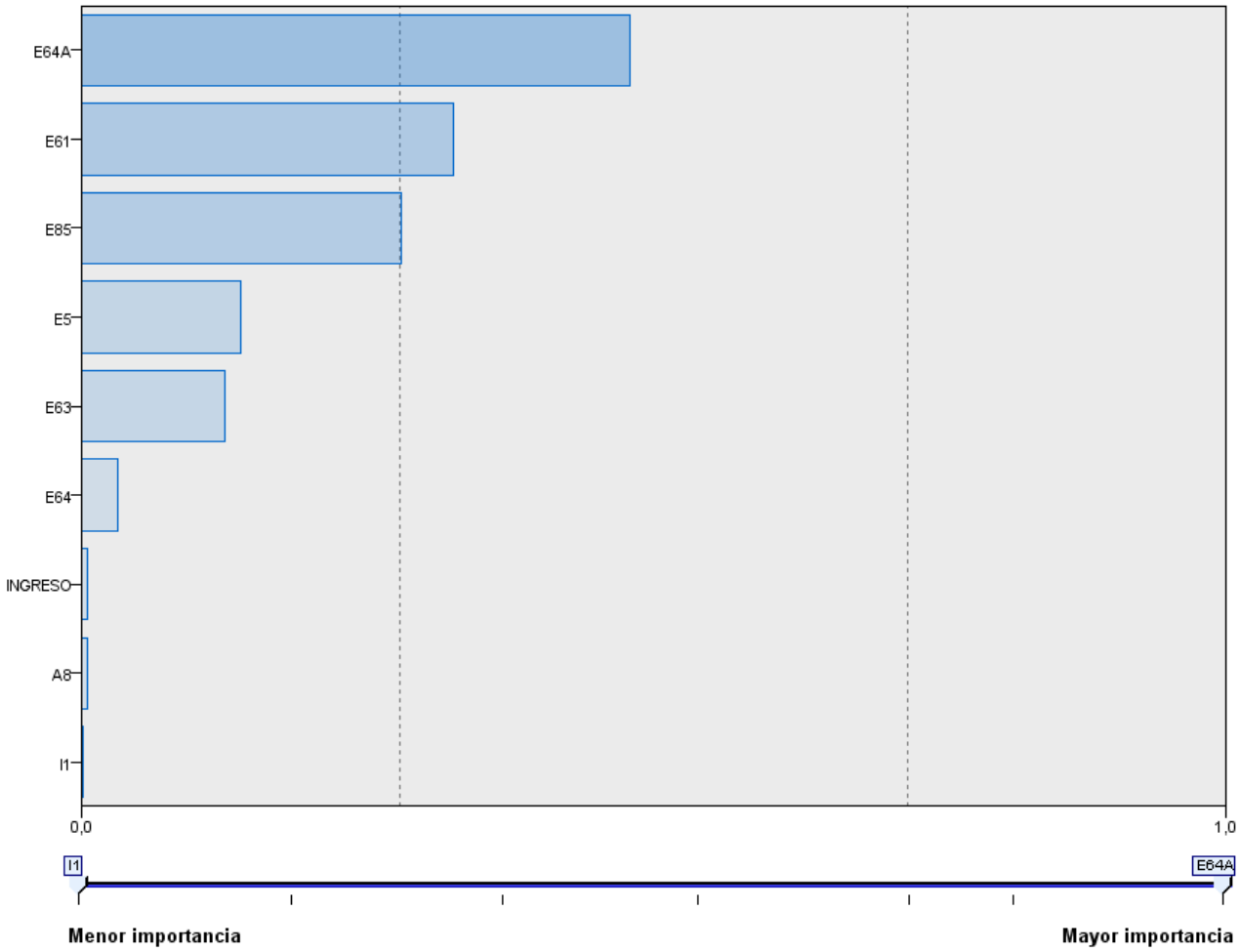
Figura 12
Reglas del Árbol de Decisión (Muestra de Prueba)



Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Árbol de Decisión se pueden observar las variables escogidas por el algoritmo, pero queda más claro en la Figura 13, a continuación:

Figura 13
Importancia del Predictor (Muestra de Prueba)



Fuente: Elaboración propia a partir de los datos de la EPS 2009

Las seis preguntas escogidas fueron relevantes para la formación del Árbol de Decisión. De las variables de caracterización, el algoritmo escogió el ingreso promedio mensual, el género y el estado civil, a pesar de que en la Figura 13 su importancia es bastante baja.

A continuación están los cuadros de análisis de resultados para el árbol generado.

Cuadro 21
Comparando $R-Y con Y
(Muestra de Prueba)

Correctos	4.132	90,38%
Erróneos	440	9,62%
Total	4.572	

Fuente: Elaboración propia a partir de los datos de la EPS 2009

Se define Y como la variable dependiente, que en este caso corresponde al Nivel de Conocimiento Financiero.

En el Cuadro 21 se tiene la cantidad de observaciones que fueron predichas de forma correcta o errónea. El Árbol de Decisión predijo correctamente el 90,38%, el cual corresponde a un alto porcentaje.

Cuadro 22
Matriz de coincidencias para $R-Y
(Muestra de Prueba)

	Bajo	**Medio**	**Alto**
Bajo	1.437	51	0
Medio	81	1.221	222
Alto	0	86	1.438

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 22 se tiene la matriz de coincidencias, donde las filas muestran las reales. Es decir, de las que son NCF Bajo, 1.437 fueron predichas como NCF Bajo, 51 NCF Medio y 0 NCF Alto. De las que son NCF Medio, 81 fueron predichas como NCF Bajo, 1.221 NCF Medio y 222 NCF Alto. Y de las que son NCF Alto, 0 fueron predichas como NCF Bajo, 86 NCF Medio y 1.438 NCF Alto.

Cuadro 23
Evaluación del Rendimiento
(Muestra de Prueba)

Bajo	1,045
Medio	0,992
Alto	0,955

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 23 se muestra la evaluación del rendimiento. En este caso se puede ver que están elevados, por lo que tiene un buen nivel predictivo y no utiliza el azar, sino que cierto algoritmo.

