OPTIMIZACIÓN DE LA CAPTACIÓN POR AFILIACIONES EN UNA AFP Y SU IMPACTO EN LA ESTRATEGIA COMERCIAL

Carlos Cateriano-Yáñez y José Cueto-Portocarrero

Lima, marzo de 2015

FACULTAD DE INGENIERÍA

Área Departamental de Ingeniería Industrial y de Sistemas

Cateriano, C. y Cueto J. (2015). *Optimización de la captación por afiliaciones en una AFP y su impacto en la estrategia comercial*. Tesis de pregrado no publicado en Ingeniería Industrial y de Sistemas. Universidad de Piura. Facultad de Ingeniería. Programa Académico de Ingeniería Industrial y de Sistemas. Lima, Perú.

U N I V E R S I D A D DE P I U R A

FACULTAD DE INGENIERIA

OPTIMIZACIÓN DE LA CAPTACIÓN POR AFILIACIONES EN UNA AFP Y SU IMPACTO EN LA ESTRATEGIA COMERCIAL

Tesis para optar el Título de

Ingeniero Industrial y de Sistemas

CARLOS CATERIANO YÁÑEZ

JOSÉ ALEJANDRO CUETO PORTOCARRERO

Lima, Marzo 2015

A nuestro asesor Víctor Manco y a nuestras familias y a todas las personas que hicieron posibles la realización de esta tesis, en especial al Ingeniero Luis Flores por su importante apoyo.

Prólogo

En el año 2011, como respuesta a la Ley N° 29903 sobre la Reforma del Sistema Privado de Pensiones (SPP), se intensificó la disputa entre las Administradoras de Fondos de Pensiones (AFP) por ganar una mayor participación en el mercado, esto con el objetivo de incrementar al máximo su stock de afiliados antes de la licitación de las afiliaciones.

En este sentido resulta clave reforzar la estrategia comercial a fin de optimizar los niveles de captación en el periodo restante.

Resumen

El presente documento, se enfoca principalmente, a identificar las empresas, con mayor potencial de afiliación al SPP en base a su stock de afiliados actual al Sistema Nacional de Pensiones (SNP) y a su ciclo de contratación histórico, para, en base a esto, determinar la captación por afiliaciones esperada en el periodo de doce meses previos a la ejecución de la ley N° 29903 que reforma al SPP.

Para alcanzar el objetivo propuesto, se optimizan los tres factores de alta relevancia para las captaciones en una AFP: la cartera de empresas, el número de asesores de venta y la distribución geográfica de sus agencias. Para lo cual, se analizan fuentes de datos de gran magnitud tomando como referencia la metodología KDD, de sus siglas en inglés *"Knowledge Discovery in Databases"*. Así, se encuentran cuáles son las actividades económicas donde se proyecta mayor número de afiliaciones y los perfiles de personas más propensos a afiliarse al SPP.

Finalmente, se comprueba que el método empleado permite estimar con mayor precisión el crecimiento del mercado para el 2012 ($R^2 = 0.947$) y que fue posible establecer un perfil de posibles migradores al SPP con un ratio de precisión del 79.8%.

ÍNDICE GENERAL

Página

ÍNDICE DE FIGURAS

Figura **Página**

ÍNDICE DE TABLAS

Cuadro **Página**

Introducción

La reforma al Sistema Privado de Pensiones (SPP), aprobada en julio del 2012, enmarcó a las AFP en un contexto difícil, ya que restringiría la potestad de realizar nuevas afiliaciones a tan sólo una AFP ganadora de la licitación bianual que establecía la ley.

Dado que previamente, el Estado ya había anticipado una reforma al SPP, las AFPs tuvieron aproximadamente, un año (octubre 2011 – septiembre 2012) para reducir el impacto económico que la reforma les representaría, entre las medidas adoptadas estaría buscar obtener la mayor cantidad de afiliaciones antes que se les suspenda esa facultad; es decir, reforzar su estrategia comercial para maximizar la captación por afiliaciones y así llegar a la licitación con el mayor stock posible de afiliados. Es en este punto en donde se enfoca el siguiente estudio.

En el Capítulo I, se resume la evolución del Sistema de Pensiones en el Perú, desde la historia del nacimiento del Sistema Nacional de Pensiones (SNP) hasta la reforma al SPP (ley N° 29903).

En el Capítulo II, se describe la metodología general del trabajo. En primer lugar, se explican los factores críticos a considerar para el reforzamiento de la estrategia comercial y cómo serán abordados. Posteriormente, se da paso al proceso KDD (*Knowledge Discovery from Databases*) el cual constituye la metodología que se utiliza para el análisis de estos datos.

En el Capítulo III, se muestra el desarrollo del proceso KDD para el análisis de perfiles de las personas que, habiéndose afiliado en un principio al SNP, migran al SPP. Así, se identificará el stock de potenciales afiliaciones, lo cual constituye un hito importante para alcanzar el objetivo propuesto.

En el Capítulo IV, se explica el desarrollo del proceso KDD, esta vez para el análisis de las actividades económicas. Esto da como resultado un estimado de las afiliaciones que se podrían obtener.

En el Capítulo V, se analizan los aspectos de una estrategia comercial que se verían afectados por la redefinición del mercado potencial de los capítulos tres y cuatro.

Finalmente, el Capítulo VI, se exponen las conclusiones a partir de los resultados obtenidos.

Capítulo 1

El sistema de pensiones

1.1 Origen del sistema de pensiones

El sistema de pensiones en el Perú tuvo como predecesores al Seguro Social Obrero Obligatorio (Ley N° 8433) y al Seguro Social del Empleado (Ley N° 13724) creados en 1936 dando origen a la seguridad social en el país, brindando protección ante los riesgos de salud y proporcionando pensión de invalidez y vejez. Durante las décadas posteriores a este suceso, hubo mejoras al sistema como el establecimiento de seguro para empleados estables, mejoras salariales, entre otros.

El primero de mayo de 1973 se creó finalmente el Sistema Nacional de Pensiones (Ley N° 19990) que incluiría por primera vez a los trabajadores independientes y agruparía el Seguro Social Obrero Obligatorio, el Seguro Social del Empleado así como también al Fondo de Jubilación de Empleados Particulares creado por el Decreto Ley 17262, con el fin de otorgar las mismas condiciones pensionarias a los diversos tipos de trabajadores.

1.2 El sistema privado de pensiones

En 1975, luego de una expansión desmesurada de los servicios sociales, se desencadenó el inicio de una crisis económica que llevaría al Perú a un periodo de desamparo social, el cual se agravaría en la segunda mitad de la década de los 80s por la inadecuada gestión pública. Debido a esto, los indicadores económicos del Perú a inicios de 1990 no eran nada alentadores y era imperativo el esfuerzo para estabilizar la situación económica del país. En este punto llegaron dos reformas principales; la primera referida a la privatización de empresas públicas y la segunda enfocada a la reestructuración del sistema de pensiones. Como respuesta a este último punto, el 6 de diciembre de 1992, por Decreto Ley N.° 25897 se crea el Sistema Privado de Pensiones (SPP) y por Decreto Legislativo 25967 se crea la Oficina de Normalización Previsional (ONP) el 19 de diciembre de 1992 con el objetivo de administrar los regímenes de pensiones del Sistema Nacional de Pensiones (SNP) y dejar al Instituto

Peruano de Seguridad Social (IPSS) a cargo exclusivamente de las prestaciones de salud, sociales y económicas.

El SPP permite a cada afiliado tener una cuenta individual, la cual se capitaliza por sus aportes y la rentabilidad producto de la administración de empresas privadas.

Por otro lado, el SNP está basado en un esquema de reparto de aportes obligatorios que no permite acumulación de fondos personales por lo que la aportación colectiva de los afiliados financia las pensiones. Por defecto los trabajadores dependientes quedarían afiliados a este sistema si es que en un plazo de 10 días no optaron por una Administradora Privada de Pensiones (AFP). De esta manera, el afiliado tenía la opción de optar por el sistema que considere más conveniente y por la empresa administradora que encuentre más eficiente, sin perder las prestaciones de salud encargadas en ese momento al IPSS.

En julio de 1995, llegaría un nuevo suceso normativo con la Ley N° 26504, el cual realizaría una serie de ajustes a la ley que se encontraba vigente y abriría la posibilidad de migrar de una AFP a otra en un proceso que se denomina "Traspaso" (ver término número 77 del anexo A).

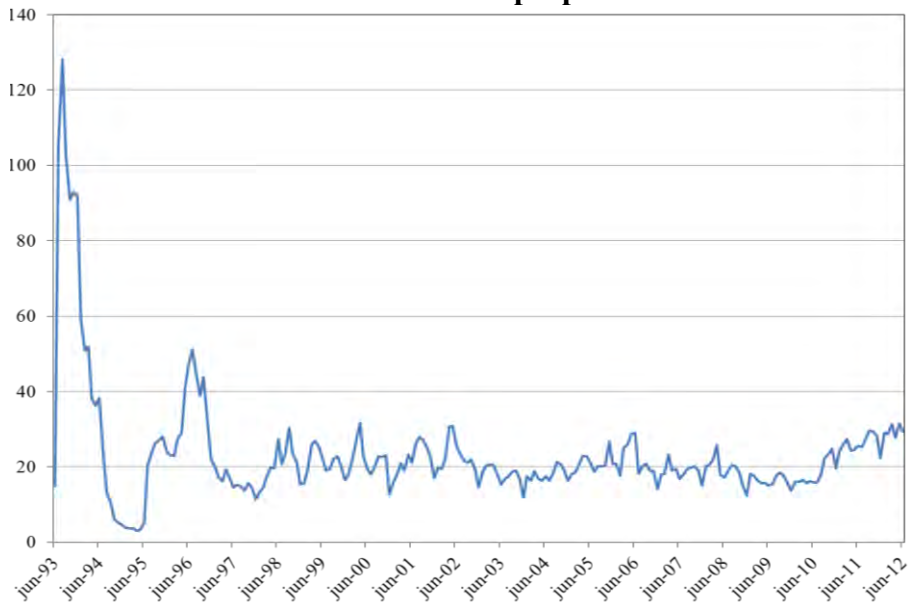
En los inicios del SPP, existieron grandes niveles de afiliación los cuales persistieron durante los 3 primeros años (ver figura 1). A partir del cuarto año, el sistema ingresa en una etapa más estable, con niveles afiliación alrededor de los 20,000 contratos. Esta situación se mantuvo estable hasta el ingreso de una nueva AFP al sistema: Prima AFP, que al iniciar sus operaciones en Agosto del 2005 causaría una revuelta en el SPP principalmente, por dos factores:

En primer lugar, dado que llegaba al mercado con una comisión considerablemente menor (1,50%) al promedio de la competencia (2.27%), el resto de AFP´s se vio obligado a bajar sus precios a niveles más competitivos.

En segundo lugar, debido a que Prima no partió con una base de afiliados, adoptó una agresiva campaña de traspasos lo que dio origen a un periodo conocido como "Guerra de Traspasos" (ver figura 2), la cual conllevaría a la asimilación de la AFP Unión Vida por Prima AFP en Diciembre del 2006, dejando a sólo 4 AFP´s en el SPP (Horizonte, Integra, Prima y Profuturo).

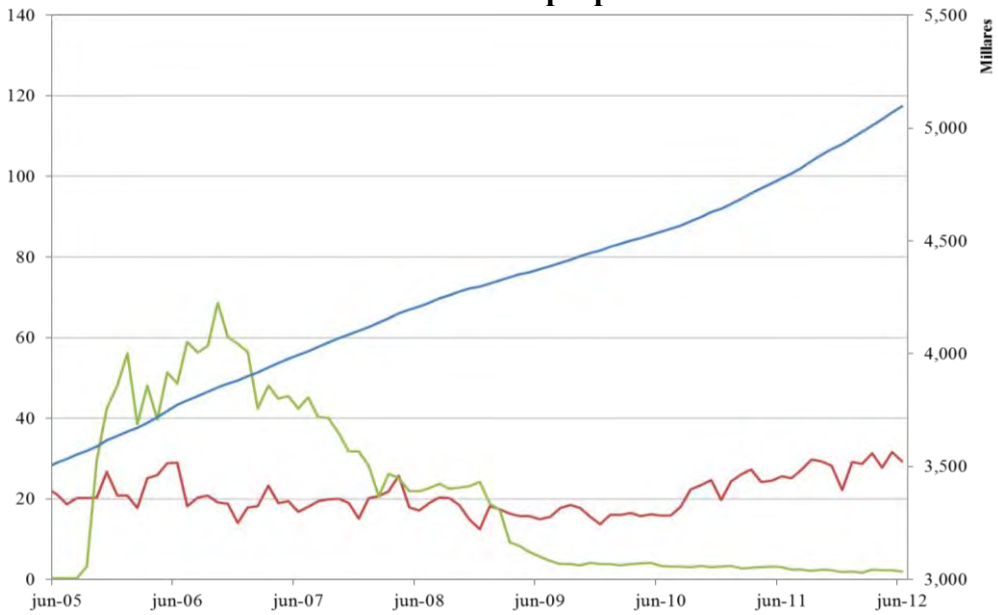
Sin embargo, la "Guerra de Traspasos" no concluyó con la caída de Unión Vida sino que ésta se mantuvo hasta mediados del año 2009 (ver figura 2) cuando la situación se tornó económicamente insostenible; debido a que solo generaba un constante desgaste entre las AFP´s, sin significar un real aumento en el stock de afiliados para cada una. En este punto, es cuando las AFPs redirigieron sus esfuerzos a expandir el sistema en su conjunto; es decir, se enfocaron a captar afiliaciones en lugar de darle prioridad a la captación por traspasos, aquí la competencia entre las AFP sería por obtener la mayor participación en el número de afiliaciones dentro del sistema.

Figura 1
Total de afiliaciones
Elaboración propia



Nota: En el eje vertical se muestra el número de afiliaciones en millares y en el eje horizontal se muestra el mes y el año al que corresponden.

Figura 2
Evolución del SPP
Elaboración propia



Nota: El eje vertical izquierdo muestra el número de ocurrencias en millares y corresponde al gráfico verde y rojo. El eje vertical derecho muestra el número de afiliados en el SPP y corresponde al gráfico azul. El eje horizontal presenta la escala

de tiempo (mes-año). El gráfico rojo representa las afiliaciones históricas del SPP y el verde los traspasos.

1.3 Reforma al sistema de pensiones (Ley Nº 29903)

Las AFP habían llegado a un periodo de estabilidad, trabajaban en conjunto para hacer crecer el sistema privado y se auto regulaban el número de traspasos mensuales que hacían para no caer en el desgaste innecesario de años anteriores. Este equilibrio no duraría mucho, las cuatro AFP, en su conjunto, empezaron a ser vistas como un gran monopolio, el cual les permitía fijar comisiones altas, además, se consideró injusto que la comisión sea independiente de la rentabilidad que las AFP generaran sobre el fondo lo que les permitía cobrar lo mismo ya sea esta rentabilidad positiva o negativa. La intervención del estado no se hizo esperar y en julio del año 2012 se aprobó la Reforma al Sistema Privado de Pensiones (ver Ley Nº 29903), la cual modificaría drásticamente, las condiciones del SPP.

Dentro de todos los cambios establecidos por la ley, se estableció que cada dos años se realizaría lo denominado "Licitación de las Afiliaciones", tras lo cual la entidad financiera ganadora de la licitación, por cobrar la menor comisión, se haría la única acreedora, siempre que la rentabilidad de los fondos que maneje no caiga debajo el promedio del mercado, de las nuevas afiliaciones del sistema por los próximos dos años, periodo luego del cual se convocaría a la próxima licitación.

La reforma afectaría considerablemente a las AFPs ya que aquellas que no ganen la licitación perderían las 120 mil afiliaciones que en promedio esperarían realizar los próximos dos años. Afortunadamente, la promulgación de esta reforma no tomó por sorpresa a las AFPs ya que desde mediados del 2011 este era un tema pendiente en la agenda del gobierno, esto les dio un margen de tiempo, aproximadamente, de un año, para optimizar sus procesos, plantearse escenarios frente a la reforma próxima y como es lógico, hacerse de una mayor cantidad de afiliaciones en el periodo previo a ésta.

1.4 Objetivos

En vista al escenario descrito y a la necesidad surgida entre las AFPs, el presente documento se enfoca en los siguientes objetivos:

General

- Identificar las empresas, con mayor potencial de afiliación al SPP en base a su stock de afiliados actual al SNP y a su ciclo de contratación histórico.

Específicos

- Determinar los perfiles de los clientes más propensos a migrar del SNP al SPP.
- Definir valores para los factores de la estrategia comercial impactados por el redimensionamiento del mercado.

Capítulo 2

Metodología de trabajo

2.1 Metodología general del estudio

Está comprendida por dos etapas. En la primera, se hará un estudio del mercado potencial mediante dos análisis complementarios, el primero, referido a la identificación del perfil de aquellas personas que han realizado la migración al SPP, lo que posteriormente, permitirá reconocer a los migradores en potencia dentro del stock de afiliados al SNP. El segundo análisis será sobre las afiliaciones históricas, a partir de las cuales, se obtendrá una estimación de las afiliaciones esperadas para cada empresa en el periodo que va desde octubre de 2011 a septiembre de 2012. Tras esta primera etapa se contará con el volumen total de afiliaciones potenciales al SPP.

En la segunda etapa, la cual corresponde directamente al impacto sobre la estrategia, se establecerá la meta, en número, de afiliaciones para una AFP de referencia y en base a esto, se determinará el tamaño del equipo de ventas necesario para cumplir ese propósito. Además, considerando el potencial de ventas y cercanía geográfica de los departamentos se establecerá el número y locación de agencias de ventas necesarias para realizar una adecuada supervisión de la labor de los vendedores asignados.

2.2 Metodología para manejo de datos

La evolución de la informática en las últimas décadas ha permitido, entre otras cosas, que los datos que las empresas solían colocar en grandes archivadores que acumulaban en cuartos depósito y luego de algunos años desechaban, pasen a ser registrados en bases de datos, las cuales, hoy en día, se han convertido en una gran fuente de datos históricos.

Teniendo en cuenta que muchas veces las decisiones empresariales se basan en resultados anteriores, que las decisiones colectivas tienen su justificación en características comunes del grupo y que se cuenta con esta información en las bases de datos, resulta lógico plantearse la posibilidad de extraer conocimiento a partir de la data con la que se cuenta. Es en este punto, durante la segunda mitad del siglo XX,

donde aparece la metodología para el descubrimiento de información a partir de bases de datos más conocida como KDD por sus siglas en inglés "*Knowledge Discovery from Databases*" la cual se compone de una serie de pasos estándar adaptables y flexibles a cualquier casuística orientados a la extracción de conocimiento que en un inicio era desconocido y podría ser de gran utilidad en la toma de decisiones tácticas y estratégicas en la empresa.

Dado el alcance del presente estudio y las extensas fuentes de información que se necesitan analizar, se utilizará el proceso KDD a nivel de guía, para llevarlo a cabo. Se debe tener en cuenta que antes de iniciar con este proceso es necesario reconocer los datos a partir de los cuales se podría generar conocimiento útil para los objetivos planteados.

2.3 El proceso KDD

Con el objetivo de interpretar los datos y encontrar patrones o relaciones significativas entre ellos, el proceso KDD se compone de una serie de fases, durante las cuales se deben tomar una serie de decisiones basadas tanto en el conocimiento del evento a estudiar como en análisis intuitivo, ambos enmarcarán un contexto específico a evaluar, esto hace de KDD un proceso iterativo producto de los diferentes escenarios planteados.

Las etapas por las que se conforma el proceso KDD son las siguientes:

2.3.1 Integración y recopilación

Tras delimitar el estudio estableciendo su alcance y objetivos, empieza la recopilación y clasificación de todas las variables (atributos) relacionadas al hecho a evaluar, las cuales pueden provenir de diversos orígenes, como bases transaccionales, archivos planos, bases de datos externas compradas a empresas privadas, etc. Es necesaria la integración de estas fuentes en un mismo repositorio de datos el cual variará su complejidad dependiendo del volumen de datos pudiendo ser desde un archivo consolidado en Excel hasta una base de datos de gran rendimiento instalada en un servidor dedicado.

En este punto cabe precisar que los datos que se analizaron para la realización del presente documento provenían de diferentes archivos de texto y de Excel que en suma reunían aproximadamente, 3 gigabytes de información. Todo este volumen de información fue almacenado en la base de datos ORACLE modelada (ver diagrama en anexo B).

2.3.2 Selección, Limpieza y Transformación

Implica la selección de las variables a partir de las cuales se espera descubrir conocimiento. Si bien, el que se produzca o no un evento puede guardar relación con un gran número variables, es importante buscar incluir en el análisis sólo aquellas más importantes, es decir; las que guarden más relación con el hecho o suceso, esto debido a que a mayor número de variables aumenta el riesgo de la obtención de patrones imprecisos y por el contrario, dificulta encontrar patrones de calidad. Dado que es posible no tener clara la importancia relativa de cada variable sobre otra, es útil formar diferentes subconjuntos de variables independientes sobre las que se

realizará el procedimiento de análisis, estos grupos de variables se proponen en base al conocimiento de negocio y a la opinión de expertos, pudiendo plantease diferentes conjuntos de datos o *data sets* a evaluar.

En este punto, se deberán definir además, las acciones a tomar frente a datos no disponibles e imprecisos así como la transformación que sufrirán los datos antes de su procesamiento por el algoritmo de minería de datos, por ejemplo, de considerar una variable de intervalo, se deberá establecer los rangos para cada subconjunto.

Es importante precisar que estos primeros pasos concentran el mayor esfuerzo del proceso KDD, y son los de mayor relevancia ya que de las variables elegidas y de la transformación que se les aplique dependerá que las etapas posteriores sean capaces de encontrar el conocimiento deseado.

2.3.3 Minería de datos

Es la fase del modelamiento, aquí se generará el modelo que permitirá la detección de patrones desconocidos, los cuales, para ser dados como válidos, deberán resultar potencialmente útiles y comprensibles. Existen diversas técnicas de minería de datos (ver anexo C), la dedición de optar por una u otra dependerá del tipo de objetivo del estudio que básicamente, puede ser de asociación, clasificación o predicción.

2.3.3.1 Asociación

Ocurre cuando se quiere encontrar si existe una alta probabilidad de que dado un evento A se produzca también un evento B; es decir, si dos hechos suelen producirse habitualmente de manera conjunta. Ejemplo: El 75% de clientes que compran pan también compran mantequilla.

Para este fin, las técnicas existentes para detección de reglas de asociación y las de agrupamiento o *clustering* son normalmente sugeridas.

2.3.3.2 Clasificación

Es la agrupación de los datos en clases según características en común, lo que permite conocer, por ejemplo, que el 85% de los clientes morosos de tarjeta de crédito tienen una relación de deuda/patrimonio mayor al 35% o que el 80% de personas entre 18 y 20 años tienen preferencia por el SPP, de este último ejemplo tenemos que la clasificación puede ser empleada para reducir el número de clases existentes ya que si se tuvieran clientes de edades que van desde 18 a 40 años, en lugar de tener 23 valores posibles para la variable edad, se podría obtener, mediante la clasificación, intervalos de edades que agrupen los 23 valores iniciales, en el ejemplo, según su preferencia por el SPP, esto restaría complejidad al modelo.