Cuadro 24
Informe de valores de confianza para $RC-Y
(Muestra de Prueba)

Rango	0,517-0,997
Media para correctos	0,91
Media para incorrectos	0,717
Siempre correctos por encima de	0,989 (17,04% de casos)
Siempre incorrectos por encima de	0,517 (0% de casos)
90,16% precisión por encima de	0
2,0 veces correctas por encima de	0,953 (72,09% de casos)

Fuente: Elaboración propia a partir de los datos de la EPS 2009

El Cuadro 24 muestra un informe de los valores de confianza. Primero se muestra el rango de confianza, luego la confianza media para los registros que se han clasificado correctamente, que es de un 91%. En tercer lugar se muestra la confianza media para los registros que se han clasificado de forma incorrecta, que es de un 71%. En cuarto y quinto lugar, se muestran los umbrales desde que las predicciones son siempre correctas e incorrectas. Después, el umbral de 0% de confianza, posee una precisión de 90,16%. Y finalmente, 95% es el umbral de confianza, donde la precisión es 2 veces mejor de lo que es para el conjunto de datos global.

5.2.2 Utilizando todas las variables disponibles en la EPS

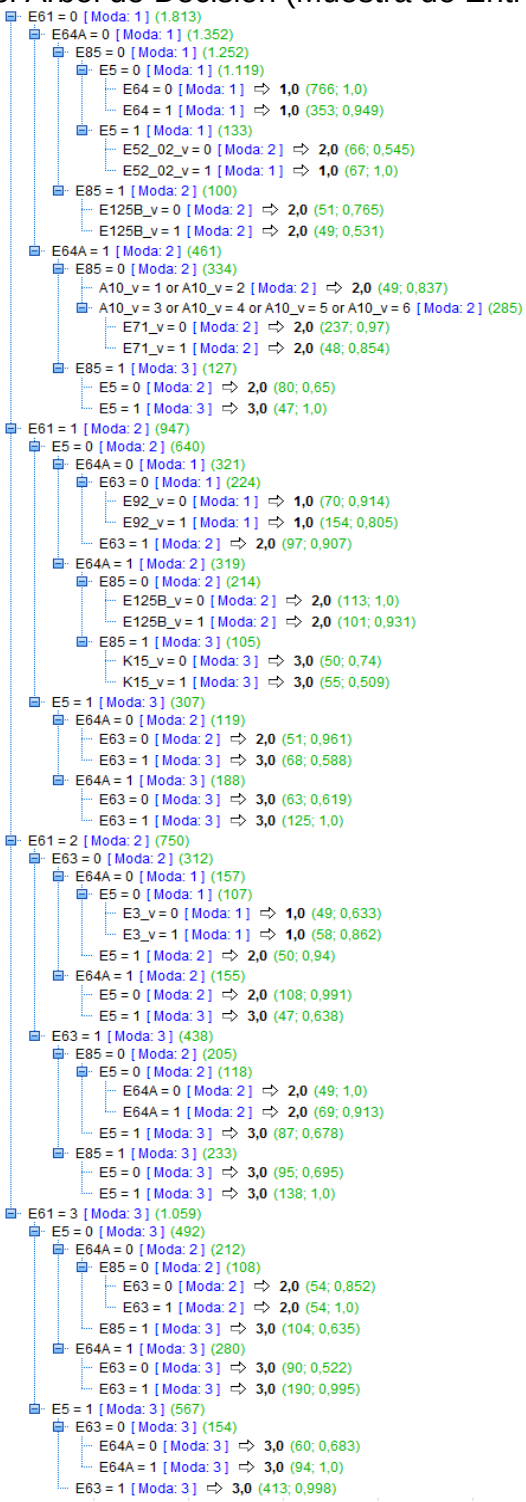
Luego de esto, la pregunta que surge es ¿Serán las únicas variables que puedan predecir el nivel de conocimiento? ¿Qué otras variables también podrían ser consideradas? Es por esto, que se realiza el mismo procedimiento de árbol de decisión de 5 niveles, CHAID con validación cruzada en IBM SPSS Modeler 14.2 generando un nuevo árbol. Esto para determinar qué variables de todas las que contiene la Encuesta de Protección Social, ayudan a predecir el nivel de conocimiento financiero de los individuos.

Para esto, el único cambio que se realiza es recodificar algunas variables, así como también eliminar otras que podrían ensuciar la muestra al no ser relevantes y no aportar nueva información.

A continuación, en la Figura 14, se tienen las reglas del Árbol de Decisión con la Muestra de Entrenamiento.

Figura 14

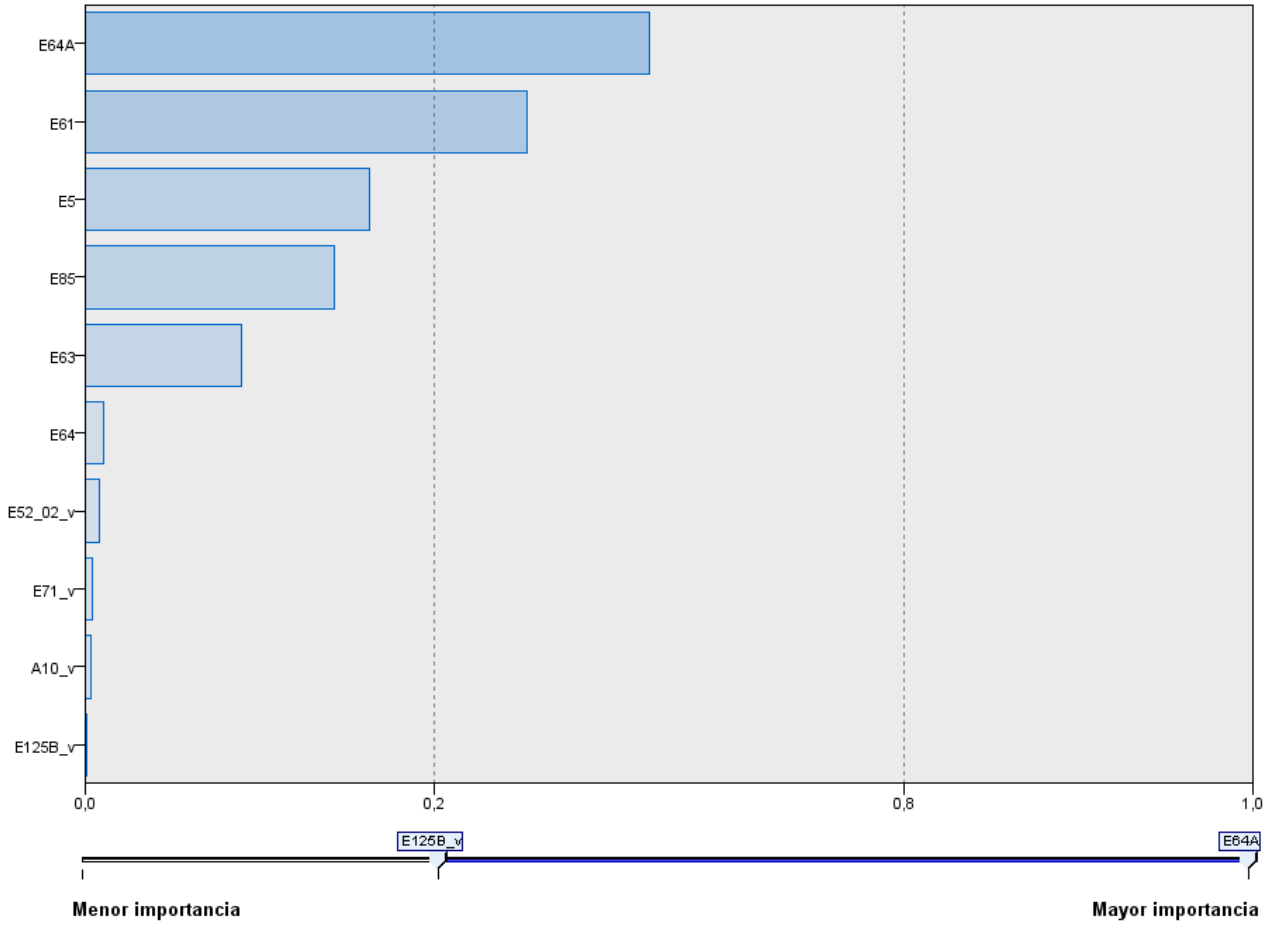
Reglas del Árbol de Decisión (Muestra de Entrenamiento)



Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Árbol de Decisión se pueden observar las variables escogidas por el algoritmo, pero queda más claro en la Figura 15, a continuación:

Figura 15
Importancia del Predictor (Muestra de Entrenamiento)



Fuente: Elaboración propia a partir de los datos de la EPS 2009

Incluyendo todas las variables con que cuenta la EPS, las mismas seis preguntas fueron escogidas con la mayor importancia. Esta vez, ninguna variable de caracterización es relevante para la formación del Árbol de Decisión.

Pero sí surgieron nuevas, estas están descritas a continuación junto a los valores que toman en el Cuadro 25.

Cuadro 25
Nuevas variables de conocimiento financiero
(Muestra de Entrenamiento)

Variables	Descripción	Valor
e52_02_v una vez que deje de trabajar, ¿Cómo piensa financiar su vejez? Con una pensión del INP	Sí	1
	No; No sabe; No responde	0
e71_v ¿Conoce o ha escuchado hablar de los multifondos?	Sí	1
	No; No sabe; No responde	0
a10 ¿Cuál es su percepción de salud?	Excelente	6
	Muy buena	5
	Buena	4
	Regular	3
	Mala	2
	Muy Mala	1
	No sabe; No responde	0
e125b_v ¿Ud. Ha sido entrevistado para la ficha de protección social?	Sí	1
	No; No sabe; No responde	0

Fuente: Elaboración propia a partir de los datos de la EPS 2009

Por lo que ya se tiene un atisbo de que sí existen otras variables que también explican el Nivel de Conocimiento Financiero y pueden complementar a las seis preguntas anteriores para lograr un mejor modelo de predicción.

A continuación están los cuadros de análisis de resultados para el árbol generado.

Cuadro 26
Comparando $R-Y con Y
(Muestra de Entrenamiento)

Correctos	4.070	89,08%
Erróneos	499	10,92%
Total	4.569	

Fuente: Elaboración propia a partir de los datos de la EPS 2009

Se define Y como la variable dependiente, que en este caso corresponde al Nivel de Conocimiento Financiero.

En el Cuadro 26 se tiene la cantidad de observaciones que fueron predichas de forma correcta o errónea. El Árbol de Decisión predijo correctamente el 89,08%, el cual corresponde a un alto porcentaje.

Cuadro 27
Matriz de coincidencias para $R-Y
(Muestra de Entrenamiento)

	Bajo	**Medio**	**Alto**
Bajo	1.437	86	0
Medio	80	1.175	268
Alto	0	65	1.458

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 27 se tiene la matriz de coincidencias, donde las filas muestran las reales. Es decir, de las que son NCF Bajo, 1.437 fueron predichas como NCF Bajo, 86 NCF Medio y 0 NCF Alto. De las que son NCF Medio, 80 fueron predichas como NCF Bajo, 1.175 NCF Medio y 268 NCF Alto. Y de las que son NCF Alto, 0 fueron predichas como NCF Bajo, 65 NCF Medio y 1.458 NCF Alto.

Cuadro 28
Evaluación del Rendimiento
(Muestra de Entrenamiento)

Bajo	1,044
Medio	0,978
Alto	0,930

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 28 se muestra la evaluación del rendimiento. En este caso se puede ver que están elevados, por lo que tiene un buen nivel predictivo y no utiliza el azar, sino que cierto algoritmo.