Las técnicas utilizadas para la clasificación son el c*lustering* y los árboles de decisión.

2.3.3.3 Predicción

Se estará hablando de predicción cuando el objetivo sea anticipar los futuros valores que tomará una variable, ya sea en el transcurso del tiempo o producto de la combinación de una serie de variables independientes. Por ejemplo, si se intenta pronosticar las ventas, número de clientes o las afiliaciones mensuales de una empresa para el próximo año, o, si dado cierto valor de edad, ingreso económico mensual y nivel de educación se quiere predecir si un cliente será moroso o no.

Para este caso sería necesario aplicar técnicas de regresión, redes neuronales o árboles de decisión. Las series de tiempo también son aplicables en este caso pero solo cuando se busque predecir un valor continuo en función de la evolución del tiempo.

2.3.4 Evaluación e Interpretación

En esta fase, en primer lugar, se comprueba que el modelo planteado tenga una confiabilidad aceptable. Los criterios de aceptación para cada modelo dependerán de su aplicación final, para trabajos orientados a campañas comerciales o de marketing se estila aceptar modelos que sean capaces de explicar como mínimo el comportamiento del 70% del universo. Posteriormente, habiendo aceptado el modelo es necesario interpretarlo, es aquí donde se identifican los patrones obtenidos y se seleccionan aquellos que aportan conocimiento útil.

2.3.5 Difusión y Uso

Esta fase implica hacer llegar el conocimiento generado a los posibles usuarios de la organización. Esto finalmente, traerá como consecuencia la aplicación del conocimiento en la toma de decisiones tácticas y estrategias.

Capítulo 3

Identificación de afiliados potenciales mediante el proceso KDD

3.1 Integración y recopilación.

Dado que se pretende encontrar prospectos de clientes potenciales en el stock de actuales afiliados al SNP, se tomaron en cuenta para este proceso todas aquellas variables que, según el conocimiento de negocio o la opinión de experto, influyen en la decisión de migración al SPP de una persona afiliada al SNP (ver tabla 1). En el capítulo II, en el punto referente a "Integración y recopilación de información", se hizo referencia a la construcción de la base de datos donde se almacena toda esta información. Adicionalmente, para el trabajo de este modelo se creó una nueva tabla, en esta base de datos, llamada "MODELO_PERFILES" que consolida estas variables a partir de las tablas donde previamente ya han sido cargados los datos.

Tabla 1: Principales variables para describir perfil de afiliados migradores: Tabla "Modelo_Perfiles"

N°	Campo	Campo Origen	Tabla Origen	Tipo	Descripción	MF
1	EDAD	EDAD	FACT_CLIENTES	Independiente Cualitativa	A partir de la fecha de nacimiento de la persona.	Sí
2	CATEGORIA	CATEGORIA	FACT_CLIENTES	Independiente Cualitativa	Indica la escala económica a la que pertenece la persona.	Sí
3	GENERO	GENERO	FACT_CLIENTES	Independiente Cualitativa	Sexo de la persona.	No
4	DEVENGUE	DEVENGUE	FACT_CLIENTES	Independiente Cualitativa	Fecha de actualización de la información de aportes pensionarios de la persona.	No
5	RUC	RUC	FACT_ONP	Independiente Cualitativa	Indica el ruc de la empresa donde se realizó la afiliación al SNP.	Sí
6	DISTRITO	DISTRITO	FACT_UBIGEO	Independiente Cualitativa	Distrito al que pertenece la localización geográfica de la empresa.	No
7	PROVINCIA	PROVINCIA	FACT_UBIGEO	Independiente Cualitativa	Provincia a la que pertenece la localización geográfica de la empresa.	No
8	DEPARTAMENTO	DEPARTAMENTO	FACT_UBIGEO	Independiente Cualitativa	Departamento al que pertenece la localización geográfica de la empresa.	Sí
9	AFP_AFILIACION	AFP_AFILIACION	FACT_AFP	Independiente Cualitativa	AFP a la que se afilió la persona por primera vez.	No
10	AFILIACIÓN	FEC_AFILIACION	FACT_AFP	Independiente Cualitativa	Fecha en la que la persona se afilió al SPP.	Sí
11	TIPO_FONDO_AFIL	TIPO_FONDO_AFIL	FACT_AFP	Independiente Cualitativa	Tipo de fondo en el que ingresó la persona en el momento de su afiliación al SPP.	No
12	CAT_AFIl_AFP	CAT_AFIL_AFP	FACT_AFP	Independiente Cualitativa	Indica la escala económica a la que pertenecía la persona al momento de su afiliación al SPP.	No
13	AFP_ACTUAL	AFP_ACTUAL	FACT_AFP	Independiente Cualitativa	AFP a la que se encuentra afiliada la persona en su último devengue.	No
14	TIPO_FONDO_ACTUAL	TIPO_FONDO_ACTUAL	FACT_AFP	Independiente Cualitativa	Tipo de fondo en el que se encuentra la persona en su último devengue.	No
15	MIGRACIÓN	FEC_AFILIACION	FACT_ONP	Independiente Cualitativa	Fecha en la que la persona se migró al SNP.	Sí
16	CAT_AFIL_ONP	CAT_AFIL_ONP	FACT_ONP	Independiente Cualitativa	Indica la escala económica a la que pertenecía la persona al momento de su afiliación al SNP.	No
17	SP	SP	FACT_AFP, FACT_ONP	Dependiente Dicotómica	Es la variable que se quiere explicar: toma el valor de "1" si se migró de SNP al SPP si no toma "0".	Sí

3.2 Selección, Limpieza y Transformación.

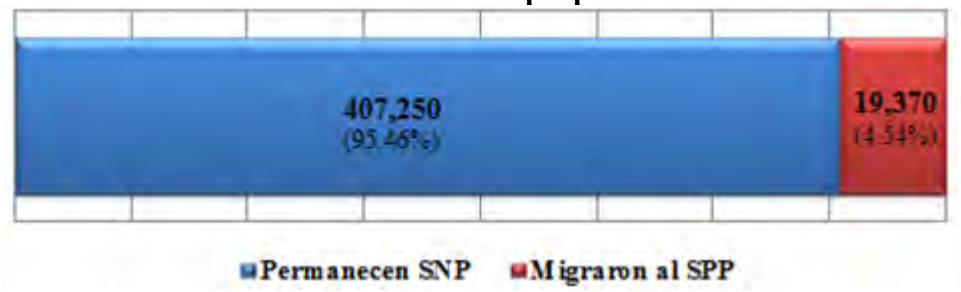
A partir del conjunto inicial de variables, mostrado en la tabla 1, se propusieron diversos subconjuntos (*data sets)* sobre los cuales se realizó de forma paralela todo el proceso KDD. A partir de este punto, se detalla el desarrollo sólo para el conjunto de variables que conforman el *data set* que finalmente, permitió generar el modelo con mayor capacidad predictiva. Estas variables son aquellas que han sido marcadas en la tabla 1 con "X" en la columna "MF".

3.2.1 Análisis de la Variable Dependiente

Se tiene la variable generada "SP" la cual funciona como una marca que asigna el valor de uno (1) a todas las personas que realizaron la migración al SPP y el valor cero (0) a los afiliados que aún permanecen en el SNP.

Se trabajó sobre un universo de 426,620 personas conformado por afiliados entre 18 y 47 años que o ya realizaron la migración al SPP o están en facultad de hacerlo. La figura 3 muestra la distribución de la variable "SP".

Figura 3
Distribución de la variable "SP"
Elaboración propia



Se espera encontrar patrones que permitan diferenciar al 4.54% que migró al SPP para identificar a los afiliados que cumpliendo con esas características no han realizado su migración, esto con el objetivo de contactarse con ellos y captarlos al SPP.

Dado que el porcentaje de los migradores es aproximadamente, 20 veces menor que el de los afiliados que permanecen en el SNP, al momento de realizar la minería de datos, el algoritmo optaría por afirmar que ningún afiliado migrará al SPP ya que esto daría una precisión del 95.46% lo cual resultaría bastante alto, pero para la finalidad del estudio no resultaría útil. Es por esto, que en casos como éste, cuando la diferencia proporcional entre la ocurrencia y no ocurrencia del evento es muy alta, es necesario realizar un balance de datos antes de iniciar con el modelamiento.

Para generar la base de datos se consideró al 100% de los afiliados que realizaron la migración, mientras que para seleccionar a las personas que permanecen en el SNP se realizó una selección aleatoria con semilla[1] 7562119 entre los registros con valor "SP"=0. Después de realizar pruebas

[1]Semilla: Valor inicial ingresado al simulador de números aleatorios.

más efectivo para forzar al algoritmo a diferenciar a los afiliados que migran al SPP de los que no lo hacen.

3.2.2 Análisis de las Variables Independientes

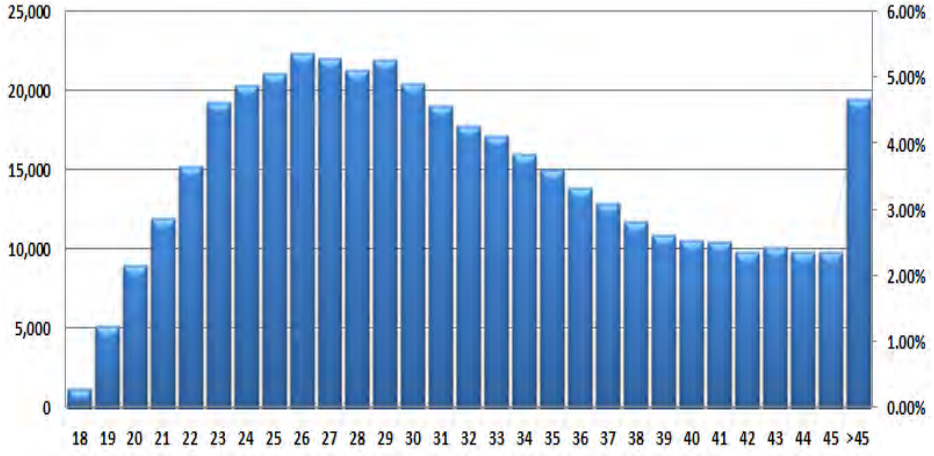
Para cada variable independiente se analizará la distribución y las frecuencias de sus categorías iniciales, a fin de adquirir mayor conocimiento sobre el comportamiento de cada variable. Posteriormente, sobre la base de datos balanceada se utilizará árboles de clasificación con la finalidad de reducir las categorías originales que presenta cada variable, asegurando que se cuenten con frecuencias suficientes en cada una de estas. Para dar como válidas las categorías obtenidas se realizará una regresión logística binomial comprobando el valor del coeficiente Beta; su error estándar; un parámetro, denominado de chi² Wald, que permite contrastar si el coeficiente es significativamente diferente de 0, la significancia de la categoría y el exponencial de los Betas. Este proceso se realizó con el software SPSS Clementine versión 11.1, en el anexo D se muestra el esquema desarrollado.

Finalmente, las nuevas categorías formadas darán paso a la generación de la variable transformada, de la cual se analizará el comportamiento dentro del universo de la población y la relación que pudiera tener con la decisión de un afiliado de migrar al SPP.

3.2.2.1 Variable Edad

En la figura 4, se muestra la distribución de la variable "Edad" donde se observa que las frecuencias más altas están entre los 26 y 29 años. En la tabla 2, se muestran los valores de frecuencia y porcentaje específicos.

Figura 4
Distribución de la variable edad
Elaboración propia



Nota: En el eje principal se muestra la frecuencia y en el secundario el porcentaje.

Tabla 2
Estadístico descriptivo de frecuencias para la variable edad.

Categoría	Frecuencia.	Porcentaje.
18	1,223	0.29%
19	5,188	1.22%
20	9,028	2.12%
21	11,972	2.81%
22	15,261	3.58%
23	19,319	4.53%
24	20,361	4.77%
25	21,093	4.94%
26	22,366	5.24%
27	22,066	5.17%
28	21,328	5.00%
29	22,009	5.16%
30	20,438	4.79%
31	19,067	4.47%
32	17,804	4.17%
33	17,244	4.04%
34	15,977	3.75%
35	15,023	3.52%
36	13,920	3.26%
37	12,901	3.02%
38	11,778	2.76%
39	10,926	2.56%
40	10,652	2.50%
41	10,484	2.46%
42	9,812	2.30%
43	10,162	2.38%
44	9,910	2.32%
45	9,824	2.30%
>45	19,484	4.57%
Total	**426,620**	**100.00%**

Las más de 20 categorías de la variable Edad se redujeron a 9 mediante el empleo de un árbol de decisión utilizando como objetivo la variable dicotómica “SP” y como variable independiente “Edad”. Estas nuevas categorías serán los valores para la variable “Edad” transformada a la que llamaremos “EDADn”. En el anexo E se puede ver el árbol formado. Los nodos obtenidos se muestran en la tabla 3.

Tabla 3
Recategorización de variable edad.

Categoría	Nodo	Descripción	Probabilidad SP="1"
1	Nodo1	<=21	68.22%
2	Nodo2	(21;24]	67.67%
3	Nodo3	(24;26]	60.20%
4	Nodo4	(26;28]	53.62%
5	Nodo5	(28;30]	47.80%
6	Nodo6	(30;32]	43.49%
7	Nodo7	(32;35]	37.18%
8	Nodo8	(35;40]	29.52%
9	Nodo9	>40	19.00%

Para verificar la validez de la transformación realizada a la variable "Edad" para explicar la variable objetivo se aplicó una regresión de tipo logístico con variable objetivo "SP" y con "Edadn" como variable dependiente.

Los resultados obtenidos en la regresión logística se observan en la tabla 4. Se observa como los primeros nodos presentan los valores de Beta y Wald más altos, lo cual corrobora el hecho de que estos son los que diferencian, en mayor medida, a las personas que migraron al SPP de los que permanecen en el SNP. Al analizar los valores de significancia de la tabla 4, se observa como la variable "EDADn" como tal, presenta un valor de significancia de 0.000 menor a 0.05 por lo cual se rechaza la hipótesis nula que los *Betas son iguales a 0* o lo que es lo mismo*:* "Se rechaza que la variable Edadn no es significativa a la hora de explicar el comportamiento de la variable dicotómica SP".

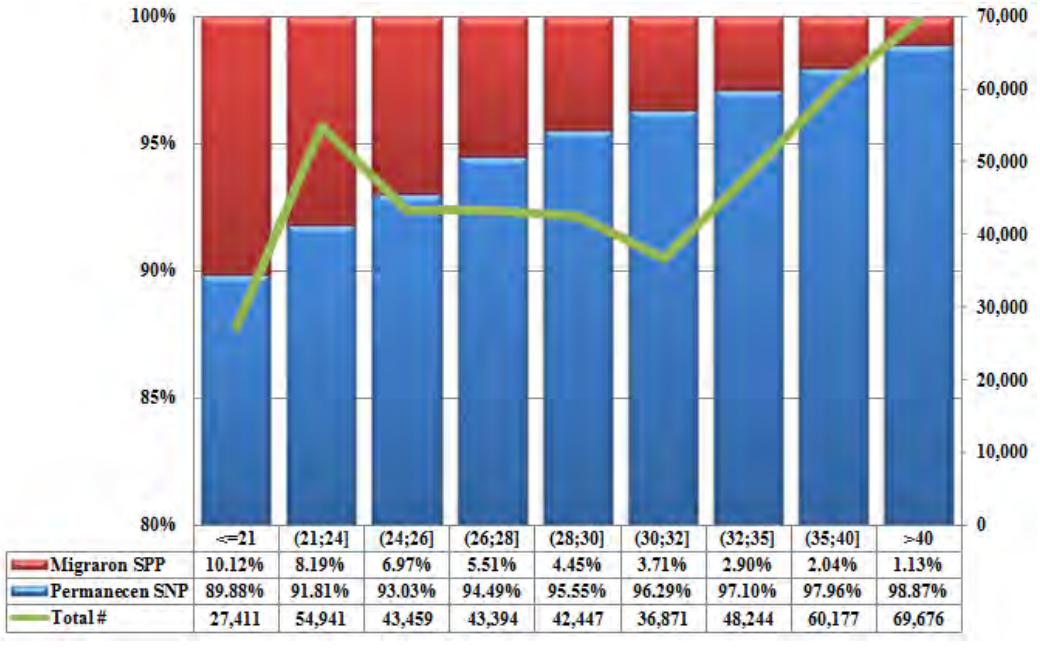
Tabla 4
Resultado de la Regresión Logística para la variable "EDADn".

Variable	B	E.T.	Wald	Gl	Sig.	Exp(B)
EDADn			3557.857	8	.000	
EDADn(1)	2.213	.052	1816.848	1	.000	9.147
EDADn(2)	2.011	.047	1859.721	1	.000	7.470
EDADn(3)	1.863	.049	1451.375	1	.000	6.445
EDADn(4)	1.595	.050	1032.056	1	.000	4.928
EDADn(5)	1.362	.051	719.856	1	.000	3.903
EDADn(6)	1.188	.053	493.923	1	.000	3.281
EDADn(7)	0.925	.052	316.901	1	.000	2.523
EDADn(8)	0.580	.052	123.618	1	.000	1.785
Constante	-1.450	.040	1344.848	1	.000	0.235

Nota: donde B indica los valores Betas; E.T, el error típico; Wald, el valor del estadístico de Wald; Gl, los grados de libertad; Sig, la significancia de la categoría y Exp(B), el exponencial de los Betas.

En el figura 5, se muestra, en el eje principal, cómo influye el rango de edad en la decisión de migrar al SPP. Notamos que a menor edad es mayor el porcentaje de migradores. Por otro lado, en el eje secundario, tenemos cómo se distribuye la variable "EDADn" en las nueve categorías obtenidas, de aquí observamos que el grupo de personas entre 22 y 24 años además de tener un porcentaje importante de migradores, cuenta con una alta participación dentro de la población en estudio.

Figura 5
Comportamiento de la variable EDADn.
Elaboración propia



3.2.2.2 Variable Ruc

Se considera que el lugar de trabajo puede tener relación con la decisión de migrar o no al SPP porque se tiene la hipótesis que las empresas que realizan actividades económicas relacionadas al Estado tendrían una mayor vinculación con el SNP. Para conocer la actividad económica que realiza cada empresa fue necesario obtener de la página web de la SUNAT su código CIIU (Clasificación Internación Industrial Uniforme) el cual agrupa a las empresas en secciones y a su vez sub agrupa estas secciones en divisiones y finalmente, en giros económicos (ver anexo F), esto permitió la generación de tres variables más: "Sección", "División" y "Giro". Dado que las tres variables describen la actividad económica de una empresa sólo que a diferente nivel de detalle, resultaría redundante utilizar simultáneamente las tres, debido a esto se optó por utilizar la variable "Sección", la cual agrupa de manera más general, con solo 21 clasificaciones, la actividad económica de las empresas.

3.2.2.3 Variable Sección

En primer lugar, se codifican las secciones para mayor facilidad en el análisis. Las empresas a analizar se reparten en sólo 17 secciones.

Tabla 5
Codificación de las secciones

Sección	**Código**
Agricultura, ganadería, caza y silvicultura	S1
Explotación de minas y canteras	S2
Pesca	S3
Industrias manufactureras	S4
Suministros de electricidad, gas y agua	S5
Construcción	S6
Comercio al por mayor y menor	S7
Hoteles y restaurantes	S8
Transporte, almacenamiento y comunicaciones	S9
Intermediación financiera	S10
Actividades inmobiliarias, empresariales y de alquiler	S11
Administración pública y defensa de planes de seguridad social de afiliación obligatoria	S12
Enseñanza	S13
Servicios sociales y salud	S14
Otras actividades de servicios comunitarios, sociales y personales	S15
Hogares privados con servicios domésticos	S16
Organizaciones y órganos extraterritoriales	S17

En la figura 6 se muestra la distribución de la variable "Sección" donde se observa que las secciones dominantes son el S4, S11 y S12 concentrando aproximadamente el 60% de la población a analizar. En la tabla 6 se muestran los valores de frecuencia y porcentaje específicos.

Figura 6
Distribución de la variable "Sección"
Elaboración propia

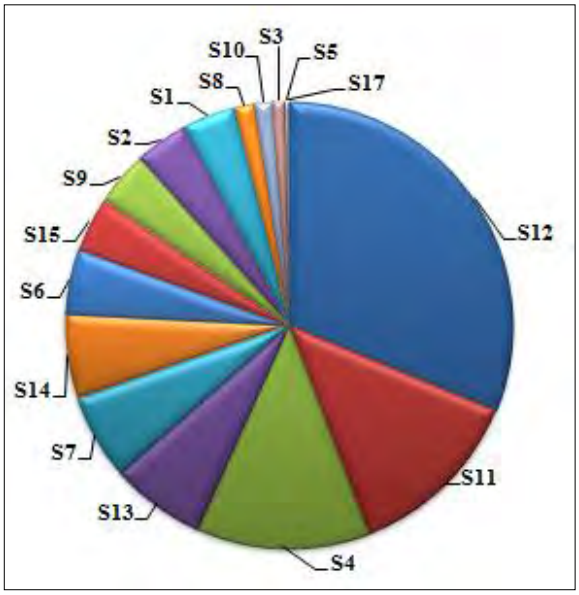


Tabla 6
Estadístico descriptivo de frecuencias para la variable sección

Categoría	Frecuencia.	Porcentaje.
S12	107,930	25.30%
S11	65,161	15.27%
S4	59,222	13.88%
S7	31,649	7.42%
S14	25,314	5.93%
S13	24,449	5.73%
S6	22,070	5.17%
S15	18,807	4.41%
S1	18,091	4.24%
S2	17,831	4.18%
S9	17,167	4.02%
S8	7,759	1.82%
S10	6,365	1.49%
S3	3,924	0.92%
S5	763	0.18%
S17	118	0.03%
Total	**426,620**	**100.00%**

Las 16 categorías de la variable Sección se redujeron a 11 mediante el empleo de un árbol de decisión utilizando como objetivo la variable dicotómica “SP” y como variable independiente “Sección”. Estas nuevas categorías serán los valores para la variable “Sección” transformada a la que llamaremos “SECCIONn”. En el anexo E se puede ver el árbol formado. Los nodos obtenidos se muestran en la tabla 7.

Tabla 7
Recategorización de variable sección

Categoría	Nodo	Descripción	Probabilidad SP="1"
1	Nodo1	S1	68.42%
2	Nodo2	S10	80.00%
3	Nodo3	S11	55.99%
4	Nodo4	S12	31.41%
5	Nodo5	S13	26.65%
6	Nodo6	S14	17.92%
7	Nodo7	S15; S17; S3	51.75%
8	Nodo8	S2	42.64%
9	Nodo9	S4; S9	47.77%
10	Nodo10	S5; S6; S8	62.05%
11	Nodo11	S7	58.69%

Para la validación se realizó una regresión logística binaria con variable objetivo “SP” y con “SECCIONn” como variable dependiente. Los resultados se muestran en la tabla 8 en donde se observa como los dos primeros nodos , por tener los valores de Beta y Wald más altos, son los que diferencian, en mayor medida, a las personas que migraron al SPP de los que permanecen en el SNP.

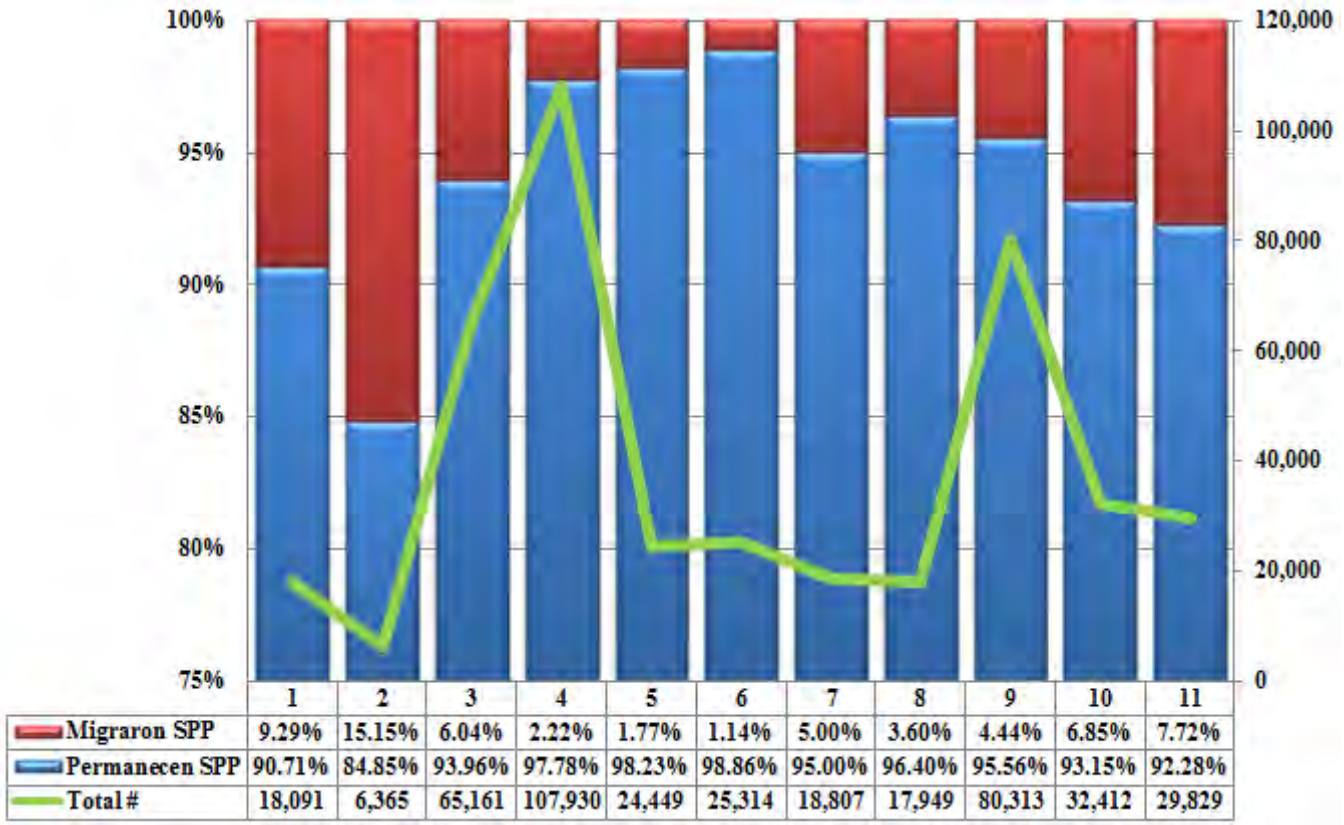
Tabla 8
Resultado de la regresión logística para la variable “SECCIONn”

	B	E.T.	Wald	Gl	Sig.	Exp(B)
SECCIONn			3027.940	10	.000	
SECCIONn(1)	0.422	.055	59.398	1	.000	1.525
SECCIONn(2)	1.035	.079	170.106	1	.000	2.816
SECCIONn(3)	-0.110	.041	7.192	1	.007	0.896
SECCIONn(4)	-1.132	.041	744.709	1	.000	0.322
SECCIONn(5)	-1.364	.065	435.549	1	.000	0.256
SECCIONn(6)	-1.873	.073	658.445	1	.000	0.154
SECCIONn(7)	-0.281	.054	26.975	1	.000	0.755
SECCIONn(8)	-0.648	.062	109.349	1	.000	0.523
SECCIONn(9)	-0.441	.041	115.415	1	.000	0.644
SECCIONn(10)	0.141	.047	8.843	1	.003	1.151
Constante	0.351	.033	110.767	1	.000	1.421

Nota: donde B indica los valores Betas; E.T, el error típico; Wald, el valor del estadístico de Wald; Gl, los grados de libertad; Sig, la significancia de la categoría y Exp(B), el exponencial de los Betas.

En la figura 7 se muestra, en el eje principal, cómo influye cada categoría de la variable "SECCIONn" en la decisión de migrar al SPP. Notamos que la categoría 2 presenta el mayor porcentaje de migradores y que, por el contrario, las categorías 4, 5 y 6 tienen el porcentaje más bajo de afiliados al SNP que decidieron pasarse al SPP. Por otro lado, en el eje secundario, tenemos cómo se distribuye la variable "SECCIONn" en las once categorías obtenidas, de aquí observamos que el grupo de personas pertenecientes a las categorías 4 y 9 concentran con una alta participación dentro de la población en estudio.

Figura 7
Comportamiento de la variable SECCIONn.
Elaboración propia



	1	2	3	4	5	6	7	8	9	10	11
Migraron SPP	9.29%	15.15%	6.04%	2.22%	1.77%	1.14%	5.00%	3.60%	4.44%	6.85%	7.72%
Permanecen SPP	90.71%	84.85%	93.96%	97.78%	98.23%	98.86%	95.00%	96.40%	95.56%	93.15%	92.28%
Total #	18,091	6,365	65,161	107,930	24,449	25,314	18,807	17,949	80,313	32,412	29,829

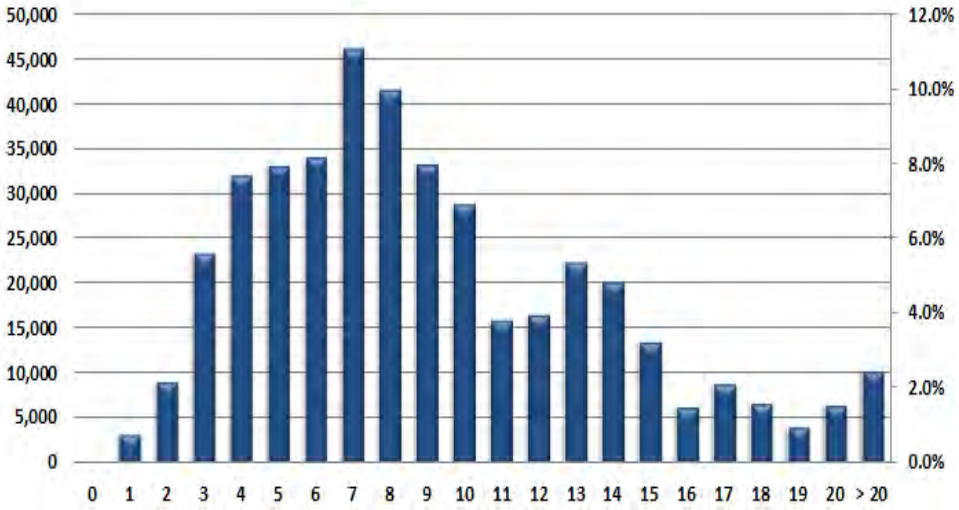
3.2.2.4 Variables Afiliación y Migración

Para el caso de las variables "Afiliación" (fecha de afiliación al SNP) y "Migración" (fecha de afiliación al SPP, si hubiera) se consideró que por sí solas no daban una información relevante por lo que se transformaron creando la variable "Años_SNP", la cual será analizada. Está variable cuantitativa indica la diferencia en años entre la fecha de migración al SPP y afiliación al SNP, o en su defecto, para aquellos que no hayan realizado la migración al SPP, el tiempo transcurrido desde la fecha de su afiliación al SNP.

3.2.2.5 Variable Años_SNP

En la figura 8 se muestra la distribución de la variable "Años_SNP" donde se observa que el promedio de antigüedad en el SNP es de 9 años aproximadamente, en la tabla 9 se muestran los valores de frecuencia y porcentaje específicos.

Figura 8
Distribución de la variable Años_SNP.
Elaboración propia



Nota: En el eje principal tenemos la frecuencia y en el secundario el porcentaje.

Tabla 9
Estadístico descriptivo de frecuencias para la variable Años_SNP

Categoría	Frecuencia.	Porcentaje.
0	84	0.02%
1	3,320	0.78%
2	9,346	2.19%
3	23,902	5.60%
4	32,869	7.70%
5	33,966	7.96%
6	34,873	8.17%
7	47,290	11.08%
8	42,613	9.99%
9	34,011	7.97%
10	29,512	6.92%
11	16,294	3.82%
12	16,915	3.96%
13	23,012	5.39%
14	20,741	4.86%
15	13,884	3.25%
16	6,377	1.49%
17	9,043	2.12%
18	6,927	1.62%
19	4,153	0.97%
20	6,635	1.56%
> 20	10,853	2.50%
Total	**426,620**	**100.00%**

Las más de 20 categorías de la variable Años_SNP se redujeron a 8 mediante el empleo de un árbol de decisión utilizando como objetivo la variable dicotómica "SP" y como variable independiente "Años_SNP". Estas nuevas categorías serán los valores para la variable "Años_SNP" transformada a la que llamaremos "AÑOS_SNPn". En el anexo E se puede ver el árbol formado. Los nodos obtenidos se muestran en la tabla 10.

Tabla 10
Recategorización de variable Años_SNP.

Categoría	Nodo	Descripción	Probabilidad SP="1"
1	Nodo1	<=2	75.86%
2	Nodo2	(2;4]	65.75%
3	Nodo3	(4;6]	59.87%
4	Nodo4	(6;7]	50.60%
5	Nodo5	(7;9]	38.55%
6	Nodo6	(9;11]	35.22%
7	Nodo7	(11;14]	25.66%
8	Nodo8	>14	22.19%

Para la validación se realizó una regresión logística binaria con variable objetivo "SP" y con "AÑOS_SNPn" como variable dependiente. Los resultados se muestran en la tabla 11 en donde se observa como los cuatro primeros nodos son los que diferencian, en mayor medida, a las personas que migraron al SPP de los que permanecen en el SNP.

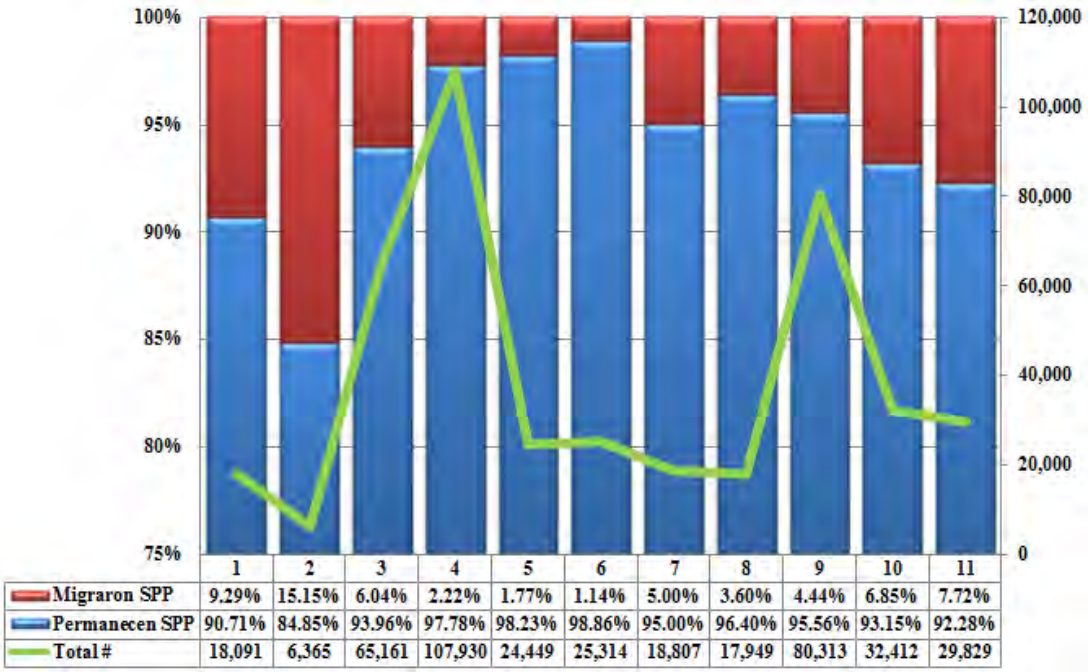
Tabla 11
Resultado de la regresión logística para la variable "AÑOS_SNPn".

Variable	B	E.T.	Wald	Gl	Sig.	Exp(B)
AÑOS_SNPn			4068.942	7	.000	
AÑOS_SNPn(1)	2.399	.063	1471.220	1	.000	11.017
AÑOS_SNPn(2)	1.906	.047	1653.851	1	.000	6.728
AÑOS_SNPn(3)	1.655	.046	1282.612	1	.000	5.231
AÑOS_SNPn(4)	1.278	.050	651.713	1	.000	3.591
AÑOS_SNPn(5)	0.788	.048	267.773	1	.000	2.199
AÑOS_SNPn(6)	0.645	.054	142.438	1	.000	1.906
AÑOS_SNPn(7)	0.191	.054	12.578	1	.000	1.210
Constante	-1.254	.040	970.835	1	.000	.285

Nota: donde B indica los valores Betas; E.T, el error típico; Wald, el valor del estadístico de Wald; Gl, los grados de libertad; Sig, la significancia de la categoría y Exp(B), el exponencial de los Betas.

En la figura 9 se muestra, en el eje principal, como influye cada categoría de la variable "AÑOS_SNPn" en la decisión de migrar al SPP. Notamos que a menor tiempo en la ONP es mayor el porcentaje de migradores. Por otro lado, en el eje secundario, tenemos como se distribuye la variable "AÑOS_SNPn" en las nueve categorías obtenidas, de aquí observamos que el grupo de personas con antigüedad entre 5 y 9 años cuenta con una alta participación dentro de la población en estudio.

Figura 9
Comportamiento de la variable AÑOS_SNPn.
Elaboración propia

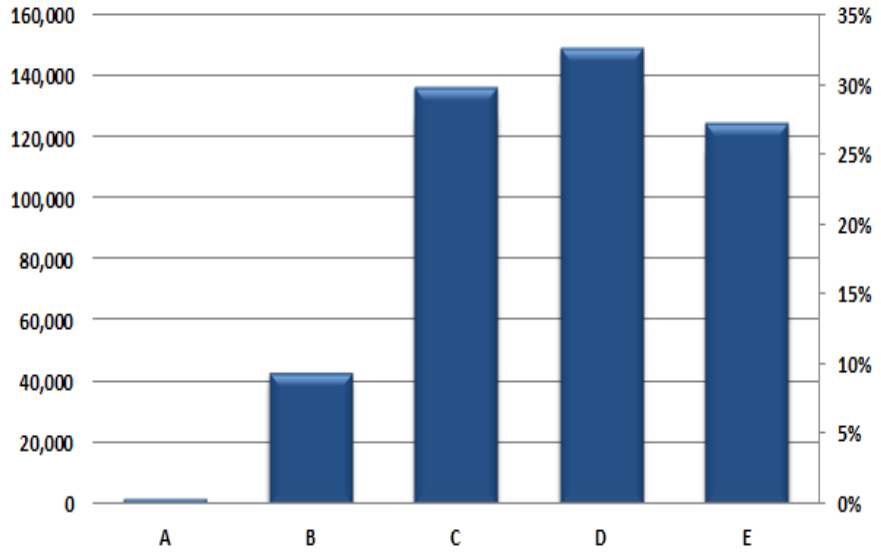


	1	2	3	4	5	6	7	8	9	10	11
Migraron SPP	9.29%	15.15%	6.04%	2.22%	1.77%	1.14%	5.00%	3.60%	4.44%	6.85%	7.72%
Permanecen SPP	90.71%	84.85%	93.96%	97.78%	98.23%	98.86%	95.00%	96.40%	95.56%	93.15%	92.28%
Total #	18,091	6,365	65,161	107,930	24,449	25,314	18,807	17,949	80,313	32,412	29,829

3.2.2.6 Variable categoría

La variable "Categoría" representa el salario promedio de cada individuo de la población (A=S./10,622; B=S./ 5,126; C=S./ 3,261; D=S./ 1,992; E=S./ 1,027), estos valores se tomaron del estudio sobre niveles socioeconómicos realizado por IPSOS Apoyo en el año 2011. En la figura 10 se muestra la distribución donde se observa que la población a analizar se concentra en las escalas económicas C y D. En la tabla 12 se muestran los valores de frecuencia y porcentaje específicos.

Figura 10
Distribución de la variable categoría
Elaboración propia



Nota: En el eje principal tenemos la frecuencia y en el secundario el porcentaje.

Tabla 12
Estadístico descriptivo de frecuencias para la variable categoría

Categoría	Frecuencia.	Porcentaje.
A	2,016	0.47%
B	40,707	9.54%
C	127,589	29.91%
D	139,602	32.72%
E	116,706	27.36%
Total	**426,620**	**100.00%**

Las 5 categorías de la variable se redujeron a 4 mediante el empleo de un árbol de decisión utilizando como objetivo la variable dicotómica "SP" y como variable independiente "Categoría". Estas nuevas categorías serán los valores para una nueva variable que nombraremos "CATEGORIAn". En el anexo E se puede ver el árbol formado. Los nodos obtenidos se muestran en la tabla 13.

Tabla 13
Recategorización de variable categoría

Categoría	Nodo	Descripción	Probabilidad SP="1"
1	Nodo1	A; B	40.96%
2	Nodo2	C	35.98%
3	Nodo3	D	32.75%
4	Nodo4	E	67.54%

Para la validación se realizó una regresión logística binaria con variable objetivo "SP" y con "CATEGORIAn" como variable dependiente. Los resultados se muestran en la tabla 14.

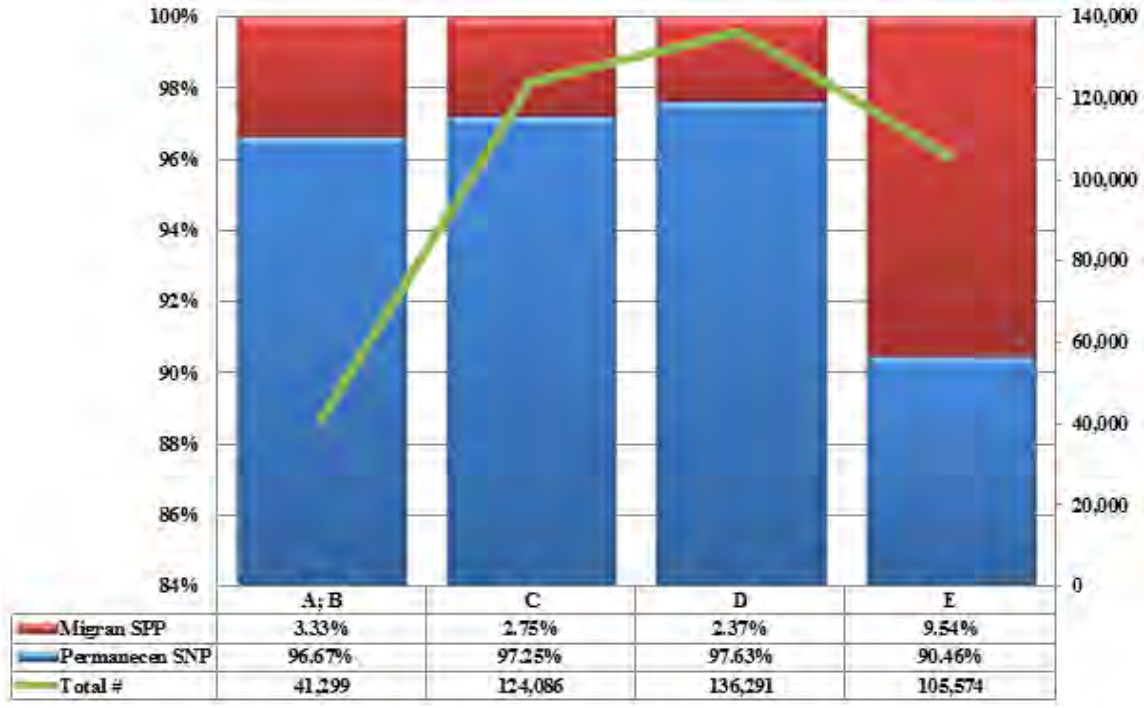
Tabla 14
Resultado de la regresión logística para la variable "CATEGORIAn".

Variable	B	E.T.	Wald	Gl	Sig.	Exp(B)
CATEGORIA			3913.862	1	.000	
CATEGORIAn(1)	-1.098	.038	822.906	1	.000	0.333
CATEGORIAn(2)	-1.309	.027	2369.744	1	.000	0.270
CATEGORIAn(3)	-1.452	.027	2904.324	1	.000	0.234
Constante	0.733	.017	1939.256	1	.000	2.080

Nota: donde B indica los valores Betas; E.T, el error típico; Wald, el valor del estadístico de Wald; Gl, los grados de libertad; Sig, la significancia de la categoría y Exp(B), el exponencial de los Betas.

En la figura 11 se muestra, en el eje principal, cómo influye cada categoría de la variable "CATEGORIAn" en la decisión de migrar al SPP. Notamos que las menores escalas salariales ("C", "D" y "E") poseen mayor porcentaje de migradores. Por otro lado, en el eje secundario, tenemos cómo se distribuye la variable "CATEGORIAn" en las cuatro categorías obtenidas, de aquí observamos que el grupo de personas en las escalas salariales "C" y "D" concentran una alta participación dentro de la población en estudio.

Figura 11
Comportamiento de la variable CATEGORIAn.
Elaboración propia



	A; B	C	D	E
Migran SPP	3.33%	2.75%	2.37%	9.54%
Permanecen SNP	96.67%	97.25%	97.63%	90.46%
Total #	41,299	124,086	136,291	105,574

3.2.2.7 Variable Departamento

Se consideró importante la variable "Departamento" porque puede estar relacionada al nivel cultural de la zona lo que se piensa influye en la decisión de migración, además, existen actividades económicas, como la agricultura, que no son uniformes en todo el Perú sino constituyen una realidad diferente en cada zona o departamento. Para iniciar el análisis, en primer lugar se codifican los departamentos para mayor facilidad en el análisis.