Cuadro 29
Informe de valores de confianza para $RC-Y
(Muestra de Entrenamiento)

Rango	0,5-0,997
Media para correctos	0,904
Media para incorrectos	0,667
Siempre correctos por encima de	0,993 (16,77% de casos)
Siempre incorrectos por encima de	0,5 (0% de casos)
90,16% precisión por encima de	0,516
2,0 veces correctas por encima de	0,953 (66,66% de casos)

Fuente: Elaboración propia a partir de los datos de la EPS 2009

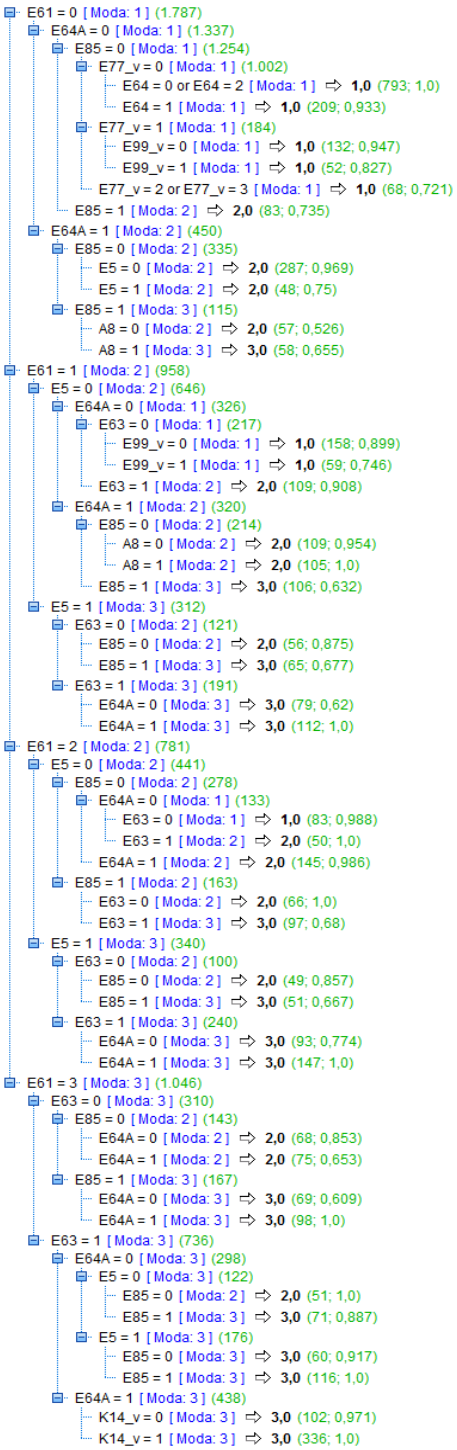
El Cuadro 29 muestra un informe de los valores de confianza. Primero se muestra el rango de confianza, luego la confianza media para los registros que se han clasificado correctamente, que es de un 90%. En tercer lugar se muestra la confianza media para los registros que se han clasificado de forma incorrecta, que es de un 66%. En cuarto y quinto lugar, se muestran los umbrales desde que las predicciones son siempre correctas e incorrectas. Después, el umbral de 51% de confianza, posee una precisión de 90,16%. Y finalmente, 95% es el umbral de confianza, donde la precisión es 2 veces mejor de lo que es para el conjunto de datos global.

Luego de realizar el Proceso de Entrenamiento, se realiza la Prueba para observar el error real del modelo.

Para esto se procede a generar un Árbol de Decisión de 5 niveles, manteniendo los mismos parámetros que anteriormente.

A continuación, en la Figura 16, se tienen las reglas del Árbol de Decisión con la Muestra de Prueba.

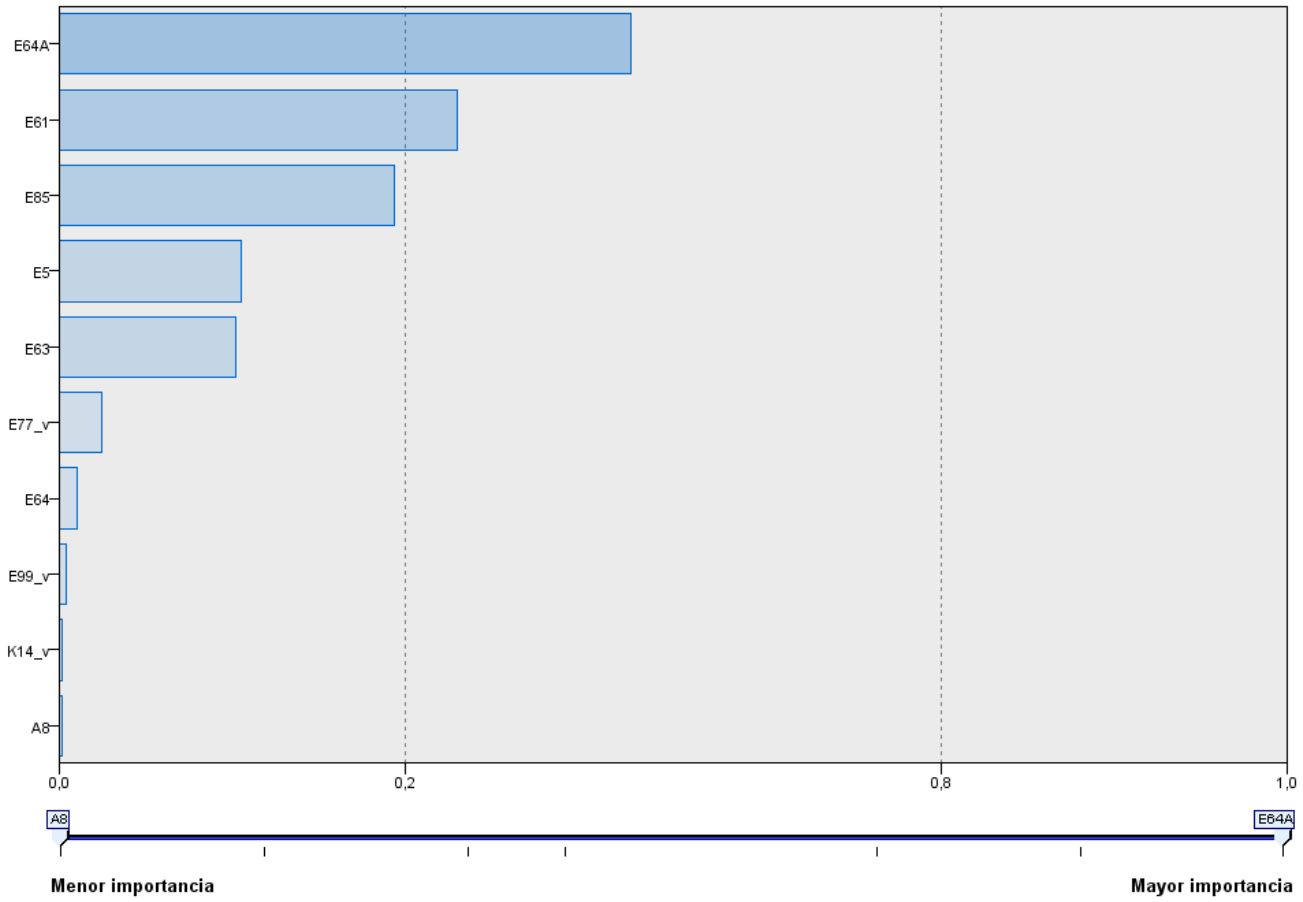
Figura 16
Reglas del Árbol de Decisión (Muestra de Prueba)



Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Árbol de Decisión se pueden observar las variables escogidas por el algoritmo, pero queda más claro en la Figura 17, a continuación:

Figura 17
Importancia del Predictor (Muestra de Prueba)



Fuente: Elaboración propia a partir de los datos de la EPS 2009

Incluyendo todas las variables con que cuenta la EPS, las mismas seis preguntas fueron escogidas con la mayor importancia. Pero, hay una nueva variable que tiene una mayor importancia que una de las escogidas en un comienzo. Esta vez, el género es la única variable de caracterización que es relevante para la formación del Árbol de Decisión.

Las nuevas variables estas están descritas a continuación junto a los valores que toman en el Cuadro 30.

Cuadro 30
Nuevas variables de conocimiento financiero
(Muestra de Prueba)

Variables	Descripción	Valor
e77_v Al afiliarse al sistema o cuando los multifondos fueron introducidos en 2002, ¿Eligió usted el tipo de fondo para sus ahorros previsionales?	Sí, realicé trámite e indique a que fondos quería pertenecer	2
	No, no supe elegir y por lo tanto dejé que la AFP me asignara	1
	No; No sabe; No responde	0
e99_v Un trabajador dependiente del sector privado, con un contrato indefinido, que pierde su empleo, ¿Puede recibir beneficios del seguro de cesantía?	Sí	1
	No; No sabe; No responde	0
k14_v ¿Con qué frecuencia Ud. lleva un control de los gastos?	Siempre; La mayor parte del tiempo	1
	Rara vez; Nunca; No sabe; No responde	0

Fuente: Elaboración propia a partir de los datos de la EPS 2009

Finalmente, se logra llegar a nuevas variables que sí son relevantes para la predicción del conocimiento financiero, de esta forma se puede realizar un modelamiento más preciso.

A continuación están los cuadros de análisis de resultados para el árbol generado.

Cuadro 31

Comparando $R-Y con Y

Correctos	4.132	90,38%
Erróneos	440	9,62%
Total	4.572	

Fuente: Elaboración propia a partir de los datos de la EPS 2009

Se define Y como la variable dependiente, que en este caso corresponde al Nivel de Conocimiento Financiero.

En el Cuadro 31 se tiene la cantidad de observaciones que fueron predichas de forma correcta o errónea. El Árbol de Decisión predijo correctamente el 90,38%, el cual corresponde a un alto porcentaje.

Cuadro 32

Matriz de coincidencias para $R-Y

	Bajo	**Medio**	**Alto**
Bajo	1.437	51	0
Medio	81	1.221	222
Alto	0	86	1.438

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 32 se tiene la matriz de coincidencias, donde las filas muestran las reales. Es decir, de las que son NCF Bajo, 1.437 fueron predichas como NCF Bajo, 51 NCF Medio y 0 NCF Alto. De las que son NCF Medio, 81 fueron predichas como NCF Bajo, 1.221 NCF Medio y 222 NCF Alto. Y de las que son NCF Alto, 0 fueron predichas como NCF Bajo, 86 NCF Medio y 1.438 NCF Alto.

Cuadro 33

Evaluación del Rendimiento

Bajo	1,045
Medio	0,992
Alto	0,955

Fuente: Elaboración propia a partir de los datos de la EPS 2009

En el Cuadro 33 se muestra la evaluación del rendimiento. En este caso se puede ver que están elevados, por lo que tiene un buen nivel predictivo y no utiliza el azar, sino que cierto algoritmo.

Cuadro 34

Informe de valores de confianza para $RC-Y

Rango	0,517-0,997
Media para correctos	0,91
Media para incorrectos	0,711
Siempre correctos por encima de	0,973 (37,34% de casos)
Siempre incorrectos por encima de	0,517 (0% de casos)
90,16% precisión por encima de	0
2,0 veces correctas por encima de	0,952 (70,42% de casos)

Fuente: Elaboración propia a partir de los datos de la EPS 2009

El Cuadro 34 muestra un informe de los valores de confianza. Primero se muestra el rango de confianza, luego la confianza media para los registros que se han clasificado correctamente, que es de un 91%. En tercer lugar se muestra la confianza media para los registros que se han clasificado de forma incorrecta, que es de un 71%. En cuarto y quinto lugar, se muestran los umbrales desde que las predicciones son siempre correctas e incorrectas. Después, el umbral de 0% de confianza, posee una precisión de 90,16%. Y finalmente, 95% es el umbral de confianza, donde la precisión es 2 veces mejor de lo que es para el conjunto de datos global.

Tomando en consideración los cuadros de análisis de las muestras de prueba, de V.2.1 y V.2.2., se concluye que los modelos generados son útiles para evaluar el nivel de conocimiento financiero. Esto porque tienen un alto nivel predictivo (alrededor del 90%), en la matriz de coincidencias se observa con claridad esto, ya que en la diagonal se concentra el número de observaciones predichas y además, los niveles extremos están en cero. En cuanto al rendimiento, se tienen valores altos, lo que implica que el modelo no es azaroso, sino que el algoritmo funciona como modelo de predicción. Finalmente, en el análisis de los valores de confianza, tiene un 90% de confianza, un nivel alto para el modelo de árbol de decisión. Además, en la muestra de prueba con todas las variables de la EPS, el valor “Siempre correctos por encima de”, aumenta al doble con la entrada de nuevas variables, lo que es relevante para el estudio.

6. Conclusiones

Considerando la situación del sistema previsional chileno, se puede ver que presenta serias dificultades en el ámbito de la competencia. Esto es principalmente por la concentración existente, dado que existen pocas AFP, es un mercado muy regulado, y la competencia se da principalmente por los agentes de venta y no por la rentabilidad o comisiones ofertadas, lo que provoca que las personas que no conozcan en profundidad el sistema de pensiones, se dejen llevar por los “regalos” en vez de tomar una decisión en base a datos que afecten en su futuro retiro. El principal problema para los individuos se da porque el sistema es complejo y obligatorio, entonces no se dan el tiempo de conocer e informarse acerca de las AFP y sus planes de retiro, para analizar cuál es la que más se ajusta a sus necesidades.

En el comienzo de este estudio, al analizar las variables, quedó en evidencia que un alto porcentaje de los encuestados posee un nivel de conocimiento financiero bastante pobre. Dada la representatividad de la Encuesta de Protección Social, y además de la evidencia que existe, se puede decir que en nuestro país existe un bajo nivel de alfabetización financiera y se deben tomar acciones al respecto. Esto es no sólo porque afecta su participación en el sistema de pensiones, sino que afecta en todos los aspectos financieros de su vida, como por ejemplo el crédito.

Este trabajo intenta determinar qué variables son las que afectan en el conocimiento financiero de los individuos para así, predecir según ellas cuál su nivel de alfabetización financiera. Para esto, se utilizó la metodología CRISP-DM, la cual consta de seis etapas a seguir: Entendimiento del Negocios, Entendimiento de los datos, Preparación de los datos, Modelación, Evaluación e Implementación. En la primera etapa, se define el problema a estudiar, la modelación del nivel de conocimiento financiero de los cotizantes del sistema

de AFP en base a las respuestas disponibles en la EPS a través de un árbol de decisión. En la segunda etapa, se analizaron todas las variables disponibles en la EPS, cerca de 900, para formar una base más acotada con la información relevante para el estudio. Se observó la estadística descriptiva, como la media y desviación estándar para saber si aportaban información. Ya en la tercera etapa, se procede a limpiar y recodificar la base. Por ejemplo, variables que tienen una respuesta abierta se sacaron ya que es difícil medir y asignarle un puntaje al set de respuestas de las personas, también se recodificaron todas las variables ya que sus valores comenzaban desde el 1 y no tenían una relación asociada a algún orden, por lo que para formar un buen árbol se codificaron las variables desde 0 asociadas a las respuesta correctas, es decir, mientras mejor sea la respuesta, mayor será el valor asociado. Y de acuerdo a las variables de caracterización, como la educación, se asocia a mayor nivel mayor valor.

Se realizó un análisis de clúster utilizando el procedimiento de X-Means, con el fin de agrupar personas con características similares como género, edad, educación y por ende conocimiento financiero. Para así, cuando ingrese otro individuo a la muestra, según algunas características y respuestas, ingrese a un clúster y se pueda predecir su nivel de conocimiento.

Como la muestra queda desbalanceada en cuando a nivel de conocimiento, para realizar un buen árbol de decisión y evitar el sesgo, se realizó un sub-muestreo, dejando así dos muestras con el mismo número de observaciones con nivel de conocimiento financiero alto, medio y bajo. En la Etapa de Modelación se procede a generar el árbol, de 5 niveles, con las variables escogidas en un comienzo (ver Cuadro 1 y Cuadro 2). Se realizó con método CHAID con validación cruzada en IBM SPSS Modeler 14.2. El algoritmo seleccionó las variables que describen las seis preguntas seleccionadas, el género, el ingreso promedio mensual y el estado civil. Luego, con el fin de encontrar nuevas variables para complementar el modelo, se realiza el mismo

procedimiento de árbol pero con todas las variables disponibles en la EPS, y finalmente se obtienen cuatro variables relevantes para complementar el modelo de Berstein y Ruiz (2005), que son: Al afiliarse al sistema o cuando los multifondos fueron introducidos en 2002, ¿Eligió usted el tipo de fondo para sus ahorros previsionales?; Un trabajador dependiente del sector privado, con un contrato indefinido, que pierde su empleo, ¿Puede recibir beneficios del seguro de cesantía?; ¿Con qué frecuencia Ud. lleva un control de los gastos?.

En la Etapa de Evaluación se procedió a verificar tanto la calidad de los clústers o segmentos creados, como la capacidad de los árboles de decisión de clasificar y describir los patrones existentes entre las variables de la encuesta y el nivel de conocimiento financiero. Para estudiar la calidad de los clústers se utilizó el método de interpretación y Validación de Silueta, el cual es una representación gráfica de cuan bien cada dato se ajusta en los grupos o clústers encontrados refiriéndose a su cohesión y separación. Para la evaluación de la predicción del árbol de decisión, se utilizó el Nodo de Análisis que está disponible en los Nodos de Resultados del IBM SPSS Modeler 14.2.

Analizando los resultados de las muestras de prueba, se concluye que los árboles de decisión generados son útiles para evaluar el nivel de conocimiento financiero. Esto porque tienen un alto nivel predictivo (alrededor del 90%), en la matriz de coincidencias se observa con claridad esto, ya que en la diagonal se concentra el número de observaciones predichas y además, los niveles extremos están en cero. En cuanto al rendimiento, se tienen valores altos, lo que implica que el modelo no es azaroso, sino que el algoritmo funciona como modelo de predicción. Finalmente, en el análisis de los valores de confianza, tiene un 90% de confianza, un nivel alto para el modelo de árbol de decisión. Es destacable, que un modelo simple, fácil de comprender, tenga resultados como tales.