Tabla 15
Codificación de los departamentos

Departamento	Código
AMAZONAS	DEP1
ANCASH	DEP2
APURIMAC	DEP3
AREQUIPA	DEP4
AYACUCHO	DEP5
CAJAMARCA	DEP6
CALLAO	DEP7
CUZCO	DEP8
HUANCAVELICA	DEP9
HUANUCO	DEP10
ICA	DEP11
JUNIN	DEP12
LA LIBERTAD	DEP13
LAMBAYEQUE	DEP14
LIMA	DEP15
LORETO	DEP16
MADRE DE DIOS	DEP17
MOQUEGUA	DEP18
PASCO	DEP19
PIURA	DEP20
PUNO	DEP21
SAN MARTIN	DEP22
TACNA	DEP23
TUMBES	DEP24
UCAYALI	DEP25

En la figura12 se muestra la participación de cada departamento en el universo de individuos a analizar. Se observa que en el departamento de Lima (DEP15) se concentra la población, otros departamentos con participación importante son Arequipa (DEP4), la provincia constitucional del Callao (DEP7) y La Libertad (DEP13). Entre estos cuatro se reúne aproximadamente, el 74% de la muestra para el análisis. En la tabla 16 se muestran los valores de frecuencia y porcentaje específicos.

Figura 12
Distribución de la variable "Departamento"
Elaboración propia

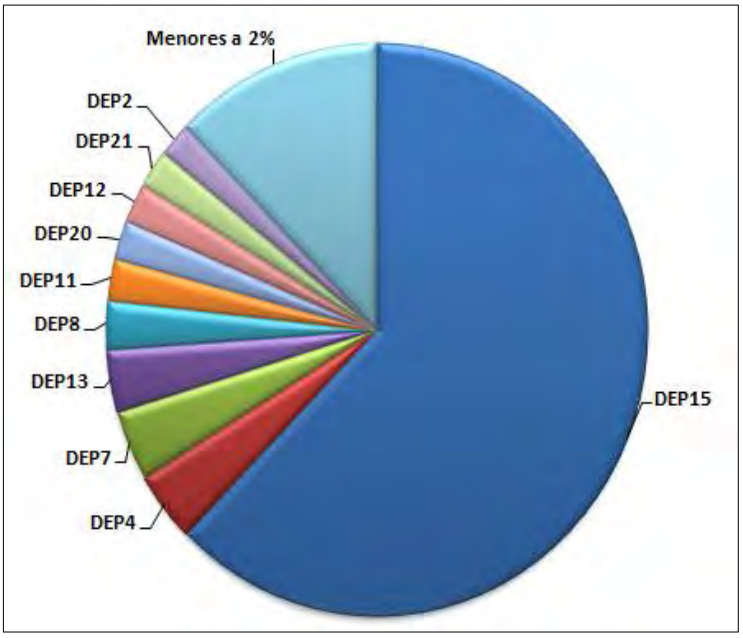


Tabla 16
Estadístico descriptivo de frecuencias para la variable "Departamento"

Categoría	Frecuencia.	Porcentaje.
DEP1	1,665	0.39%
DEP2	8,398	1.97%
DEP3	3,510	0.82%
DEP4	17,074	4.00%
DEP5	4632	1.09%
DEP6	6,846	1.60%
DEP7	16,938	3.97%
DEP8	11,588	2.72%
DEP9	5,280	1.24%
DEP10	4,425	1.04%
DEP11	9,901	2.32%
DEP12	9,457	2.22%
DEP13	15,139	3.55%
DEP14	6,629	1.55%
DEP15	266,252	62.41%
DEP16	4,100	0.96%
DEP17	706	0.17%
DEP18	2,437	0.57%
DEP19	2,620	0.61%
DEP20	9,472	2.22%
DEP21	9,237	2.17%
DEP22	3,545	0.83%
DEP23	3,051	0.72%
DEP24	1,165	0.27%
DEP25	2,553	0.60%
Total	**426,620**	**100.00%**

Las 25 categorías de la variable Sección se redujeron a 8 mediante el empleo de un árbol de decisión utilizando como objetivo la variable dicotómica "SP" y como variable independiente "Departamento". Estas nuevas categorías serán los valores para la variable "Departamento" transformada a la que llamaremos "DEPARTAMENTOn". En el anexo E se puede ver el árbol formado. Los nodos obtenidos se muestran en la tabla 17.

Tabla 17
Recategorización de variable "Departamento"

Categoría	Nodo	Descripción	Probabilidad SP="1"
1	Nodo1	DEP1; DEP2; DEP5; DEP9	19.50%
2	Nodo2	DEP10; DEP12; DEP19	30.09%
3	Nodo3	DEP11; DEP15; DEP24	51.36%
4	Nodo4	DEP13; DEP22; DEP7	48.49%
5	Nodo5	DEP14; DEP6	38.30%
6	Nodo6	DEP16; DEP17; DEP 25; DEP8	60.63%
7	Nodo7	DEP18; DEP20; DEP23; DEP4	45.55%
8	Nodo8	DEP21; DEP3	24.60%

Para la validación se realizó una regresión logística binaria con variable objetivo "SP" y con "DEPARTAMENTOn" como variable dependiente. Los resultados se muestran en la tabla 18 en donde se observa como los nodos tres y seis son los que diferencian, en mayor medida, a las personas que migraron al SPP de los que permanecen en el SNP.

Tabla 18
Resultado de la regresión logística para la variable "DEPARTAMENTOn"

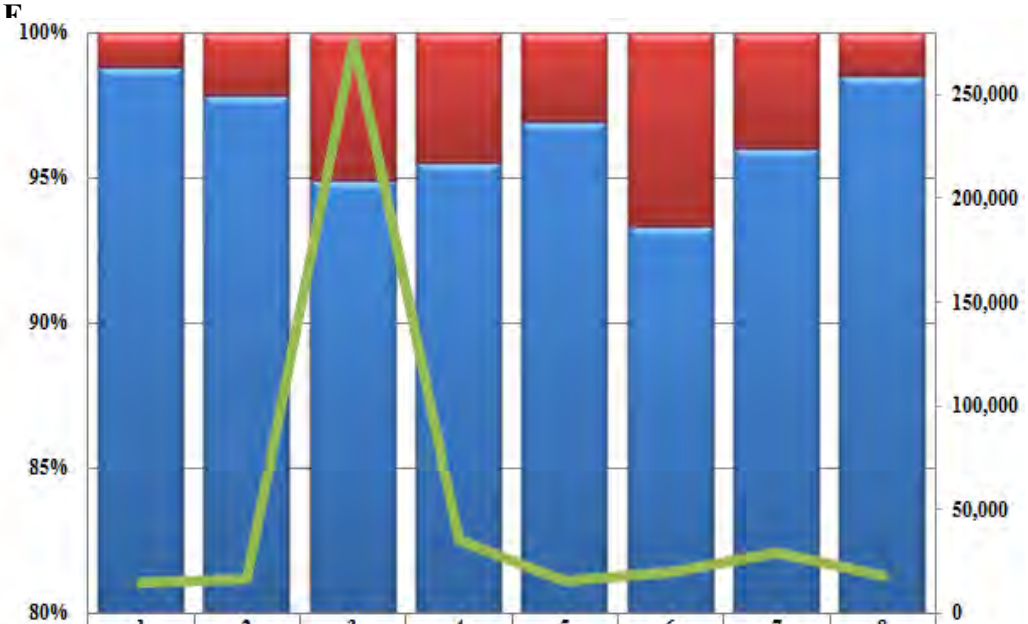
Variable	B	E.T.	Wald	gl	Sig.	Exp(B)
DEPARTAMENTOn			950.533	7	.000	
DEPARTAMENTOn(1)	-0.298	.108	7.608	1	.006	0.742
DEPARTAMENTOn(2)	0.277	.103	7.218	1	.007	1.319
DEPARTAMENTOn(3)	1.175	.083	202.545	1	.000	3.237
DEPARTAMENTOn(4)	1.060	.089	142.527	1	.000	2.885
DEPARTAMENTOn(5)	0.643	.104	38.462	1	.000	1.903
DEPARTAMENTOn(6)	1.552	.093	278.457	1	.000	4.720
DEPARTAMENTOn(7)	0.942	.090	109.356	1	.000	2.565
Constante	-1.120	.082	188.274	1	.000	0.326

Nota: donde B indica los valores Betas; E.T, el error típico; Wald, el valor del estadístico de Wald; Gl, los grados de libertad; Sig, la significancia de la categoría y Exp(B), el exponencial de los Betas.

En la figura 11 se muestra, en el eje principal, como influye cada categoría de la variable "DEPARTAMENTOn" en la decisión de migrar al SPP. Notamos que la categoría 6 presenta el mayor porcentaje de migradores y que, por el contrario, las categorías 1 y 8 tienen el porcentaje más bajo de afiliados al SNP que decidieron pasarse al SPP. Por otro lado, en el eje secundario, tenemos cómo se distribuye la variable "DEPARTAMENTOn" en las ocho categorías

obtenidas, de aquí observamos que el grupo de personas pertenecientes a la categoría 3 concentran con una alta participación dentro de la población en estudio.

Figura 13
Comportamiento de la variable DEPARTAMENTOn



	1	2	3	4	5	6	7	8
Migran SPP	1.20%	2.18%	5.05%	4.49%	3.05%	6.67%	3.99%	1.51%
Permanecen SNP	98.80%	97.82%	94.95%	95.51%	96.95%	93.33%	96.01%	98.49%
Total #	14,695	16,502	276,153	35,622	15,912	20,112	29,597	18,027

3.3 Minería de datos.

En este punto, producto de transformaciones aplicadas a las variables seleccionadas inicialmente, se tienen cinco variables independientes para la construcción del modelo de minería de datos: EDADn, SECCIONn, AÑOS_SNPn, CATEGORIAn y DEPARTAMENTOn. El objetivo es anticipar la decisión de un afiliado al SNP de migrar o no al SPP; es decir, la predicción de un evento futuro. Para este fin, sobre la base de datos balanceada, se generaron paralelamente, tres modelos. El primero, utilizando técnicas de árboles de decisión, el segundo aplicando redes neuronales y el tercero empleando regresión logística. El modelamiento fue un proceso iterativo en busca del algoritmo y de la configuración de parámetros que dotara al modelo con una predicción más precisa para cada una de las técnicas seleccionadas. A continuación, se detallará la configuración de cada modelo y se evaluará su capacidad predictiva mediante el análisis de su matriz de confusión y el área bajo la curva ROC (ver anexo G). El esquema de los modelos desarrollados se muestran en el anexo H.

3.3.1 Modelo 1: Árbol de decisión

Se realizaron varios árboles de decisión para la variable dependiente “SP” utilizando diferentes algoritmos (C&R, QUEST, CHAID y CHAID Exhaustivo) descritos en el anexo C. En la tabla 19, se muestra el mejor resultado obtenido para cada uno de ellos.

Tabla 19
Resumen de resultados para algoritmos de árbol de decisión

Algoritmo empleado	Precisión Global	Área bajo la curva ROC
CHAID Exhaustivo	74.96%	.8267
CHAID	74.92%	.8257
QUEST	71.58%	.7571
C&R	71.03%	.7497

El algoritmo CHAID Exhaustivo se mostró superior al momento de predecir la variable objetivo al tener una precisión global y la mayor área bajo la curva ROC. En la tabla 20, se muestran los parámetros de configuración para este modelo.

Tabla 20
Parámetros del modelo 1

Parámetro	Valor
Alfa para división	.05
Épsilon para convergencia	.001
Niveles por debajo de la raíz	5
Número mínimo de registros en rama parental	25
Número mínimo de registros en rama parental	10
Niveles por debajo de la raíz	5

En la tabla 21, se muestra la matriz de confusión del modelo planteado.

Tabla 21
Matriz de confusión del modelo 1

			Pronosticado	
			SP	
			0	1
Real	SP	0	14,728	5,706
		1	4,262	15,108

A partir de la matriz generada podemos concluir lo siguiente:

- La precisión global del modelo es de 74.96%.
- Su precisión para predecir eventos positivos es de 78.00%.
- Su precisión para predecir eventos negativos es de 72.08%.
- Finalmente, el coeficiente de Dice es 0,75. Esto nos indica que el modelo puede darse por válido.

3.3.2 Modelo 2: Red neuronal

Se realizaron diferentes modelos de redes neuronales utilizando diversos métodos (Rápido, Dinámico, Múltiple, Poda Exhaustiva y RBFN) descritos en el anexo C. En la tabla 22, se muestra el mejor resultado obtenido para cada uno de ellos.

Tabla 22
Resumen de resultados para algoritmos de redes neuronales

Algoritmo empleado	Precisión Global	Área bajo la curva ROC
PODA	75.44%	.8274
MULTIPLE	74.43%	.8203
RAPIDO	74.45%	.8175
DINAMICO	73.76%	.7828
RFBN	73.76%	.7828

El método de PODA posee una mejor precisión global y una mayor área bajo la curva ROC lo que lo convierte en el método idóneo a utilizar. Se configuró la poda para un análisis exhaustivo de los datos, esta modalidad no admite parámetros de configuración. En la tabla 23, se tiene la matriz de confusión del modelo planteado:

Tabla 23
Matriz de confusión del modelo 2

			Pronosticado	
			SP	
			0	1
Real	SP	0	14,572	5,862
		1	3,912	15,458

A partir de la matriz generada podemos concluir lo siguiente:

- La precisión global del modelo es de 75.44%.
- Su precisión para predecir eventos positivos es de 79.80%.
- Su precisión para predecir eventos negativos es de 71.31%.
- Finalmente, el coeficiente de Dice es 0,7598. Esto nos indica que el modelo puede darse por válido.

3.3.3 Modelo 3: Regresión logística

Se realizaron diversos modelamientos. Los métodos empleados, descritos en el anexo C, fueron: Introducir, Por pasos, Adelante, Por pasos hacia atrás. Todos estos métodos proporcionaron un precisión global de 72.32% y un área bajo la curva ROC de 0,79. En la tabla 24 se muestran los parámetros de configuración que se utilizaron:

Tabla 24
Parámetros del modelo 3

Parámetro	Valor
Tipo de modelo	Efectos Principales
Máxima subdivisión por pasos	5
Convergencia de los parámetros	1.0E-6
Tolerancia para la singularidad	1.0E-8

En la tabla 25 se muestra la matriz de confusión del modelo planteado:

Tabla 25
Matriz de confusión del modelo 1

			Pronosticado	
			SP	
			0	1
Real	SP	0	15,106	5,328
		1	5,688	13,682

A partir de la matriz generada podemos concluir lo siguiente:

- La precisión global del modelo es de 72.32%.
- Su precisión para predecir eventos positivos es de 70.63%.
- Su precisión para predecir eventos negativos es de 73.93%.
- Finalmente, el coeficiente de Dice es 0,7130. Esto nos indica que el modelo puede darse por válido.

3.4 Evaluación e interpretación

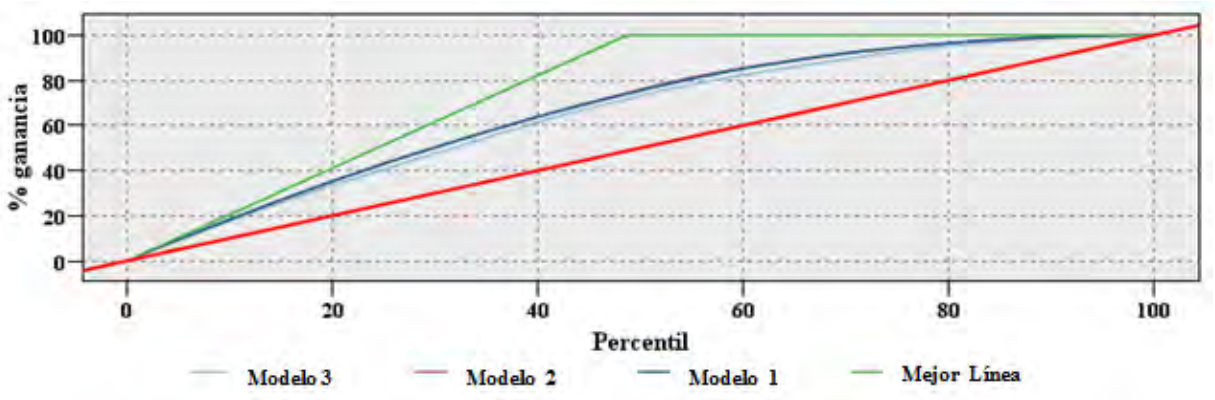
Dado que los tres modelos elegidos son considerados válidos es necesario determinara cual será el modelo a utilizar comparando sus indicadores de calidad, en la tabla 26 se muestra el resumen de estos.

Tabla 26
Principales indicadores de calidad de los modelos planteados

	Técnica empleada	Precisión Global	Área bajo la curva ROC	Coeficiente de Dice
Modelo 1	Árbol de decisión	74.96%	.8267	.7500
Modelo 2	Red neuronal	75.44%	.8274	.7598
Modelo 3	Regresión logística	72.32%	.7926	.7130

Tenemos que el modelo 2 correspondiente a la técnica de redes neuronales presenta una precisión global superior, una mayor área bajo la curva ROC y así mismo, un más elevado coeficiente Dice. Adicionalmente, la figura 14 muestra el gráfico de ganancias (ver anexo I) para los tres modelos propuestos:

Figura 14
Gráfico de ganancia de los modelos seleccionados



Si bien los tres modelos tienen un comportamiento similar en el gráfico, el modelo 3 es el más alejado a la mejor línea, dicho de otro modo, el que presenta menor ganancia.

A pesar que los dos primeros modelos son muy cercanos, el modelo 2 es superior, ofreciendo, en el 54% de los datos, una ganancia de 80.3% frente al 79.7% del modelo 1 y 77.2% del modelo 3.

A partir de los indicadores analizados concluimos que el modelo a utilizar para la identificación de afiliados potenciales será el modelo 2, el cual corresponde a la técnica de redes neuronales. Esta técnica se caracteriza por la complejidad en la interpretación de los modelos que genera, en respuesta a esto, el SPSS Modeler ofrece un análisis de sensibilidad a las variables de entrada, del cual se obtiene un número entre 0 y 1 por cada una de ellas como una medida de su importancia relativa. En la tabla 27 se muestra la importancia de cada variable de entrada para el modelo seleccionado.

Tabla 27
Importancia relativa de las variables de entrada del modelo 2

Variable	Importancia Relativa
SECCIONn	0.53
EDADn	0.49
AÑOS_ONPn	0.43
CATEGORIAn	0.37
DEPARTAMENTOn	0.33

Se observa que la variable “SECCIONn” es la que tiene mayor efecto en el pronóstico y, por el contrario, la variable “DEPARTAMETOn” es la que menos impacta. Cabe precisar que debido a la complejidad de esta técnica se útilizará además, el modelo 1 (técnica de árbol de decisión) para la identificación e interpretación de patrones.

Tras analizar el árbol de decisión se desprendieron diferentes patrones, o características que permitieron al modelo identificar a las personas migradoras al SPP. Las principales reglas se observan en la tabla 28.

Tabla 28
Principales condiciones para identificar afiliados migradores al SPP

Id	Condición
1	SECCIONn = 2 y CATEGORIAn = E
2	EDADn = (26;28] y AÑOS_ONPn = 3 y CATEGORIAn = A; B
3	SECCIONn = 8 y CATEGORIAn = E
4	AÑOS_ONPn = 2 y SECCIONn = 2
5	EDADn = (28;30] y AÑOS_ONPn = 2 y DEPARTAMENTOn = DEP18; DEP20; DEP23; DEP4
6	EDADn = (26;28] y AÑOS_ONPn = 3
7	AÑOS_ONPn = 1 y CATEGORIAn = D y DEPARTAMENTOn = DEP13; DEP22; DEP7
8	EDADn = (26;28] y AÑOS_ONPn = 4 y SECCIONn = 2 y DEPARTAMENTOn = DEP11; DEP15; DEP24
9	EDADn = (21;24] y AÑOS_ONPn = 7
10	EDADn = (28;30] y AÑOS_ONPn = 8
11	SECCIONn = 10 y CATEGORIAn = E
12	AÑOS_ONPn = 2 y CATEGORIAn = E y DEPARTAMENTOn = DEP11; DEP15; DEP24
13	SECCIONn = 1 y CATEGORIAn = E
14	EDADn = (21;24] y AÑOS_ONPn = 2
15	EDADn = (26;28] y AÑOS_ONPn = 4 y CATEGORIAn = C y DEPARTAMENTOn = DEP16; DEP17; DEP 25; DEP8
16	EDADn = (26;28] y CATEGORIAn = E y DEPARTAMENTOn = DEP11; DEP15; DEP24
17	AÑOS_ONPn = 1
18	EDADn = (28;30] y AÑOS_ONPn = 4 y CATEGORIAn = C y DEPARTAMENTOn = DEP18; DEP20; DEP23; DEP4
19	EDADn = (28;30] y AÑOS_ONPn = 2 y DEPARTAMENTOn = DEP16; DEP17; DEP 25; DEP8
20	CATEGORIAn = E y DEPARTAMENTOn = DEP16; DEP17; DEP 25; DEP8
21	EDADn = (28;30] y AÑOS_ONPn = 3 y CATEGORIAn = A; B
22	EDADn = (28;30] y AÑOS_ONPn = 4 y DEPARTAMENTOn = DEP16; DEP17; DEP 25; DEP8
23	SECCIONn = 9 y CATEGORIAn = E y DEPARTAMENTOn = DEP13; DEP22; DEP7
24	EDADn = (24;26] y AÑOS_ONPn = 3
25	AÑOS_ONPn = 2 y SECCIONn = 11
26	EDADn = <=21 y SECCIONn = 7
27	EDADn = (26;28] y AÑOS_ONPn = 4 y SECCIONn = 10 y CATEGORIAn = C y DEPARTAMENTOn = DEP11; DEP15; DEP24
29	EDADn = (28;30] y AÑOS_ONPn = 3
30	EDADn = (24;26] y AÑOS_ONPn = 2
31	EDADn = (30;32] y AÑOS_ONPn = 5 y CATEGORIAn = A; B
32	EDADn = (28;30] y AÑOS_ONPn = 4
33	EDADn = (28;30] y AÑOS_ONPn = 6

Adicionalmente, se detectaron las siguiente reglas generales:

- A mayor antigüedad en el SNP menor probabilidad de migrar.
- Cuanto menor es un afiliado, es mayor su índice de migración.
- Las categorías socioeconómicas C, D y E son las más propensas a migrar.
- Es notoria una mayor resistencia a la migración en las secciones más relacionadas a las actividades económicas propias del estado (S12, S13, S14).
- Las personas pertenecientes a la sección S10 presentan mayor disposición a la migración al SPP.

Estas generalizaciones resultan lógicas por varias razones. En primer lugar, dado que los 65 años es la edad esperada de jubilación, la edad de un afiliado tiene relación directa con el número de años de vida laboral que tiene por delante, lo que toma relevancia ya que según esto será el número estimado de aportes pendientes por realizar. En otras palabras, cuanto mayor sea una persona, menos serán los años estimados de aporte a su fondo, que es lo mismo a decir, menor será su beneficio al migrar al SPP. Esta misma relación se puede encontrar al analizar la cantidad de años que lleva una persona aportando al SNP, dado que cuanto menor sea este tiempo, mayor será el tiempo de aportaciones al SPP en caso decida migrar y por lo tanto, mayor su beneficio.

Por otro lado, en lo que respecta a las secciones, podría suponerse que las personas que laboran en las empresas vinculadas al Estado estén relacionadas a los servicios que él brinda, lo que incluiría la administración de su fondo de pensiones, caso contrario ocurría en las empresas, que por pertenecer al rubro de finanzas, sus trabajadores tienen un mayor conocimientos de los beneficios financieros del SPP, lo que facilitaría su migración. Finalmente, se podría explicar el hecho que las personas de categorías “A” y “B”, se afilien directamente al SPP porque las administradoras de fondos de pensiones priorizan elaborar estrategias para captar a este mercado por resultar evidentemente, más rentable.