En comparación con el trabajo de Berstein y Ruiz (2005), donde se estima la probabilidad de que según las variables de caracterización tenga cierto nivel de conocimiento financiero de AFP, se obtienen resultados similares para el índice, ya que la mayoría de la población se concentra en el Nivel Medio, pero un porcentaje relevante está en el Nivel Bajo. Para las variables de caracterización, los resultados de este trabajo no son comparables ya que acá se segmentó la población para analizar más en detalle qué grupos poseen cierto nivel de conocimiento financiero.

Finalmente, es importante destacar que a pesar de que aún no se tienen resultados sobre una generación que haya cotizado toda su vida activa, han pasado más de 30 años desde su implementación y aun así las personas no tienen claro de qué forma funciona. Lo que claramente los afectará al momento de jubilarse, ya que gran parte no se ha preocupado de ahorrar lo suficiente para poder obtener una pensión que cumpla el objetivo principal que es mantener el nivel de ingresos de sus últimos años trabajados. Esto merece la preocupación del gobierno, porque cada vez la población se hace más longeva, por lo que el estado tendrá que realizar mayor gasto en ellos y puede aumentar la situación de pobreza. Por lo tanto, se requiere aún más regulación a las AFP para que informen de forma más clara, se rebajen las comisiones y se muevan más los fondos para generar rentabilidad. Además, se necesita que se eduque a la población desde la etapa escolar sobre finanzas básicas, ya que dado el contexto actual es necesario que cada individuo se interese sobre su propia salud y pensión, sepa cómo funciona el sistema y cuál es la mejor opción para cada uno.

Otro método para ayudar a la educación de las personas en este tema, es la creación de cursos on-line y para las personas de más escasos recursos o de tercera edad, los municipios se podrían encargar de tener cursos para el manejo financiero en todos los aspectos de la vida.

En este último tiempo, Shawn Cole, profesor de Harvard University, estudió de qué forma los niños aprenden sobre el dinero. Concluyó que es mejor enseñarles matemáticas que finanzas. Por lo que en Chile, se debería realizar un experimento para averiguar qué es lo que debe ir en el currículum educativo para que los niños tengan conciencia sobre los temas financieros. Un estudio interesante sería formarles una base matemática para que en los últimos años de educación se les integren asignaturas de finanzas básicas. Otro estudio sería analizar cómo afecta el comportamiento financiero de los padres en los hijos, debería existir una alta correlación, y ver si un periodo de educación lo puede alterar.

Bibliografía

[1] Arrau, P. y Valdés, S. (2002): “Para desconcentrar los fondos de pensiones y aumentar la competencia en su administración”. Estudios Públicos N° 85, Centro de Estudios Públicos Chile, Verano 2002.

[2] Alessie, R., Lusardi, A. y Van Rooij, M. (2007): “Financial Literacy and Stock Market Participation”. Journal of Financial Economics, Vol. 101 N°2, p. 449-472, Agosto 2007.

[3] Berlanga, V., Rubio, M. y Vila, R. (2013): “Cómo aplicar árboles de decisión en SPSS”. REIRE, Revista d’Innovació i Recerca en Educació, Vol. 6 N°1, p. 65-79, Enero 2013.

[4] Berstein, S. (2011): “Implementación de la reforma previsional”. Documento de trabajo N° 45, Superintendencia de Pensiones Chile, Abril 2011.

[5] Bernstein, S. y Cabrita, C. (2007): “Los determinantes de la elección de AFP en Chile: Nueva evidencia a partir de datos individuales”. Estudios de Economía, Vol. 34 N°1, p. 53-72, Junio 2007.

[6] Berstein, S. y Micco, A. (2002): “Turnover and Regulation: The Chilean Pension Fund Industry”. Documento de trabajo N° 180, Banco Central de Chile, Septiembre 2002.

[7] Berstein, S. y Ruiz, J. (2005): “Sensibilidad de la Demanda con Consumidores Desinformados: El Caso de las AFP en Chile”. Documento de trabajo N° 4, Superintendencia de Administradoras de Fondos de Pensiones, Abril 2005.

[8] Bertranou, F., Gana, P. y Vásquez, J. (2006): “Pensiones no contributivas. Su relevancia en la reforma previsional”. Serie OIT Notas N°3, OIT, Mayo 2006.

[9] Cerda, R. (2006): “Movilidad en la Cartera de Cotizantes por AFP: La importancia de ser primero en rentabilidad”. Documento de Trabajo N° 309, Instituto de Economía, Pontificia Universidad Católica de Chile, Abril 2006.

[10] Cerullo, M. y Cerullo, V. (1999): “Using neural networks to predict financial reporting fraud”. Computer Fraud & Security, Vol. 1999 N°5, p. 14-17, Mayo 1999.

[11] Cole, S. y Kartini, G. (2008): “If You Are So Smart, Why Aren’t You Rich? The Effects of Education, Financial Literacy and Cognitive Ability on Financial Market Participation”. Working Paper 09-071, Harvard Business School, Diciembre 2008.

[12] Cole, S., Kartini, G., Paulson, A. (2014): “Smart Money? The Effect of Education on Financial Outcomes”. Rev. Financ. Stud. 2014: hhu012v1-hhu012, Junio 2014.

[13] Corbo, V. y Schmidt-Hebbel, K. (2003): "Efectos Macroeconómicos de la Reforma de Pensiones en Chile". Presentado en la Federación Internacional de Administradoras de Fondos de Pensiones, Cancún, Mayo 2003.

[14] Courchane, M., Gailey, A. y Zorn, P. (2007): “Consumer Credit Literacy: What Price Perception?”. Journal of Economic and Business, Vol. 60 N°1-2, p. 125-138, Abril 2007.

[15] Díaz, D., Sapaio, P., Theodoulidis, B. (2011): “Analysis of Stock Market Manipulations Using Knowledge Discovery Techniques Applied to Intraday Trade Price”. Expert Systems with Application Journal, Vol. 38, N° 10, p. 12757-12771, Junio 2011.

[16] Dunham, M. (2003), “Data Mining, Introductory an Advanced Topics”. Prentice Hall, 2003, 315 páginas.

[17] Dvorak, Tomas y Hanley, Henry (2010): “Financial literacy and the design of the retirement plans”. Journal of Socio-Economics, Vol. 39, p. 645-652, Abril 2010.