3.5 Difusión y uso.

El modelo seleccionado se aplicó al universo de los datos (ver esquema en anexo K). Se obtuvieron 131,619 potenciales migradores al SPP. Tomando en consideración la precisión del modelo para predecir eventos positivos (79.80%), se estima que del listado de prospectos obtenidos aproximadamente, 105,032 personas son las que realizarán la migración. Finalmente, esta información será utilizada para reforzar la estrategia comercial. Los impactos sobre esta se desarrollarán en el capítulo V de este estudio.

Capítulo 4

Proyección de las afiliaciones en el SPP mediante el proceso KDD

4.1 Integración y recopilación.

La variable que se proyectará es el número de afiliaciones mensuales a nivel de empresa para el periodo comprendido entre octubre 2011 a septiembre 2012, para esto, se analizará el historial de los 36 meses previos, lo cual otorga un horizonte lo suficientemente amplio para detectar componentes estacionales anuales con un nivel de confiabilidad aceptable.

En el capítulo II, en el punto referente a "Integración y recopilación de información", se hizo referencia a la construcción de la base de datos donde se almacena toda esta información. Para el trabajo de este modelo se creó una nueva tabla, en esta base de datos, llamada "MODELO_PROYECCION" que consolida la información seleccionada a partir de las tablas donde previamente ya han sido cargados los datos.

4.2 Selección, limpieza y transformación.

Inicialmente, se planteó obtener la proyección de afiliaciones a nivel de empresa, pero al ser el promedio de afiliación mensual por empresa de 1.3, implicaría contar con muy pocas ocurrencias mensuales del evento a proyectar, lo que finalmente no permitiría obtener una confiabilidad aceptable. Debido a esto se utilizó el código CIIU (ver anexo E) que permite agrupar a las empresas según el giro, división y sector económico al que pertenece. De este modo al utilizar el giro como base para la proyección, teniendo 78 afiliaciones mensuales en promedio, se asegurará una mayor confiabilidad. La tabla 29 muestra las variables implicadas en la proyección.

Tabla 29
Variables utilizadas para el análisis de proyección

	Nombre	Tipo de Variable	Descripción	Tabla Origen
1	Afiliación mensual proyectada	DEPENDIENTE: Cuantitativa	Es el número de afiliaciones proyectada mes a mes (oct 2011-sep2012).	FACT_AFP
2	Afiliación mensual histórica	DEPENDIENTE: Cuantitativa	Es el número de afiliaciones mensual (oct 2008-sep2011).	FACT_AFP
3	Ruc	INDEPENDIENTE: Cualitativa	Indica el ruc de la empresa donde se realizó la afiliación al SPP.	FACT_AFP
4	FEC_AFILIACION	INDEPENDIENTE: Cualitativa	Indica la fecha en que se realizó la afiliación al SNP.	FACT_ONP
5	FEC_AFILIACION	INDEPENDIENTE: Cualitativa	Indica la fecha en que se realizó (si fuera el caso) la migración al SPP.	FACT_AFP
6	CATEGORIA	INDEPENDIENTE: Cualitativa	Indica la escala económica a la que pertenecen los posibles prospectos.	FACT_CLIENTES
7	Departamento	INDEPENDIENTE: Cualitativa	Indica el departamento donde se encuentra ubicada la empresa.	FACT_UBIGEO

4.3 Minería de datos.

Como se explicó anteriormente, en base a la información mensual de afiliaciones para el periodo comprendido entre octubre 2008 a septiembre 2011 se realizó una proyección de los 12 meses siguientes para cada giro de negocio. Para ello, se emplearon algoritmos de series de tiempo empleando el software para minería de datos Clementine 11.1. La tabla 30 muestra los resultados obtenidos.

Tabla 30
Resumen de resultados proyección giros

Modelo	Giros	R2 >= 70%
Estacional simple	202	23.8%
Aditivo de Winters	109	39.4%
ARIMA	6	0%
Multiplicativo de Winters	4	100%
Simple	1	0%
Total	**322**	**38.2%**

Nota: En la primera columna se muestran los distintos modelos de serie de tiempo empleados por el software para el pronóstico de cada giro (ver anexo C: Criterios de suavizado exponencial de series temporales). En la segunda columna se muestra la distribución de los giros en los modelos que les generaron un R cuadrado más elevado. En la tercera columna se muestra el porcentaje de giros para cada modelo que obtuvieron un R cuadrado mayor o igual a 70%.

Las empresas de las que se obtuvo afiliaciones se reparten en 322 giros económicos. Se aprecia que la gran mayoría presentan modelos con estacionalidad y tendencia constante (Estacional simple y Aditivo de Winters), lo cual resulta lógico debido a que las afiliaciones están sujetas a la demanda del mercado laboral, que es de carácter estacional. Sin embargo, se puede apreciar que solo un 38.2% de los giros tienen un modelo con un nivel de R cuadrado aceptable (superior o igual a 70%). Esto se debe a que la cantidad de ocurrencias en las muestras de estos giros no es la suficiente para realizar un modelamiento adecuado.

Para poder contrarrestar el impacto en la confiabilidad producida en los giros con poco volumen de afiliaciones, se utilizará una agrupación más general como es el caso de las divisiones y los sectores, siguiendo el mismo procedimiento de modelamiento. Las tablas 31 y 32 muestran los resultados obtenidos utilizando agrupaciones a nivel se divisiones y sectores respectivamente.

Tabla 31
Resumen de resultados proyección divisiones

Modelo	Divisiones	R2 >= 70%
Aditivo de Winters	32	59.4%
Estacional simple	25	28.0%
Multiplicativo de Winters	1	100.0%
Total	**58**	**46.6%**

Nota: En la primera columna se muestran los distintos modelos de serie de tiempo empleados por el software para el pronóstico de cada división. En la segunda columna se muestra la distribución de las divisiones en los modelos que les generaron un R

cuadrado más elevado. En la tercera columna se muestra el porcentaje de divisiones para cada modelo que obtuvieron un R cuadrado mayor o igual a 70%.

Tabla 32
Resumen de resultados proyección sectores

Modelo	Sectores	R2 >= 70%
Aditivo de Winters	9	88.9%
Estacional simple	7	57.1%
Total	**16**	**75.0%**

Nota: En la primera columna se muestran los distintos modelos de serie de tiempo empleados por el software para el pronóstico de cada sector. En la segunda columna se muestra la distribución de los sectores en los modelos que les generaron un R cuadrado más elevado. En la tercera columna se muestra el porcentaje de sectores para cada modelo que obtuvieron un R cuadrado mayor o igual a 70%.

Notamos que la confiabilidad de la proyección de los sectores es superior a la de las divisiones y estas, a su vez, mejor que la de giros. Esto ocurre porque según se agrupan mayor cantidad de empresas, incrementa el volumen de datos que se analiza para la aproximación del modelo, sin embargo, se va perdiendo la homogeneidad entre las empresas dentro de un mismo grupo. Esto se aprecia nuevamente, en la proyección general del sistema, la cual se observa en la tabla 33.

Tabla 33
Confiabilidad resultados proyección general del sistema

Modelo	R2
Aditivo de Winters	94.0%

Para buscar un equilibrio, en primer lugar se reducirá cada proyección de sector, división y sistema a nivel de giro de manera proporcional, luego se comparará con la proyección real. En todos los casos se priorizará la proyección del giro, si ésta no cuenta con un nivel de R cuadrado superior o igual a 70%, se procederá a una clasificación superior, es decir; se dará prioridad a la agrupación más homogénea siempre que tenga un R cuadrado mayor o igual a 70%. En la tabla 34, se muestra un resumen de las proyecciones que finalmente, se utilizarán.

Tabla 34

Distribución de proyecciones a nivel de giros

Clasificación	N° Giros	R2 Promedio
Giro	123	86.3%
División	169	89.4%
Sector	27	95.4%
Sistema	3	96.7%
Total	**322**	**88.8%**

Nota: En la primera columna se muestran los distintos niveles de agrupamiento de empresas de manera creciente. En la segunda columna se muestra la distribución de los giros en los demás niveles de agrupamiento o manteniendo su nivel de giro. En la tercera columna se muestra el promedio de R cuadrado para nada nivel de agrupamiento.

4.4 Evaluación e interpretación.

Es necesario estimar la contribución de las transformaciones y clasificaciones realizadas para el aumento de la confiabilidad en la proyección global. Para esto, una vez obtenida las proyecciones finales a nivel de giro, el siguiente paso es llevarlas a nivel de cada empresa. Para lo cual, nuevamente se hará uso de la proporcionalidad histórica, sin embargo, para esta matriz de proporcionalidad se deberá reducir el horizonte histórico a una muestra más reciente con la finalidad de evitar considerar producción en empresas que no se encuentren activas. Se considerará la información de los 12 meses previos al periodo proyectado; es decir, de octubre 2010 a septiembre 2011.

Siguiendo el procedimiento descrito anteriormente, obtenemos la proyección conjunta del sistema, la cual se muestra en la tabla 35.

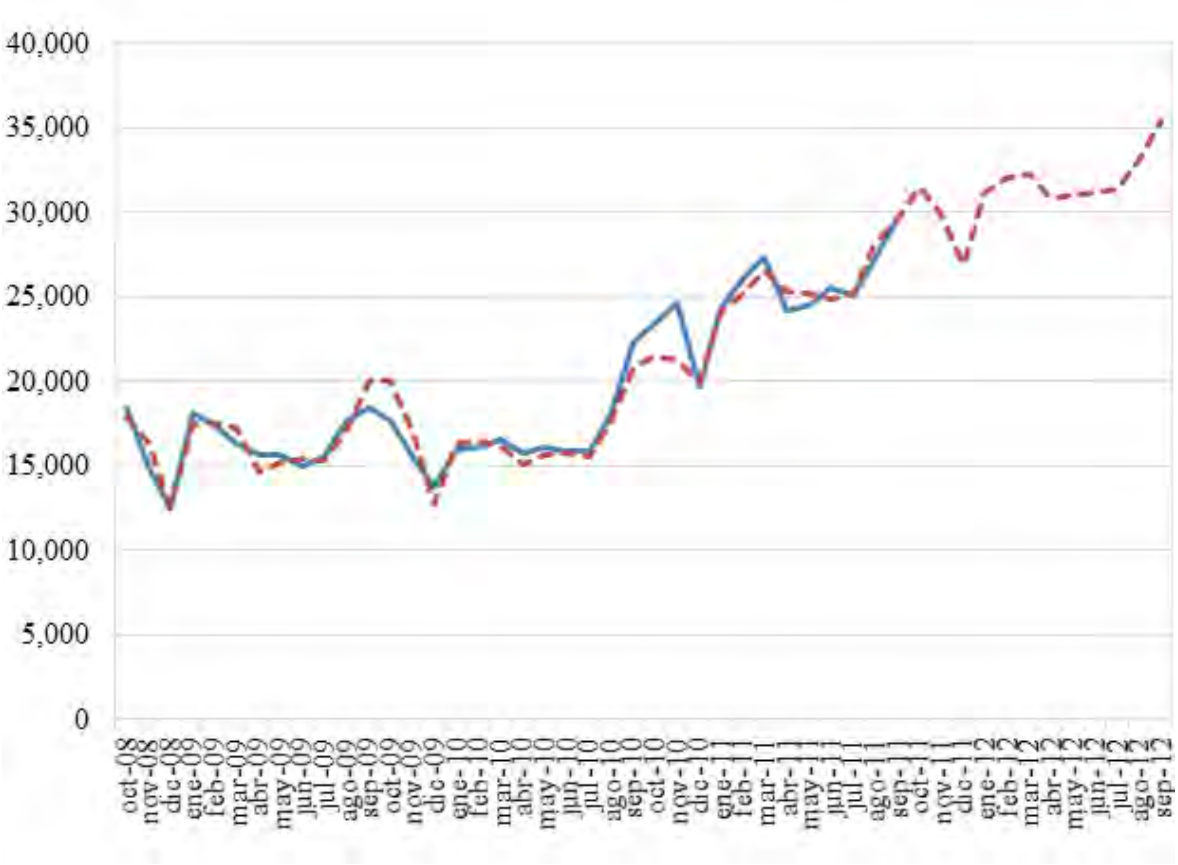
Tabla 35
Proyección global

Periodo	Real	Proyección	Diferencia
oct-08	18,524	17,854	-3.6%
nov-08	14,962	16,368	9.4%
dic-08	12,473	12,528	0.4%
ene-09	18,172	17,477	-3.8%
feb-09	17,478	17,589	0.6%
mar-09	16,325	17,235	5.6%
abr-09	15,696	14,607	-6.9%
may-09	15,706	15,149	-3.5%
jun-09	14,974	15,477	3.4%
jul-09	15,536	15,263	-1.8%
ago-09	17,715	17,108	-3.4%
sep-09	18,449	20,012	8.5%
oct-09	17,659	19,986	13.2%
nov-09	15,566	16,980	9.1%
dic-09	13,777	12,736	-7.6%
ene-10	15,965	16,342	2.4%
feb-10	16,054	16,403	2.2%
mar-10	16,503	16,220	-1.7%
abr-10	15,750	15,120	-4.0%
may-10	16,164	15,690	-2.9%
jun-10	15,871	15,774	-0.6%
jul-10	15,858	15,539	-2.0%
ago-10	17,960	17,522	-2.4%
sep-10	22,294	20,792	-6.7%
oct-10	23,425	21,522	-8.1%
nov-10	24,703	21,271	-13.9%
dic-10	19,725	19,920	1.0%
ene-11	24,298	24,068	-0.9%
feb-11	26,051	25,221	-3.2%
mar-11	27,340	26,535	-2.9%
abr-11	24,175	25,338	4.8%
may-11	24,540	25,236	2.8%
jun-11	25,550	24,868	-2.7%
jul-11	25,164	25,244	0.3%
ago-11	27,263	28,150	3.3%
sep-11	29,742	29,561	-0.6%
Total	697,407	692,705	-0.7%

Nota: La primera columna muestra el periodo (mes-año), la segunda columna el nivel de afiliación real (ver Anexo K), la tercera el nivel de afiliación pronosticado y la cuarta la variación porcentual del pronóstico frente a la cifra real.

Estos datos se pueden ver reflejados en la figura 15 se muestra la comparación entre las gráficas de la producción real versus la gráfica de la proyección.

Figura 15
Comparación gráfica real versus proyección



Nota: En el eje vertical se muestra la cantidad de afiliaciones en número y en el horizontal el mes en que se realiza dicha captación., siendo el gráfico azul la captación real y el rojo punteado en base a las proyecciones.

El R cuadrado de la proyección total conjunta es de 94.7%, 0.7% por encima de la proyección macro correspondiente al sistema. Esto se debe al efecto de haber tomado una muestra más singular y homogénea al reducir el análisis a nivel de giros, permitiéndole al modelo captar comportamientos propios de cada giro, que normalmente se hubieran disipado en un análisis más macro.

4.5 Difusión y uso.

La proyección obtenida estima un total de 376,854 nuevas afiliaciones para el periodo que va desde octubre del 2011 a septiembre del 2012. Para el empleo de esta información es necesario obtener la distribución geográfica del potencial de afiliaciones obtenido, para la cual se considerará el número de trabajadores vigentes de cada empresa a Septiembre 2011. De esta manera, partiendo del supuesto de que la necesidad de contratación es proporcional al tamaño del negocio de las empresas en cada localidad, se obtiene la repartición del potencial, tanto de afiliaciones nuevas como en empresas, el cual se muestra en la tabla 36.

Tabla 36
Distribución geográfica del potencial de afiliaciones nuevas al SPP

Departamento	Total Afiliaciones	Número Sucursales	Potencial Sucursales
Lima	179,158	12,912	13.9
La Libertad	24,762	1,957	12.7
Cuzco	15,236	1,301	11.7
Piura	18,877	1,845	10.2
Ica	13,991	1,400	10.0
Loreto	11,556	1,201	9.6
San Martin	7,729	934	8.3
Tacna	5,440	662	8.2
Puno	7,463	910	8.2
Cajamarca	9,805	1,214	8.1
Arequipa	16,184	2,007	8.1
Lambayeque	10,821	1,403	7.7
Huánuco	5,461	786	6.9
Madre De Dios	1,526	250	6.1
Ucayali	4,522	764	5.9
Callao	18,939	3,259	5.8
Amazonas	1,643	294	5.6
Junín	7,584	1,392	5.4
Tumbes	2,148	410	5.2
Ancash	5,513	1,053	5.2
Huancavelica	1,811	364	5.0
Moquegua	2,009	412	4.9
Ayacucho	2,270	513	4.4
Apurímac	1,113	332	3.4
Pasco	1,294	411	3.1
Total	**376,854**	**37,986**	**9.9**

Nota: La primera columna muestra los departamentos considerados en la muestra. La segunda columna contiene el número de afiliaciones proyectadas para cada departamento. En la tercera columna se muestra el número de sucursales diferenciadas

en cada departamento. Finalmente en la cuarta columna se muestra el número de afiliaciones promedio por sucursal de cada departamento.

El potencial viene determinado por la densidad de afiliación promedio de las sucursales de las empresas en cada departamento. Este ratio permite calcular el grado de dispersión de las afiliaciones en cada departamento. Factor que resulta clave, considerando que una amplia concentración de afiliaciones en un mismo punto, reduce tiempo y recursos. Finalmente, esta información será utilizada para reforzar la estrategia comercial. Los impactos sobre esta se desarrollarán en el capítulo V de este estudio.

Capítulo 5

Análisis del impacto sobre la estrategia comercial

5.1 El mercado potencial

En capítulos anteriores se obtuvieron, por un lado, 105,032 potenciales migradores del SNP y por otro, un estimado de 376,855 futuras afiliaciones directas al SPP lo que en suma conforman un mercado potencial de 481,887 nuevos afiliados. Aquí cabe precisar que cada AFP realiza su propia estimación del potencial de afiliaciones, en su mayoría sólo considerando las afiliaciones históricas de cada empresa y suponiendo un comportamiento similar a futuro. La estimación que se considera en el presente estudio, en cambio, si bien también contempla el análisis de las afiliaciones históricas, lo hace de manera más minuciosa, evaluando en principio, todo el mercado y luego buscando la proyección más precisa para cada empresa. Además, se tomó en consideración un análisis en el stock de afiliados al SNP para identificar a los posibles migradores al SPP, lo que permitió obtener un mercado potencial más amplio y completo.

5.2 Las empresas objetivo

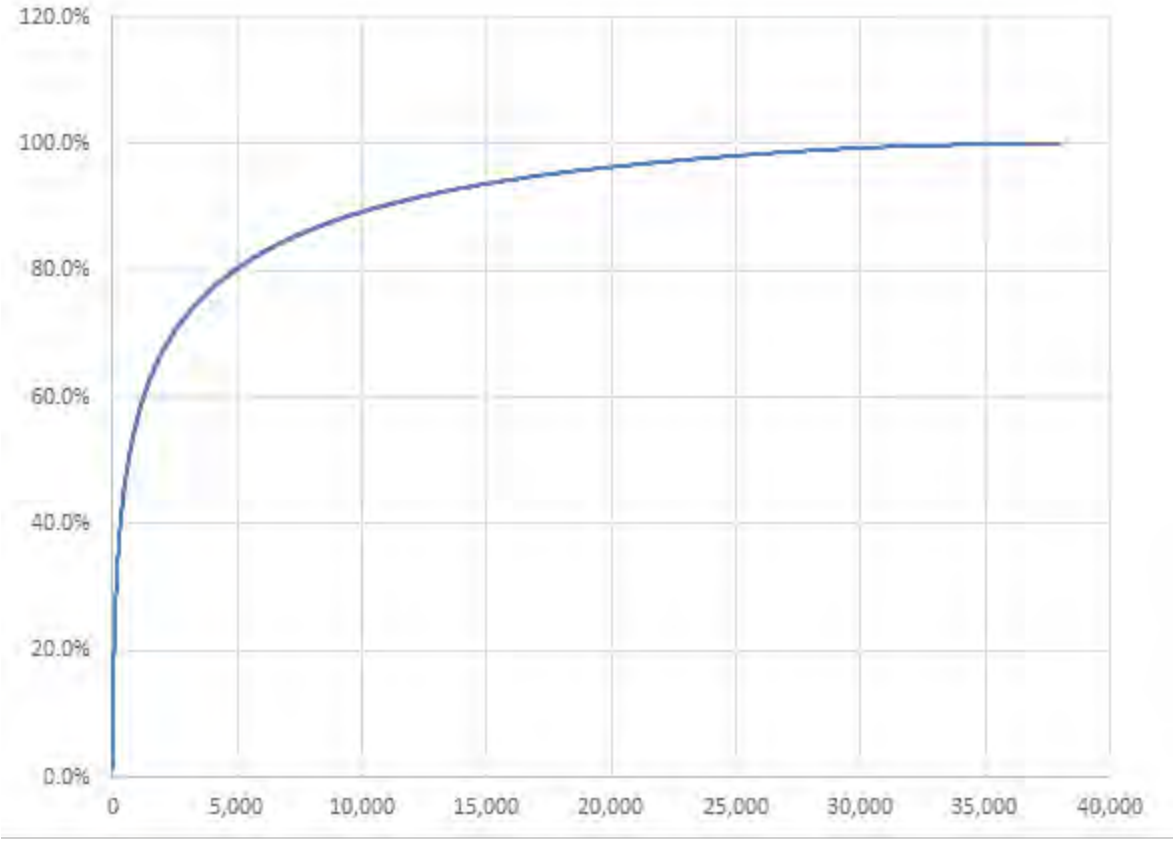
Actualmente, hay registradas aproximadamente 500 mil empresas en el Perú. Por ser un número tan elevado sería costoso e insostenible visitarlas a todas, ya que además de demandar un gran equipo de ventas, no todas las empresas tienen un volumen potencial significativo de afiliaciones por lo que no resultaría rentable trabajarlas. Esta situación hace que las AFP se deban centrar solo en un grupo de "empresas objetivo" que consideran, por diferentes motivos, las más rentables.

En la estrategia que se plantea, se entiende por "sucursal" a la filial de una empresa en un departamento. Analizando las sucursales, se detecta que si se agrupan las primeras (ordenadas de mayor a menor por nivel de afiliación) que concentran el 80% del potencial de afiliaciones directas al SPP, la pendiente acumulada se estabiliza en 45

grados y cada sucursal adicional a este grupo llevaría a una razón de cambio acumulada con un ángulo de inclinación inferior a 45 grados; siendo el aporte de esta sucursal adicional, proporcionalmente menos significativo (ver figura 16). Dicho grupo está conformado por las primeras 5,179 sucursales con mayor potencial.

Figura 16

Acumulado porcentual de afiliaciones ascendente por sucursal



Nota: En el eje vertical se muestra el porcentaje acumulado en número de afiliaciones respecto al total y en el horizontal se muestra el acumulado en número de sucursales.

Finalmente, es importante conocer la distribución de éstas sucursales entre los departamentos del Perú. La tabla 37 muestra ésta información.

Tabla 37
Distribución de empresas objetivo en los departamentos del Perú

Departamento	Frecuencia.	Porcentaje.
Amazonas	29	0.60%
Ancash	114	2.20%
Apurímac	29	0.60%
Arequipa	276	5.30%
Ayacucho	40	0.80%
Cajamarca	171	3.30%
Callao	344	6.60%
Cuzco	161	3.10%
Huancavelica	34	0.70%
Huánuco	76	1.50%
Ica	194	3.70%
Junín	138	2.70%
La Libertad	251	4.80%
Lambayeque	135	2.60%
Lima	2,278	44.00%
Loreto	150	2.90%
Madre De Dios	18	0.30%
Moquegua	30	0.60%
Pasco	24	0.50%
Piura	268	5.20%
Puno	110	2.10%
San Martin	139	2.70%
Tacna	56	1.10%
Tumbes	25	0.50%
Ucayali	89	1.70%

Nota: La primera columna muestra los departamentos considerados en la muestra. La segunda columna contiene sucursales pertenecientes a la agrupación en cada departamento. Finalmente en la tercera columna se muestra la distribución porcentual de las sucursales en los departamentos. Todos los datos anteriores fueron obtenidos a partir de la base de afiliaciones por empresas tratada en el capítulo anterior.

5.3 La meta

La meta que cada AFP se establece, guarda relación con la participación histórica que ha tenido en el mercado y sus expectativas de crecimiento. En este caso se debe fijar cuántas afiliaciones se espera conseguir en el periodo que va desde octubre de 2011 a septiembre de 2012 (doce meses). Dado que el estudio no va dirigido a una AFP en particular, se considerará el escenario en que las AFP, que en ese entonces eran cuatro (Profuturo, Prima, Integra y Horizonte), desean obtener como mínimo el 25% del

mercado de afiliaciones directas al SPP lo que en este caso, dado que se estimó un mercado potencial de 376,855 nuevos afiliados, representa 94,214 afiliaciones. Esta meta deberá ser conseguida trabajando las 5,179 sucursales calificadas como objetivo, las cuales, si bien se esperaría obtener el 25% del potencial que concentran de nuevas afiliaciones (94,214), también cuentan con más de 13,600 posibles migradores del SNP en total asignados a estas sucursales, lo que permitiría lograr cómodamente el objetivo. No se consideró en la estimación de la meta a los afiliados provenientes del SNP porque no se tiene antecedentes de trabajo de captación con ellos, sin embargo, trabajar este nicho de mercado nunca antes explotado por las AFP resulta una ventaja competitiva clara, dado que implicaría un aumento del 14% de la productividad general.

Dado que las sucursales objetivo concentran el 80% de las afiliaciones del sistema, es necesario obtener una participación de mercado del 30% en estas sucursales, para obtener una cantidad de afiliaciones equivalente a una participación de mercado de 25% del sistema total. Esa sería la meta.

5.4 El equipo de ventas

Según la información publicada por la SBS (ver anexo K), la productividad esperada de un asesor de ventas es de 30 afiliaciones al mes, considerando 20 sucursales por asesor (asumiendo 20 días útiles al mes, considerando una visita por día a una empresa). De las 5,179, obtenemos un total de 261 asesores, lo cual nos permite alcanzar el 25% del mercado total con una productividad estándar. Esta dotación es inferior en 3.6% a la dotación promedio del mercado (271), lo cual demuestra una mejor delimitación del *target*, dado que se ha considerado una productividad acorde al promedio del mercado, esta diferencia en eficiencia se vería incrementada más aun por el 14% extra de productividad que se obtendría atacando el nicho de mercado potencial SNP. Teniendo la distribución de sucursales, se realizó la estimación del número de asesores de venta necesarios por departamento. La tabla 38 muestra esta información.

Tabla 38
Distribución de asesores de venta en los departamentos del Perú

Departamento	# Asesores de Venta.
Amazonas	1
Ancash	6
Apurímac	1
Arequipa	14
Ayacucho	2
Cajamarca	9
Callao	17
Cuzco	8
Huancavelica	2
Huánuco	4
Ica	10
Junín	7
La Libertad	13
Lambayeque	7
Lima	114
Loreto	8
Madre De Dios	1
Moquegua	2
Pasco	1
Piura	13
Puno	6
San Martin	7
Tacna	3
Tumbes	1
Ucayali	4

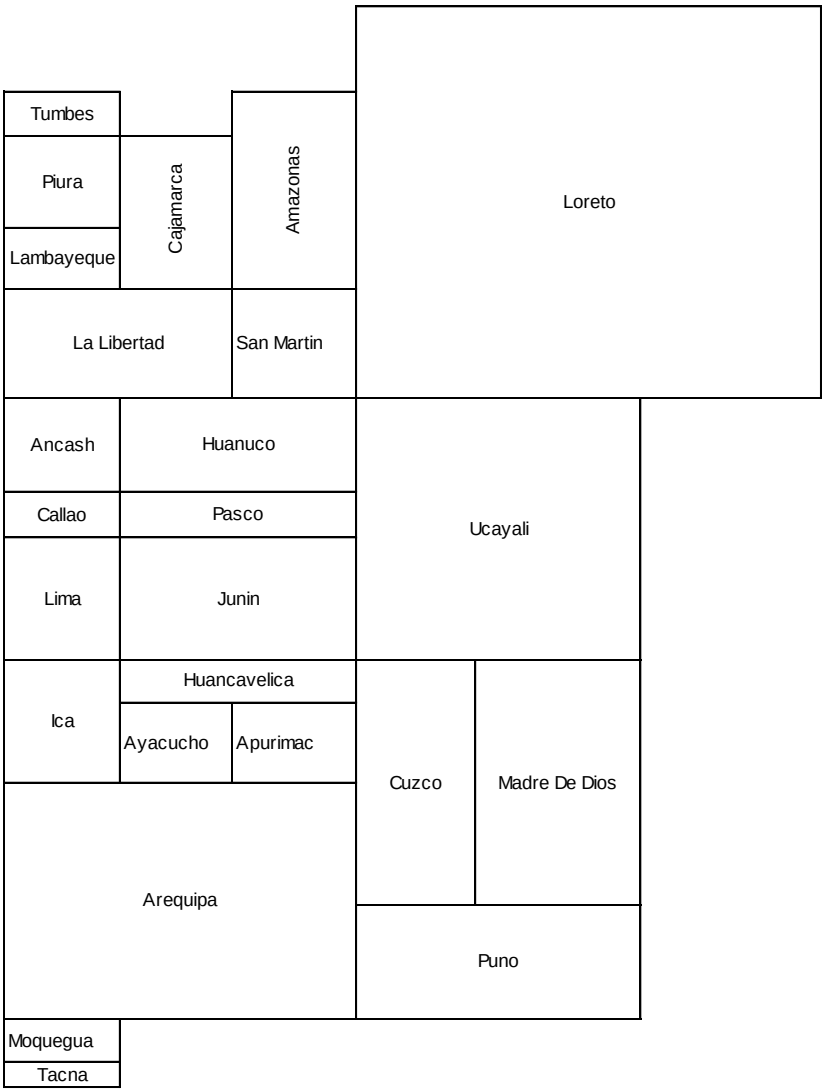
Nota: La primera columna muestra los departamentos considerados en la muestra. La segunda la distribución de los asesores en los departamentos. Esta tabla se elaboró en base a los criterios mencionados en el párrafo anterior y al número de afiliaciones por departamento.

Se debe tener en cuenta que para la estimación de asesores solo se consideró la captación esperada de afiliaciones. Para otros conceptos, propios de cada AFP, como la cobertura del servicio a sus afiliados, se deberá ajustar el cálculo.

5.5 Las agencias de venta

El número y distribución de agencias principales como puntos de concentración de los asesores se realizó a partir de un diagrama de proximidad. En la figura 17 se muestra el gráfico que representa a Perú según sus departamentos:

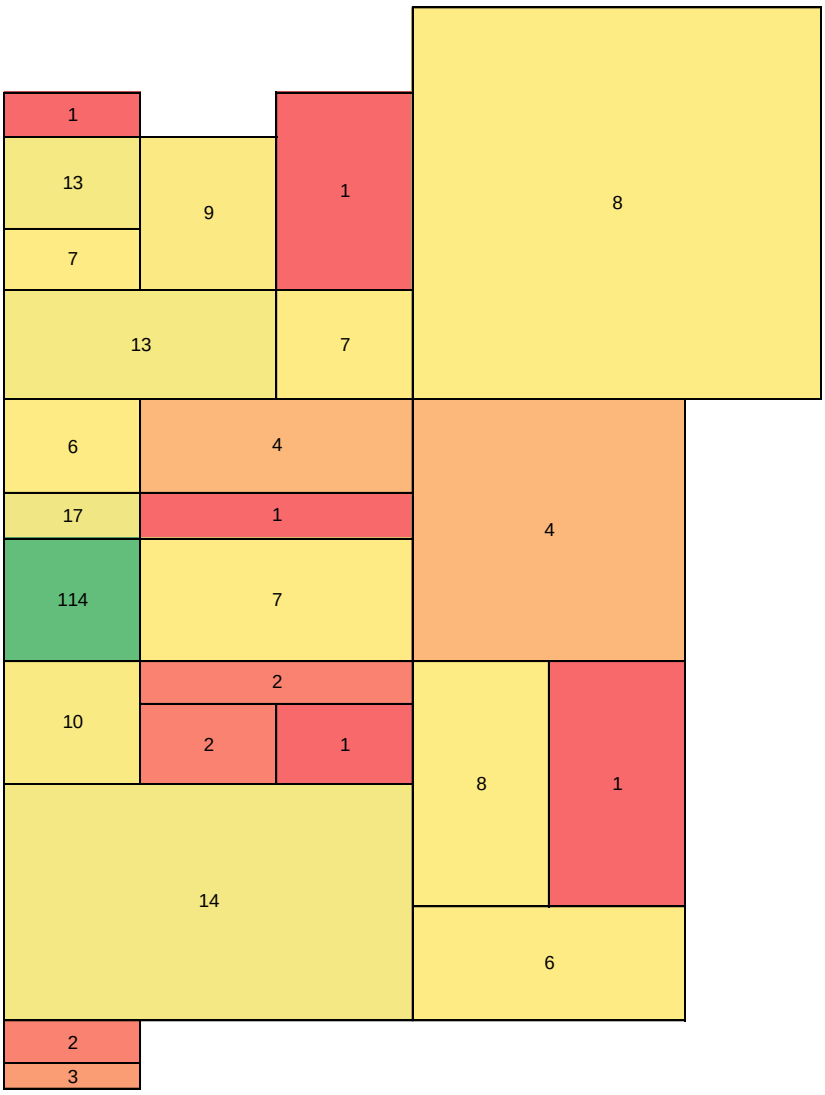
Figura 17
Leyenda del Mapa de calor de distribución sugerida de dotación



Nota: Cada recuadro representa un departamento del Perú y tiene una correspondencia en la siguiente figura de Distribución.

Figura 18
Mapa de calor de distribución sugerida de dotación

Distribución:



Nota: Cada número representa la cantidad de asesores en el departamento referenciado en la Leyenda para cada recuadro. Los colores corresponden a una escala de gradación, siendo el color verde los de mayor concentración en proporción al total de la muestra y los rojos los de menor concentración.

Finalmente, analizando los agrupamientos y proximidad entre los departamentos se determinaron las agencias comerciales, las cuales se muestran en la tabla 39.

Capítulo 6

Conclusiones

A lo largo del estudio queda en manifiesto que el uso de la minería de datos resulta ser efectivo para una aproximación más certera del potencial de ventas en una empresa, para el caso de una AFP el marco de esta investigación está dado por el número de afiliaciones realizadas.

En el análisis de perfiles, se estudiaron las características de las personas provenientes del SNP propensas a afiliarse al SPP, obteniéndose un modelo de redes neuronales con una precisión para predecir eventos positivos de 79.80% (105,032 personas) sobre los casos evaluados, siendo superior al 70% esperado para esta clase de modelos y claramente mejor que el 50% determinado por la arbitrariedad. Esto se logró debido a que se identificaron eficientemente las variables que definen los patrones de comportamiento para este grupo de personas, encontrándose que la edad, la escala económica y el rubro de trabajo son las características más representativas de las personas del perfil buscado.

En el análisis de las afiliaciones, se empleó la data histórica registrada en el mercado, gracias a la aplicación de la metodología KDD se logró identificar patrones que permitieron definir modelos de serie de tiempo que perfilaron los ciclos de contratación de las empresas, una característica que trasciende a una proyección regular pues surge de la minería de datos, que permite identificar relaciones trascendentales no evidentes en un análisis plano de los datos, logrando hacer una pronóstico más certero, estimando 376,855 afiliaciones para el periodo de Octubre 2011 a Septiembre 2012, siendo la cifra real 380,150 afiliaciones, lo que significa un R cuadrado del 94.7%.

Finalmente, se puede concluir que la estimación total de afiliaciones obtenida para el periodo de los doce meses previos a la reforma del SPP, muestra una aproximación del real potencial del mercado, lo que permite calcular el valor óptimo y la manera más eficiente de emplear los recursos claves en la estrategia comercial como son: la meta, las empresas objetivo, el número de asesores de venta y la mejor manera de distribuirlos a través de las agencias a nivel nacional.

Esto genera una diferencia significativa en comparación con las prácticas promedio del sector, la que se ve claramente reflejada en los siguientes puntos:

- Aumento del mercado potencial.- al considerar no sólo una proyección más fina del crecimiento del crecimiento del sector sino incluyendo además, la estimación de los potenciales migradores al SPP provenientes del SNP.

- Mejora en la productividad promedio por asesor.- la delimitación del target propuesta permitió una focalización y distribución más óptima de los recursos.

- Reducción del equipo de ventas.- la metodología empleada permitió reducir además, el equipo de ventas requerido en un 3.6% por debajo del promedio del mercado.

Adicionalmente, cabe resaltar que el presente estudio conjuntamente muestra la contribución del uso de técnicas de minería de datos y el proceso KDD para establecer pronósticos y predicciones con alto grado de precisión, lo que hace de esta tesis además, una guía a aplicarse en diferentes escenarios e industrias para la extracción del conocimiento en grandes bases de datos.

Bibliografía

- Andrés Fernández Romero: Dirección y Planificación Estratégicas en las Empresas y Organizaciones, 2010, Diaz de Santos.

- Adrinzén, G. Barúa, R. Tres Años del Sistema Privado de Pensiones – 1993-1996. San Isidro. Lima-Peru 1996.

- Box, G. E. P., G. M. Jenkins, y G. C. Reinsel. 1994. *Time series analysis: Forecasting and control*, 3rd ed. Englewood Cliffs, N.J.: Prentice Hall.

- Cesar Perez Lopez, Daniel Santin Gonzalez: Minería de Datos, Técnicas y Herramientas, Primera Edición, Segunda reimpresión, 2008.

- Estadísticas Sistema Privado de Pensiones. (2011, Octubre20). Recuperado 20:00, Octubre 20, 2011, de SBS (Superintendencia de Banca y Seguros y AFP): http://www.sbs.gob.pe/0/modulos/JER/JER_Interna.aspx?ARE=0&PFL=0&JER=150

- Fernando Berzal, Ignacio Blanco, Juan Carlos Cubero & Nicolás Marín: Component-based Data Mining Frameworks. Communications of the ACM, Vol. 45, No. 12, December 2002, pp. 97-100.

- Francisco Javier Garrido: "Pensamiento Estratégico: hacia el ADN de la planificacion estratégica", 2005, Primera Edición, *Executive Business School*, Santiago, Chile.

- Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

- Pena, D., G. C. Tiao, y R. S. Tsay, eds. 2001. *A course in time series analysis*. Nueva York: John Wiley and Sons.

- Usama Fayyad, Gregory Piatetsky-Shapiro & Padhraic Smyth: The KDD process for extracting useful knowledge from volumes of data, Communications of the ACM, November 1996, pp. 27-34.

Tabla 39

Agrupamiento de departamentos según agencias comerciales

Agencia	Departamento	Asesores	Total
Ancash	Ancash	6	11
	Huánuco	4	
	Pasco	1	
Arequipa	Arequipa	14	19
	Moquegua	2	
	Tacna	3	
Cajamarca	Cajamarca	9	9
Cuzco	Cuzco	8	9
	Madre de Dios	1	
Ica	Apurímac	1	15
	Ayacucho	2	
	Huancavelica	2	
	Ica	10	
Junín	Junín	7	11
	Ucayali	4	
La Libertad	La Libertad	13	20
	Lambayeque	7	
Lima	Callao	17	131
	Lima	114	
Loreto	Loreto	8	8
Piura	Piura	13	14
	Tumbes	1	
Puno	Puno	6	6
San Martin	Amazonas	1	8
	San Martin	7	

Nota: La primera columna muestra las agencias sugeridas. La segunda los departamentos incluidos en cada agencia. Finalmente en la tercera columna se muestra la distribución de los asesores en los departamentos y totalizados por cada agencia.

ANEXO A

GLOSARIO DE TÉRMINOS DEL SPP

I. Términos del SPP

1. **Afiliados Activos**: Personas incorporadas al Sistema Privado de Pensiones que no se encuentran percibiendo ninguna pensión en el SPP.

2. **Afiliados Pasivos**: Se entenderá que un afiliado tiene la condición de pasivo cuando se encuentre percibiendo una pensión bajo cualquiera de los productos previsionales que otorga el SPP. Para dicho efecto, se entenderá que el derecho a la percepción del beneficio –en el caso de jubilación- queda perfeccionado con la suscripción de la sección I de la solicitud de pensión, siempre y cuando se otorgue la conformidad de la misma. En el caso de invalidez, el derecho a la pensión queda perfeccionado con la existencia de un dictamen de invalidez, parcial o total, consentido o emitido por el Comité Médico de la Superintendencia (COMEC).

3. **Afiliación Mensual**: Se entiende por Afiliación Mensual al flujo de personas incorporadas al Sistema Privado de Pensiones en el transcurso del mes indicado.

4. **Aportes Obligatorios**: Pagos mensuales efectuados por cada afiliado sobre su remuneración asegurable y destinados a su Cuenta Individual de Capitalización. También se denomina Aportes al primer rubro del cuadro Ingresos y Egresos del Fondo de Pensiones, rubro que está conformado por la sumatoria de los aportes efectuados en el mes por todos los afiliados (correspondientes a pagos oportunos o a pagos devengados en meses anteriores) y los pagos en exceso y pagos en defecto del empleador.

5. **Aportes Voluntarios**: Son aportes que pueden realizar los afiliados o sus empleadores para complementar sus aportes obligatorios con la finalidad de incrementar su pensión y pueden ser de dos tipos aportes voluntarios con fin previsional y sin fin previsional. Los aportes voluntarios con fin previsional pueden ser realizados por el afiliado o su empleador, son inembargables, y pueden ser retirados al final de la etapa laboral activa del trabajador. Los aportes sin fin previsional pueden ser realizados por los afiliados con al menos cinco años de incorporación al SPP, o tener 50 años de edad, son embargables, se pueden retirar total o parcialmente, y pueden llegar a convertirse en aportes con fin previsional.

6. **Beneficiario**: Personas que tienen derecho a pensión de sobrevivencia al fallecimiento del afiliado activo o pasivo. Las normas del SPP contemplan como beneficiarios a los hijos menores de dieciocho (18) años, o mayores de dieciocho incapacitados de manera total y permanente para el trabajo; al cónyuge o concubino; y a los padres inválidos

total o parcialmente o mayores de sesenta (60) años y que hayan dependido económicamente del causante.

7. **Beneficiarios de afiliados activos**: Personas que tienen derecho a pensión de sobrevivencia al fallecimiento de un afiliado activo. Los beneficiarios de afiliados activos pasan por un proceso de elección de modalidad de pensión para el grupo familiar.

8. **Beneficiarios de afiliados pasivos**: Personas que tienen derecho a pensión de sobrevivencia al fallecimiento de un afiliado pasivo. Los beneficiaros de afiliados pasivos reciben pensión de sobrevivencia bajo la misma modalidad que estuvo percibiendo el afiliado, en los casos de Renta Vitalicia Familiar o Renta Temporal con Renta Vitalicia Diferida. En caso el afiliado haya estado percibiendo una pensión bajo Retiro Programado, los beneficiarios pueden continuar percibiendo pensión bajo dicha modalidad o contratar otra modalidad, de acuerdo a las disposiciones establecidas al respecto en el Título VII del Compendio de Normas Reglamentarias del SPP.

9. **Bono de Reconocimiento**: Título valor que reconoce los aportes realizados por los trabajadores al Sistema Nacional de Pensiones (SNP) antes de incorporarse al Sistema Privado de Pensiones (SPP). El Bono de Reconocimiento es redimible al cumplimiento de los 65 años del trabajador o con ocasión de su muerte, la jubilación anticipada, o la declaración de invalidez total permanente del titular. Actualmente las normas del SPP contemplan cuatro (4) tipos de Bono: el Bono de Reconocimiento 1992, el Bono de Reconocimiento 1996, Bono de Reconocimiento D.L N° 20530 y el Bono de Reconocimiento 2001. El Bono de Reconocimiento D.L. N° 20530 se encuentra pendiente de reglamentación.

10. **Cartera Administrada**: Es el total de los activos en los que se encuentran invertidos los recursos del Fondo de Pensiones más el Encaje Legal.

11. **Categoría de Riesgo Equivalente**: Clasificación de Riesgo de Inversión establecida por la Superintendencia que tiene equivalencias con las clasificaciones establecidas por las empresas clasificadoras de riesgo.

12. **Cobertura del Seguro de Invalidez, Sobrevivencia y Gastos de Sepelio**: Derecho del que gozan aquellos afiliados que no se encuentran en ninguna de las circunstancias denominadas exclusiones y cumplen con alguna de las condiciones que originan el acceso a cobertura. El afiliado cubierto tiene el beneficio de recibir una prestación de invalidez, o sus beneficiarios una pensión de sobrevivencia, la que mantiene una relación directa y proporcional con la remuneración mensual del afiliado, conforme a las regulaciones establecidas en el Título VII del Compendio de Normas Reglamentarias del SPP. El financiamiento de las pensiones es completado por el Aporte Adicional que efectúe la Empresa de Seguros.

13. **Comisión por la Gestión de los Aportes Obligatorios**: Comisión porcentual calculada sobre la Remuneración Asegurable del afiliado cobrada por la AFP como retribución por la administración de los Aportes Obligatorios del afiliado.

14. **Comisión por la Gestión de los Aportes Voluntarios**: Comisión porcentual calculada sobre el saldo de aportes voluntarios del afiliado como retribución por la administración de los Aportes Voluntarios.

15. **Cotizantes**: Afiliados que han registrado aportes a su Cuenta Individual de Capitalización en el mes, dichos aportes pueden corresponder a pagos oportunos o pagos de aportes devengados en meses anteriores.

16. **Cuenta Individual de Capitalización (CIC):** Cuenta conformada por la Libreta de Capitalización AFP y Libreta Complementaria de Capitalización AFP donde se registran los aportes obligatorios y voluntarios, y las ganancias derivadas de tales aportes

17. **Egresos del Fondo de Pensiones para el pago de Pensiones**: Rubro del cuadro Ingresos y Egresos del Fondo de Pensiones que comprende entre otros los siguientes conceptos: las salidas del Fondo originadas por pagos mensuales de Retiro Programado o Renta Temporal y transferencias de Cuentas Individuales a Empresas de Seguro para la provisión de rentas vitalicias.