[18] Fajnzylber, E. y Reyes, G. (2011): “Knowledge, Information and Retirement Saving decisions: Evidence from a large scale intervention in Chile”. Documento de Trabajo, Escuela de Gobierno, Universidad Adolfo Ibañez, Mayo 2011.

[19] Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P. (1996): “From Data Mining to Knowledge Discovery in Databases”. American Association for Artificial Intelligence, p. 37-54, Otoño 1996.

[20] Guardia, A., Clark, R. y Martner, G. (2007): “Rompiendo mitos: La reforma del sistema de pensiones en Chile”. Fundación Friedrich Ebert, p. 15-62, 2007.

[21] Holzmann, R. y Hinz, R. (2005): “Old-Age Income Support in the Twenty-first Century: An international perspective on pension systems and reform”. World Bank, Washington DC, Febrero 2005.

[22] Kirkos, E., Spathis, C. y Manolopoulos, Y. (2007): “Data mining techniques for the detection of fraudulent financial statements”. Science Direct, ELSEVIER, Expert Systems with Applications 32, p. 995–1003.

[23] Lusardi, A. (2008): “Financial literacy: An essential tool for informed consumer choice?”. NBER Working Paper N° 14084, Cambridge, MA, Junio 2008.

[24] Lusardi, A. (2012): “Numeracy, Financial Literacy, and Financial Decision-Making”. NBER Working Paper N° 17821, Febrero 2012.

[25] Lusardi, A., Keller, P. y Keller, A. (2009): “New ways to make people save: A social marketing approach”. NBER Working Paper N° 14715, Febrero 2009.

[26] Lusardi, A., Mitchell, O. y Curto, V. (2009): “Financial literacy and financial sophistication among older americans”. NBER Working Paper N° 15469, Noviembre 2009.

[27] Lusardi, A. y Tufano, P. (2008): “Debt literacy, financial experiences and overindebtedness”. NBER Working Paper N° 14808, Cambridge, MA, Diciembre 2008.

[28] Marinovic, I. y Valdés, S. (2010): “La demanda de las AFP Chilenas: 1993-2002”. Documento de Trabajo N°369, Instituto de Economía Pontificia Universidad Católica de Chile, Enero 2010.

[29] Mitchell, O. (2010): “Implications of the financial crisis for long run Retirement Security”. Pension Research Council Working Paper N° 2010-02, Enero 2010.

[30] Mitchell, O., Mottola, G., Utkus, S. y Yamaguchi, T. (2006): “The inattentive participant: Porftolio trading behavior in 401(k) plans”. Pension Research Council Working Paper N° 2006-05, Philadelphia, PA, Junio 2006.

[31] Mitchell, O., Todd, P. y Bravo, D. (2007): “Learning from the Chilean experience: the determinants of pension switching”. Serie Documentos de Trabajo N°266, Departamento de Economía, Universidad de Chile, Octubre 2007.

[32] Skog, J. (2006): “Who knows what about their pensions? Financial literacy in the Chilean individual account system”. Population Aging Research Center Working Paper Series N° 06-11, Septiembre 2006.

[33] Valdés, S. (2005): “Para aumentar la competencia entre las AFP”. Publicado en Estudios Públicos N° 98, Centro de Estudios Públicos Chile, Otoño 2005.

[34] Weber, R. (2000): “Data Mining en la empresa y en las finanzas utilizando tecnologías inteligentes”. Revista Ingeniería de Sistemas, Volumen XIV N°1, Junio 2000.

[35] Witten, I. y Frank, E. (2005): “Data Mining: Practical machine learning tools and techniques”. ELSEVIER, Segunda Edición.

[36] Zhang, D. y Zhou, L (2004): “Discovering Golden Nuggets: Data Mining in Financial Application”. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 34 N° 4, p. 513-522, 2004.

[37] IBM Knowledge Center (2008): “Modelos de árboles de decisión”. [http://www-01.ibm.com/support/knowledgecenter/SS3RA7_16.0.0/com.ibm.spss.modeler.help/clementine/nodes_treebuilding.htm?lang=es]

[38] IBM Knowledge Center (2008): “Nodos de resultados”. [http://www-01.ibm.com/support/knowledgecenter/SS3RA7_16.0.0/com.ibm.spss.modeler.help/clementine/outputnodes_container.htm?lang=es]

[39] Subsecretaría de Previsión Social: “Documentos EPS”. [http://www.previsionsocial.gob.cl/subprev/?page_id=7516]

Anexos

Anexo 1: Cajas que existían en el sistema antiguo de previsión.

Instituciones de Previsión fusionadas en el IPS (ex INP):	
S.S.S.	Servicio de Seguro Social
Canaempu	Caja Nacional de Empleados Públicos y Periodistas
Empart	Caja de Previsión de Empleados Particulares
Bancaria	Caja Bancaria de Pensiones, Sección de Previsión del Banco Central de Chile, Caja de Previsión y Estímulo del Banco de Chile
Cajaferro	Caja de Retiro y Previsión Social de los Ferrocarriles del Estado
Camuval	Caja de Previsión Social de los Empleados Municipales de Valparaíso
Caprebech	Caja de Previsión y Estímulo de los Empleados del Banco del Estado de Chile
Capremer	Caja de Previsión de la Marina Mercante Nacional Sección Oficiales y Empleados y Sección Tripulantes de Naves y Operarios Marítimos
Capremusa	Caja de Previsión de los Empleados Municipales de Santiago
Capresomu	Caja de Previsión Social de los Obreros Municipales de la República
Emos	Caja de Previsión de los Empleados y Obreros de la Empresa Metropolitana de Obras Sanitarias, Departamentos de Empleados y Departamento Obreros
Gasco	Sección de Previsión Social de los Empleados de la Compañía de Consumidores de Gas de Santiago
Gildemeister	Caja de Previsión Gildemeister
Hípica	Caja de Previsión Social de la Hípica Nacional
Hoschild	Caja de Previsión de los Empleados de Mauricio Hotschild
Salitre	Caja de Previsión para Empleados del Salitre