18. **Otros (Ingresos y Egresos del Fondo de Pensiones):** Rubro del cuadro Ingresos y Egresos del Fondo de Pensiones conformado por los siguientes conceptos: Gastos de Sepelio, Retiro de Ahorro Voluntario, Retiro Hereditario, Excedente de Pensión, Otros ingresos y Otros Egresos.

19. **Encaje Legal:** Recursos de propiedad de la AFP conformado por un porcentaje de las inversiones del Fondo de Pensiones. Se determina de acuerdo con la categoría de riesgo de los valores en los que se encuentra invertido el Fondo de Pensiones. Sirve como garantía de la rentabilidad mínima en caso que la rentabilidad caiga por debajo.

20. **Fondo de Pensiones:** A nivel específico, es el conjunto de Cuentas Individuales de Capitalización administradas bajo los criterios determinados por el Plan de Inversiones de cada Tipo de Fondo del que se trate. A nivel general, el Fondo de Pensiones es el conjunto de Fondos que una AFP administra con excepción de los Fondos Voluntarios para personas jurídicas.

21. **Gastos de Sepelio:** El monto referido al promedio de gastos de sepelio que cubre el SPP ascenderá hasta un límite de S/. 2,500 (dos mil quinientos nuevos soles), el que se actualizará trimestralmente en función al Índice de Precios al Consumidor (IPC) que

elabora el Instituto Nacional de Estadística e Informática (INEI), o el indicador que lo sustituya, tomando como base para el referido índice el número que arroje para el mes de junio de 1998.

22. **Ingresos y Egresos del Fondo de Pensiones:** Cuadro que presenta los rubros que explican la variación del Fondo de Pensiones para un mes determinado, información consignada de modo exacto en el Anexo II del Informe Diario de Inversiones.

23. **Invalidez Parcial:** Grado de invalidez bajo el cual el afiliado registra una pérdida de su capacidad de trabajo en un porcentaje igual o superior al 50% pero inferior a los dos tercios, conforme al dictamen emitido por la Comisión Médica de la Superintendencia (COMEC) o la Comisión Médica de las AFP (COMAFP).

24. **Invalidez Total:** Grado de invalidez bajo el cual el afiliado sufre la pérdida en su capacidad de trabajo en un porcentaje igual o superior a dos tercios, conforme al dictamen emitido por la COMEC o la COMAFP.

25. **Invalidez Temporal:** Naturaleza de invalidez dictaminada por el COMAFP o COMEC bajo la cual el afiliado puede recuperar su capacidad para trabajar.

26. **Invalidez Permanente:** Naturaleza de invalidez dictaminada por el COMAFP o COMEC bajo la cual un afiliado no puede recuperar su capacidad para trabajar.

27. **Inversiones Locales:** Inversiones realizadas en instrumentos financieros emitidos por el Gobierno Central, el Banco Central de Reserva, Gobiernos Regionales, Municipalidades y en empresas que realizan sus actividades mayoritariamente en el país o inviertan los recursos captados de la cartera administrada en una proporción mayor al 50% en actividades económicas realizadas en el país.

28. **Inversiones en el Exterior:** Inversiones realizadas en instrumentos emitidos por entidades cuyos activos se concentran en 50% o más, en actividades económicas realizadas en el exterior.

29. **Jubilación por Edad Legal:** Jubilación que procede cuando el afiliado alcanza los sesenta y cinco años de edad (65), cumplidos en meses y días, al momento de presentar la solicitud de pensión de jubilación.

30. **Jubilación Anticipada – Régimen Ordinario:** Beneficio al que tiene derecho el afiliado que, no habiendo cumplido con los requisitos establecidos para percibir pensión de jubilación por edad legal, obtenga una pensión igual o superior al 50% del promedio de las remuneraciones percibidas y rentas declaradas durante los últimos ciento veinte (120) meses anteriores al mes de la presentación de la solicitud de pensión de

jubilación, actualizadas sobre la base del Índice de Precios al Consumidor de Lima Metropolitana, o el indicador que lo sustituya.

31. **Jubilación Anticipada - Régimen Extraordinario:** Régimen de carácter transitorio por el cual el Estado reconoce al trabajador un beneficio extraordinario por los aportes efectuados durante su permanencia en el SNP realizando trabajo de riesgo para su vida o salud. La Jubilación por este Régimen procede para los trabajadores que cumplen las condiciones establecidas en el acápite I del artículo 4° del Reglamento de la Ley N° 27252, aprobado por Decreto Supremo N° 164-2001-EF.

32. **Jubilación Anticipada - Régimen Genérico:** Establece la jubilación anticipada para los afiliados que realizan trabajos que implican riesgo para su vida o salud y que no cumplen con los requisitos señalados en el Régimen Extraordinario. En este caso, el empleador y el afiliado realizan aportes complementarios a su (CIC), determinados en función de su edad de jubilación. Podrán acceder a este Régimen los trabajadores que cumplen las condiciones establecidas en el acápite II del artículo 4° del Reglamento de la Ley N° 27252, aprobado por Decreto Supremo N° 164-2001-EF.

33. **Régimen Especial de Jubilación Anticipada para Desempleados - Ley N° 27617 (REJ):** Régimen transitorio que permite la jubilación a los afiliados mayores de 55 años, desempleados por doce o más meses, siempre que cumplan los requisitos establecidos en el artículo 9° de la Ley N° 27617, que incorpora la Decimotercera Disposición Final y Transitoria al Texto Único Ordenado de la Ley del Sistema Privado de Administración de Fondos de Pensiones. Este régimen tendrá vigencia hasta el 1° de diciembre de 2005.

34. **Régimen Especial de Jubilación Anticipada para Desempleados - Ley N° 28991 (REJ):** Régimen transitorio que permite la jubilación a los afiliados mayores de 55 años, desempleados por doce o más meses, siempre que cumplan los requisitos establecidos en el artículo 17° de la Ley N° 28991, que incorpora la Decimotercera Disposición Final y Transitoria al Texto Único Ordenado de la Ley del Sistema Privado de Administración de Fondos de Pensiones. Este régimen tendrá vigencia hasta el 31 de diciembre de 2008.

35. **Régimen Especial de Jubilación Anticipada para Desempleados - Ley N° 29426 (REJ):** Régimen transitorio que permite la jubilación a los afiliados varones de 55 años y mujeres de 50 años, al momento de presentación de la solicitud, que se encuentren desempleados por doce o más meses, siempre que cumplan los requisitos establecidos en el artículo 1° de la Ley N° 29426. Este régimen tendrá vigencia hasta el 31 de diciembre de 2012.

36. **Jubilación Adelantada del Decreto Ley N° 19990:** Régimen que establece la jubilación adelantada para los afiliados que al momento de su incorporación al SPP

cumplían con los requisitos para acceder a Jubilación Adelantada en el SNP. Dichos afiliados pueden jubilarse adelantadamente en el SPP, siempre que cumplan con lo establecido en el artículo 9° de la Ley N° 27617, que incorpora la Decimoquinta Disposición Final y Transitoria al Texto Único Ordenado de la Ley del Sistema Privado de Administración de Fondos de Pensiones.

37. **Modalidad de Pensión:** Se denomina modalidad de pensión a cada una de las distintas formas de pago de pensión que la normativa del SPP contempla para la provisión de beneficios. Las modalidades de pensión existentes son: Retiro Programado (en soles), Renta Vitalicia Familiar (en soles o dólares), Renta Mixta (en soles y dólares), Renta Bimoneda (en soles y dólares) y Renta Temporal (en soles) con Renta Vitalicia Diferida (en soles o dólares)

38. **Monto Máximo por Gastos de Sepelio:** Monto máximo que cubre el SPP por los gastos del sepelio de un afiliado, siempre que se encuentre comprendido bajo la cobertura del seguro, conforme lo dispuesto en los artículos 64º y 65º del Título VII del Compendio de Normas Reglamentarias del SPP. En caso un afiliado no tenga la cobertura del seguro, los gastos de sepelio se cubren con cargo a los recursos existentes en la CIC del afiliado. El monto límite de gastos de sepelio que cubre el SPP asciende a S/. 2,500 (dos mil quinientos nuevos soles), el mismo que se actualiza trimestralmente en función al Índice de Precios al Consumidor (IPC) que elabora el Instituto Nacional de Estadística e Informática (INEI), o el indicador que lo sustituya, tomando como base -para el referido índice- el número que arroje para el mes de junio de 1998.

39. **Número de nuevos pensionistas:** Corresponde al número de pensionistas que han percibido pensiones pagadas por primera vez en el mes informado, independientemente del mes de devengue del pago. No incluye a los pensionistas que cambian de modalidad de pensión.

40. **Número de Pensionistas:** Corresponde al número total de pensionistas que perciben pensión en el mes informado.

41. **Oferta Privada:** Oferta de acciones a personas o instituciones antes de cotizar en una Bolsa de Valores y a la cual no tiene acceso el público en general.

42. **Operaciones en Tránsito:** Concepto constituido por las cuentas por pagar y cobrar de la Cartera Administrada

43. **Pensión:** Es la prestación que otorga el SPP a sus afiliados y, de ser el caso, a sus beneficiarios. Las pensiones son de jubilación, invalidez y sobrevivencia, siendo otorgadas por las AFP o las empresas de seguros, según corresponda.

44. **Pensión de Invalidez:** Pensión que se otorga con carácter transitorio o definitivo a aquellos afiliados que, sin haber optado aún por el goce de una pensión de jubilación, presentan una pérdida mayor o igual al 50% de su capacidad de trabajo.

45. **Pensión de Jubilación:** Se otorga desde el momento en que el afiliado alcanza los sesenta y cinco (65) años de edad o antes si es que el afiliado cumple con los requisitos y condiciones establecidos para acceder a alguno de los regímenes de jubilación anticipada. La jubilación es un acto voluntario del afiliado.

46. **Pensión de Sobrevivencia:** Es aquella que se otorga a los beneficiarios de un afiliado activo o pasivo luego del fallecimiento de éste.

47. **Pensiones de Sobrevivencia por Afiliados Activos:** Cuando al fallecimiento del afiliado éste no se encontraba percibiendo pensión de jubilación ni pensión de invalidez definitiva.

48. **Pensiones de Sobrevivencia por Afiliados Pasivos:** Cuando al fallecimiento del afiliado éste se encontraba percibiendo pensión de jubilación o de invalidez definitiva.

49. **Pensión Mínima - Ley N° 27617:** Es un beneficio creado mediante la Ley N° 27617 y representa una garantía que brinda el Estado peruano a aquellos trabajadores que cumpliendo con requisitos de años de aporte y requisitos de edad, no alcanzan a tener una pensión que supere el mínimo establecido. Podrán acceder a una pensión mínima aquellas personas que cumplan con los siguientes requisitos: Haber nacido a más tardar el 31/12/1945, contar con un mínimo de 65 años de edad, y que no se encuentren percibiendo una pensión de jubilación al momento de presentar su solicitud ante la AFP; haber realizado un mínimo de veinte (20) años completos de aportaciones efectivas en total, entre el Sistema Nacional de Pensiones (SNP) y/o el SPP y que las mencionadas aportaciones hayan sido calculadas sobre la base de la Remuneración Mínima Vital, en cada oportunidad.

50. **Pensión Mínima - Ley N° 28991:** Es un beneficio creado mediante la Ley Nº 28991 para los afiliados al Sistema Privado de Pensiones (SPP) que pertenecieron al Sistema Nacional de Pensiones (SNP) al momento de la creación del SPP, estos afiliados podrán gozar de una Pensión Mínima de jubilación equivalente en términos anuales a la que reciben los afiliados al SNP. Los afiliados al SPP que accedan a esta pensión deberán cumplir los mismos requisitos del SNP y pagar el diferencial de aportes respectivos, según las condiciones del artículo 7° de la Ley N° 28991.

51. **Pensión Pagada:** Término que se aplica cuando se emite el cheque o se realiza el depósito del monto correspondiente a la pensión -sea ésta por jubilación (legal o anticipada), invalidez o sobrevivencia- en la cuenta del pensionista.

52. **Pensión Promedio:** Monto de pensión referencial calculado como el cociente entre el monto de pensiones que corresponden a Pagos regulares mes de pago y el número de las pensiones pagadas. Los Pagos regulares mes de pago corresponden a pagos devengados y efectuados únicamente en el periodo informado (no incluye reintegros por efectos de regularización de pensiones preliminares del 80% al 100%).

53. **Prima de Seguro:** Porcentaje cobrado sobre la remuneración asegurable del trabajador con el objetivo de adquirir derecho al financiamiento de las prestaciones de invalidez, sobrevivencia y gastos de sepelio en la eventual ocurrencia de alguno de estos siniestros.

54. **Promedio Ponderado de la Comisión Variable:** Promedio ponderado de las Comisiones Variables cobradas por cada AFP de acuerdo al número de Cotizantes. La información de cotizantes utilizada en el cálculo tiene un mes de rezago respecto de la información de las Comisiones Variables.

55. **Promedio Ponderado de la Prima de Seguro:** Promedio ponderado de las Primas de Seguro cobradas por cada AFP de acuerdo al número de Cotizantes. La información de cotizantes utilizada en el cálculo tiene un mes de rezago respecto de la información de las Primas de Seguro.

56. **Promotores:** Son aquellas personas naturales que habiendo celebrado contratos de trabajo con una AFP, participan en la incorporación de trabajadores al Sistema Privado de Pensiones. Para ello, el contrato de trabajo deberá estar inscrito en el Registro de Promotores de AFP de la SBS.

57. **Remuneración Asegurable:** Es el total de las rentas provenientes del trabajo personal del afiliado percibidas en dinero, cualquiera que sea la categoría de renta a que deban atribuirse según las normas tributarias sobre renta.

58. **Remuneración Máxima Asegurable:** Monto máximo fijado trimestralmente por la Superintendencia, sobre el cual se calcula la prima del seguro de invalidez, sobrevivencia y gastos de sepelio. Dicha remuneración máxima asciende a S/. 3,000 (tres mil nuevos soles) de mayo de 1993, y se reajusta mediante el Índice de Precios al Consumidor (IPC) de Lima Metropolitana que elabora el Instituto Nacional de Estadística e Informática (INEI), o el indicador que lo sustituya, con la periodicidad que establezca la Superintendencia.

59. **Rentabilidad Ajustada por Riesgo:** Indica el retorno que ha obtenido un Fondo de Pensiones por cada unidad de riesgo asumida por su portafolio de inversiones y se obtiene de dividir el promedio simple de la rentabilidad nominal diaria durante los últimos 12 meses entre la desviación estándar de dichas rentabilidades calculada para el mismo período (riesgo). Cabe señalar que cuanto más alto resulte la rentabilidad ajustada por riesgo significará una mejor gestión del portafolio del fondo de Pensiones y viceversa.

60. **Rentabilidad Nominal Anual de la Cartera Administrada:** Mide el rendimiento nominal de las inversiones realizadas con los recursos del Fondo de Pensiones y del Encaje Legal durante un año. Se construye como la variación porcentual entre el valor cuota promedio de un mes específico y el valor cuota promedio del mismo mes del año anterior.

61. **Rentabilidad Real Anual de la Cartera Administrada:** Mide el rendimiento real anual de las inversiones realizadas con los recursos del Fondo de Pensiones y del Encaje Legal. Se construye como la rentabilidad nominal anual deflactada por la inflación del período.

62. **Renta Vitalicia Familiar:** Modalidad de pensión mediante la cual el afiliado o sus beneficiarios contratan directamente con la Empresa de Seguros una renta mensual a ser pagada hasta el fallecimiento del afiliado y el pago posterior de pensiones de sobrevivencia a sus beneficiarios. La contratación de una renta vitalicia implica la cesión irrevocable de la Cuenta Individual de Capitalización a favor de la empresa de seguros elegida, motivo por el cual los fondos que no se lleguen a utilizar para el pago de pensiones no constituyen herencia.

63. **Renta Temporal con Renta Vitalicia Diferida:** Modalidad de pensión mediante la cual el afiliado retiene en su Cuenta Individual de Capitalización los fondos suficientes para obtener de la AFP una Renta Temporal y, adicionalmente, contrata una Renta Vitalicia Familiar, con la finalidad de recibir pagos mensuales a partir de una fecha determinada (período diferido).

64. **Renta Mixta:** Modalidad de pensión por la cual el afiliado o beneficiario –con una parte del saldo de la Cuenta Individual de Capitalización (CIC)- contrata el pago de una renta mensual a cargo de una empresa de seguros bajo la modalidad de renta vitalicia familiar en dólares americanos, en tanto que con el fondo que permanezca en la CIC se otorgará una pensión bajo la modalidad de retiro programado. En este caso, la pensión total corresponde a la suma de los montos de pensión percibidos por cada una de las modalidades. Sólo pueden acogerse a esta modalidad los afiliados que, en el proceso de cotizaciones de esta modalidad puedan obtener una renta vitalicia inmediata en dólares americanos equivalente –cuando menos- al valor de la pensión mínima anualizada que garantiza el Estado a los afiliados del SPP.

65. **Renta Vitalicia Bimoneda:** Modalidad de pensión por la cual el afiliado contrata dos Rentas Vitalicias de manera simultánea: una en moneda nacional y la otra en dólares americanos, ambas otorgadas por la misma empresa de seguros. En este caso, la pensión total corresponderá a la suma de los montos percibidos por cada una de las monedas.

66. **Retiro Programado:** Modalidad de pensión mediante la cual el afiliado, manteniendo la propiedad sobre los fondos acumulados en su CIC, efectúa retiros mensuales contra el saldo de dicha cuenta hasta que la misma se extinga. El Retiro Programado tiene

carácter revocable y los fondos no utilizados para el pago de pensiones constituyen herencia siempre y cuando no queden beneficiarios.

67. **Solicitud de Bono por Emisión Ordinaria:** Solicitud presentada ante la ONP sin haber presentado paralelamente alguna solicitud de beneficios.

68. **Solicitud de Bono por Emisión y Redención Simultánea:** Solicitud presentada por el afiliado o sus beneficiarios ante la ONP y que registra, al momento de su presentación, alguna de las siguientes causales de redención: por jubilación anticipada, por invalidez total permanente o por fallecimiento del titular. Este tipo de solicitud implica la presentación paralela de alguna solicitud de beneficios ante la AFP.

69. **Solicitudes de Bono de Reconocimiento Presentadas a la ONP**: Solicitudes presentadas por las AFP ante la ONP luego de verificar el cumplimiento de los requisitos formales para tener derecho a Bono de Reconocimiento.

70. **Tasa de Aporte Obligatorio**: Porcentaje a deducir mensualmente de la remuneración del trabajador y cuyo destino es la Cuenta Individual de Capitalización. Dicho porcentaje es fijado mediante Ley y actualmente es 10% de la remuneración.

71. **Tasa de Interés Técnico:** La tasa de interés técnico (tasa de descuento) es la tasa a ser utilizada por la AFP para el cálculo del capital requerido unitario bajo la modalidad de Retiro Programado, deberá ser fijada libremente por las AFP, según los parámetros que establezca la Superintendencia.

72. **Tasa de Cotización de Rentas Vitalicias (o Tasa de Venta):** Es la tasa utilizada por la empresa de seguros para calcular el capital requerido de las pensiones bajo la modalidad de renta vitalicia a otorgar al afiliado y/o sus beneficiarios, de ser el caso. Esta tasa es fijada libremente por las empresas de seguros.

73. **Tipos de Fondo:** Los tipos de Fondo son tres: Fondo 1, o de preservación del capital, el cual presenta una baja volatilidad; Fondo 2, o mixto, de una volatilidad media; y Fondo 3, o de crecimiento, de una alta volatilidad.

74. **Tipos de Jubilación:** Son los distintos regímenes mediante los cuales puede accederse al beneficio de pensión de jubilación en el SPP: Jubilación por Edad Legal, Jubilación Anticipada – Régimen Ordinario, Jubilación Anticipada – Régimen Extraordinario, Régimen Genérico, Régimen Especial para Desempleados (Ley 27617, Ley 28991 y Ley 29426) y Régimen de Jubilación Adelantada del Decreto Ley 19990.

75. **Tipo de Trabajador:** Criterio de clasificación que define a los afiliados activos como dependientes o independientes sobre la base de la última información reportada por la AFP. Debe entenderse como afiliados dependientes, a aquellos trabajadores que registran una relación laboral vigente; y a los afiliados independientes, como aquellos que no registran relación laboral activa. No se establece la naturaleza de los ingresos que perciben los afilados.

76. **Traspaso:** Proceso que implica el traslado voluntario de la Cuenta Individual de Capitalización de una AFP (AFP de Origen) a otra (AFP de Destino) y que se inicia con la presentación de la respectiva Solicitud de Traspaso.

77. **Traspaso Efectivo:** Culminación del proceso de traspaso que se materializa cuando la AFP de origen transfiere la CIC del afiliado a la AFP de destino.

78. **Movimiento del Fondo de Pensiones por Traspasos Efectivos:** Montos de dinero desplazados de una AFP a otra como resultado de la entrada y salida de Cuentas Individuales de Capitalización por procesos de Traspaso.

79. **Traspaso Neto:** Resultado agregado del proceso de entradas y salidas por traspasos al final de un período. Los traspasos netos pueden medirse en número de afiliados o en términos de Fondo de Pensiones.

80. **Traslados:** Proceso que implica el movimiento de los aportes voluntarios, con fin previsional y/o sin fin previsional, de una AFP (AFP de Origen) a otra (AFP de Destino) y que se inicia con la presentación de la respectiva Solicitud de Traslado.

81. **Traslado Efectivo:** Culminación del proceso de traslado que se materializa cuando la AFP de origen transfiere los aportes voluntarios del afiliado a la AFP de destino.

82. **Movimiento del Fondo de Pensiones por Traslados Efectivos:** Montos de dinero desplazados de una AFP a otra como resultado de la entrada y salida de los Aportes Voluntarios por procesos de Traslado.

83. **Utilidad (Pérdida) del Encaje:** Comprende el resultado obtenido por las operaciones realizadas con el Fondo del Encaje Legal.

84. **Utilidad (Pérdida) del Fondo Complementario:** Comprende el resultado obtenido por las operaciones realizadas con el Fondo Complementario. El Fondo Complementario es aquel que una AFP debe constituir en caso opte por la administración directa de los riesgos de Invalidez y Sobrevivencia y estará integrado por los aportes que corresponda hacer a los afiliados para tener cobertura por los riesgos de invalidez, sobrevivencia y gastos de sepelio. Actualmente esta figura no está operativa.

85. **Utilidad (Pérdida) del Fondo de Longevidad:** Comprende el resultado obtenido por las operaciones realizadas con el Fondo de Longevidad. De acuerdo a la Ley del SPP, el Fondo de Longevidad se constituiría con la utilización de los saldos de las Cuentas Individuales de los afiliados que contrataron la modalidad de Renta Vitalicia Personal que hayan fallecido. En tanto la modalidad de Renta Vitalicia Personal no se encuentra operativa, tampoco el Fondo de Longevidad.

86. **Valor Cuota:** Unidad de cuenta del Sistema Privado de Pensiones de valor variable calculado diariamente por la siguiente fórmula:

$$VC_T = \frac{(Activo_T - PasivoExigible_T)}{Número total de Cuotas_T}$$

El activo está conformado por los instrumentos de inversión autorizados por Ley y que son adquiridos con los recursos del Fondo de Pensiones. El pasivo exigible está compuesto por las prestaciones de los afiliados, retiros de aportes voluntarios, traspasos por pagar, entre otras cuentas. El número total de cuotas corresponde a las cuotas que han sido adquiridas por todos los afiliados con sus aportes.

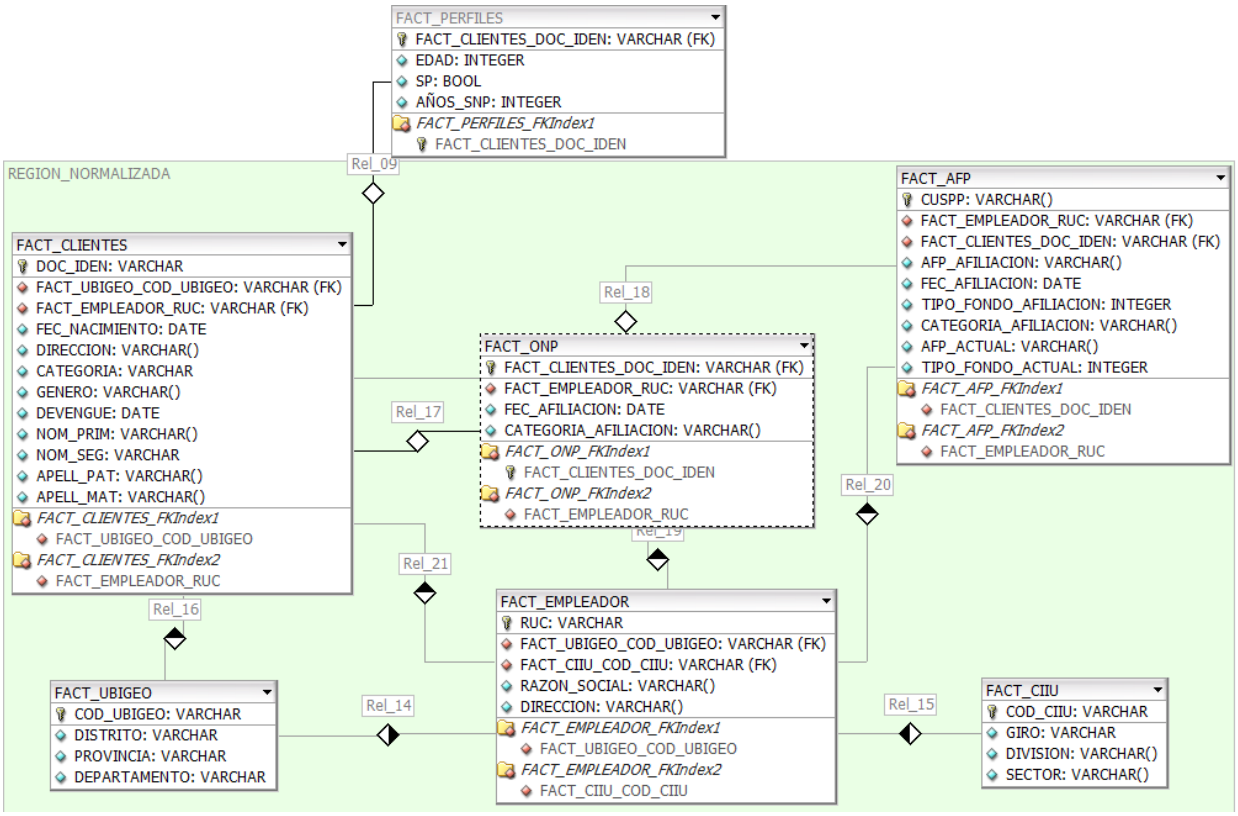
87. **Valor Cuota Ajustado:** Es el valor cuota utilizado en el cálculo de la rentabilidad nominal de los Fondos de Pensiones, el mismo que no incluye la rentabilidad generada por las inversiones que superan los límites de inversión (excesos de inversión imputables).

88. **Valor Cuota Promedio:** Promedio mensual de los valores cuota diarios, considerando sólo los valores cuota de los días hábiles, sobre la base del cual se calcula la rentabilidad de los fondos de pensiones.

Fuente:
Glosario de Términos del Sistema Privado de Pensiones. (2013, Agosto16). Recuperado 20:00, Agosto 16, 2013, de SBS (Superintendencia de Banca y Seguros y AFP): http://www.sbs.gob.pe/app/stats/Glosarios/Glosario_SPP_(Mayo2010).DOC

ANEXO B

DIAGRAMA DE BASE DE DATOS



ANEXO C

PRINCIPALES MODELOS PARA MINERÍA DE DATOS

Los modelos de **conglomerado** se centran en la identificación de grupos de registros similares.

Algoritmo de K-medias:

Agrupa conjuntos de datos en grupos distintos (o conglomerados). El método define un número fijo de conglomerados, de forma iterativa asigna registros a los conglomerados y ajusta los centros de los conglomerados hasta que no se pueda mejorar el modelo. En lugar de intentar pronosticar un resultado, los modelos de k-medias utilizan un proceso conocido como aprendizaje no supervisado para revelar los patrones del conjunto de campos de entrada.

Los modelos de **asociación** asocian una determinada conclusión (como, por ejemplo, la decisión de comprar algo) con un conjunto de condiciones:

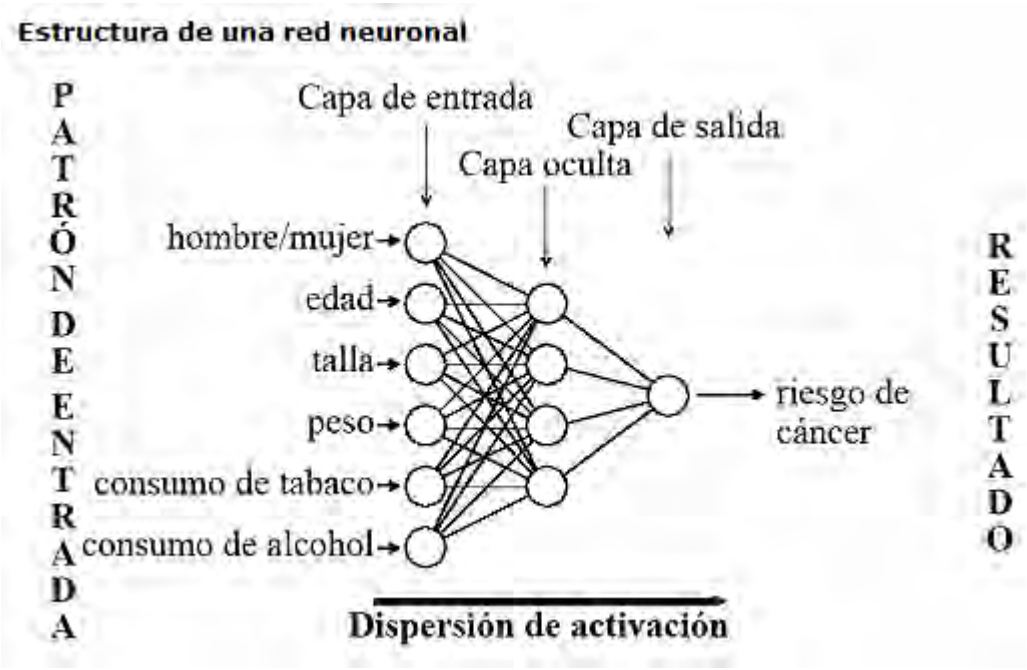
Algoritmo de Inducción de reglas generalizado (GRI):

Es capaz de encontrar las reglas de asociación existentes en los datos. Por ejemplo, los clientes que compran cuchillas y loción para después del afeitado también suelen comprar crema de afeitado. GRI extrae reglas con el mayor nivel de contenido de información basándose en un índice que tiene en cuenta tanto la generalidad (soporte) como la precisión (confianza) de las reglas.

Los modelos de **árboles de decisión** permiten desarrollar sistemas de clasificación que pronostican o clasifican observaciones futuras basadas en un conjunto de reglas de decisión.

Algoritmo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas las cuales pueden ser no binarias, lo que significa que algunas divisiones tendrán más de dos ramas. Los campos objetivo y predictor pueden ser de rango o categóricos.

Los modelos de **redes neuronales** son modelos simples del funcionamiento del sistema nervioso. Las unidades básicas son las neuronas, que generalmente se organizan en capas, como se muestra en la siguiente ilustración.



Fuente: Temas de Ayuda – Clementine 11.1

Una red neuronal, a menudo denominada perceptrón multicapa, es básicamente un modelo simplificado del modo en que el cerebro humano procesa la información. Funciona simultaneando un número elevado de unidades simples de procesamiento interconectadas que parecen versiones abstractas de neuronas.

Las unidades de procesamiento se organizan en capas. Existen, generalmente, tres capas en una red neuronal: una capa de entrada, con unidades que representan los campos de entrada; una o varias capas ocultas; y una capa de salida, con unidades que representan los campos de salida. Las unidades se conectan con fuerzas de conexión variables (o ponderaciones). Los datos de entrada se presentan en la primera capa, y los valores se propagan desde cada neurona hasta cada neurona de la capa siguiente. Al final, se envía un resultado desde la capa de salida.

La red aprende examinando los registros individuales, generando un pronóstico para cada registro y realizando ajustes a las ponderaciones cuando realiza un pronóstico incorrecto. Este proceso se repite muchas veces y la red sigue mejorando sus pronósticos hasta haber alcanzado uno o varios criterios de parada.

Al principio, todas las ponderaciones son aleatorias y las respuestas que resultan de la red son, posiblemente, disparatadas. La red aprende a través del entrenamiento. Continuamente se presentan a la red ejemplos para los que se conoce el resultado, y las respuestas que proporciona se comparan con los resultados conocidos. La información procedente de esta comparación se pasa hacia atrás a través de la red, cambiando las ponderaciones gradualmente. A medida que progresa el entrenamiento, la red se va haciendo cada vez más precisa en la replicación de resultados conocidos. Una vez entrenada, la red se puede aplicar a casos futuros en los que se desconoce el resultado.

Ejemplo. Al cribar subvenciones para el desarrollo agrícola para posibles casos de fraude, se puede utilizar una red neuronal para explorar en profundidad las desviaciones de la norma, resaltando aquellos registros que sean anómalos y dignos de una investigación más detallada. En particular, le interesan aquellas solicitudes de subvenciones que parezcan reclamar demasiado dinero teniendo en cuenta el tipo y tamaño de la granja. Demostración

Requisitos. No se aplican restricciones a los tipos de campo. Los nodos Red neuronal pueden gestionar entradas y salidas numéricas, simbólicas o de marcas. El nodo Red neuronal espera uno o varios campos con dirección Entrada y uno o varios campos con dirección Salida. Se ignorarán los campos establecidos en Ambos o Ninguno. Los tipos de campo deben estar completamente instanciados al ejecutar el nodo.

Puntos fuertes. Las redes neuronales son dispositivos eficaces de cálculo de funciones generales. Por lo general, realizan al menos las tareas de pronóstico y otras técnicas, y su rendimiento puede mejorar significativamente en determinadas ocasiones. También se precisa un conocimiento matemático o estadístico mínimo para entrenarlas o aplicarlas.

Los **modelos estadísticos** utilizan ecuaciones matemáticas para codificar información extraída de los datos.

La regresión lineal es una técnica de estadístico común utilizada para resumir datos y realizar pronósticos ajustando una superficie o línea recta que minimice las discrepancias existentes entre los valores de salida reales y los pronosticados.

La regresión logística es una técnica de estadístico para clasificar los registros en función los valores de los campos de entrada. Es análoga a la regresión lineal pero toma un campo objetivo categórico en lugar de uno numérico. Hay de dos tipos:

- Binomial. Se utiliza cuando el campo objetivo es un conjunto o una marca con dos valores discretos (dicotómicos), como sí/no, encendido/apagado, hombre/mujer.

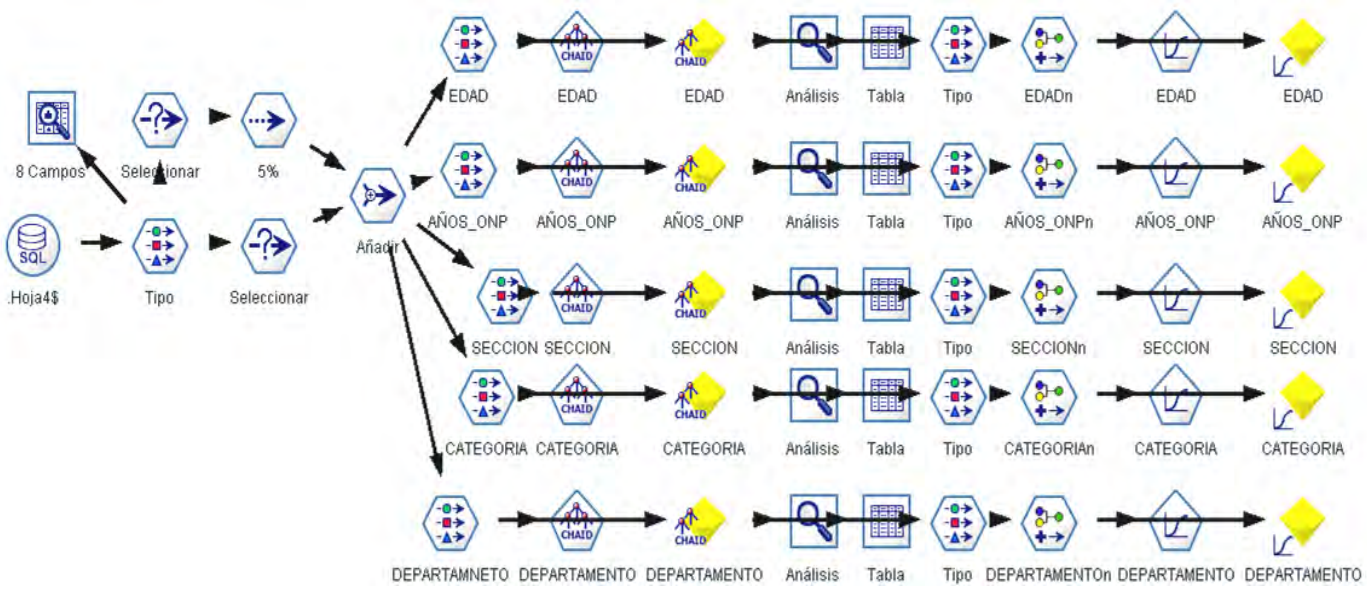
- Multinomial. Se utiliza cuando el campo objetivo es un campo de conjuntos con más de dos valores. Puede especificar Efectos principales, Factorial completo o Personalizada.

Modelos de **serie temporal** producen pronósticos de rendimiento futuro de datos de series temporales existentes. Estima modelos de suavizado exponencial, modelos autorregresivos integrados de media móvil (ARIMA) univariados y modelos ARIMA (o de función de transferencia) multivariados para series temporales y genera datos de predicciones.

Fuente:
Temas de Ayuda – Clementine 11.1

ANEXO D

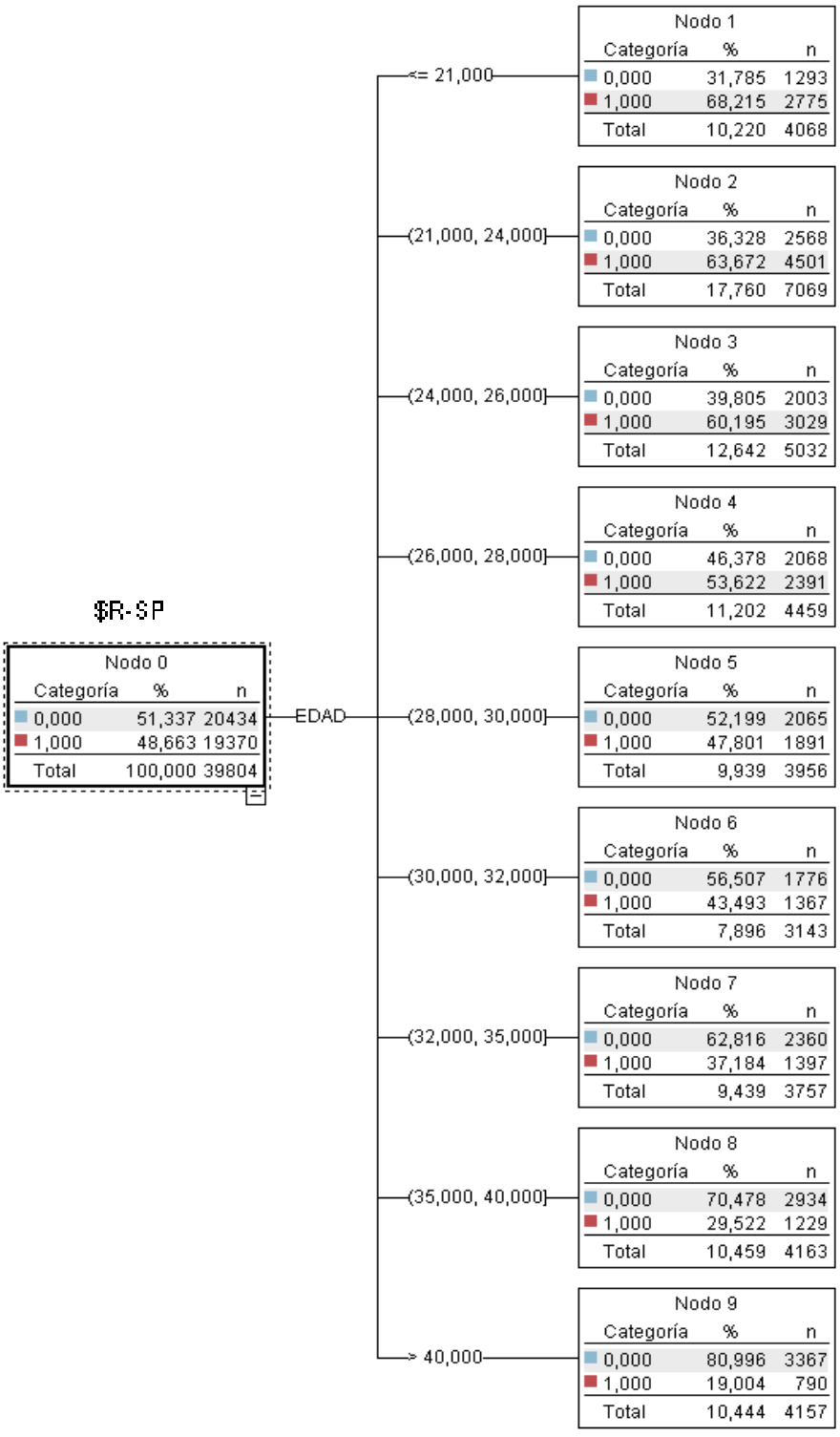
ESQUEMA EN SPSS CLEMENTINE PARA TRANSFORMACIÓN DE VARIABLES INDEPENDIENTES



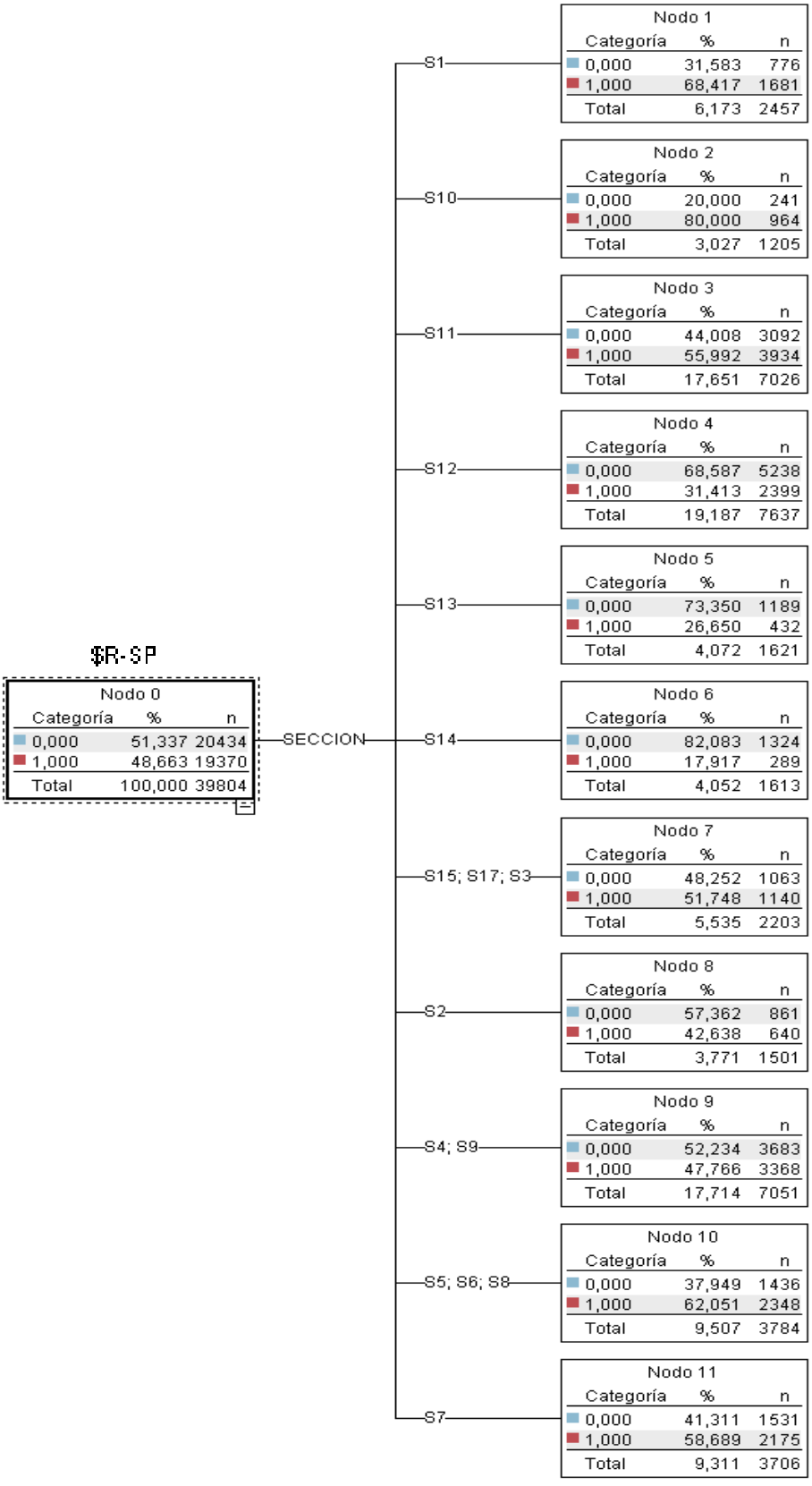
ANEXO E

RECATEGORIZACIÓN DE VARIABLES INDEPENDIENTES

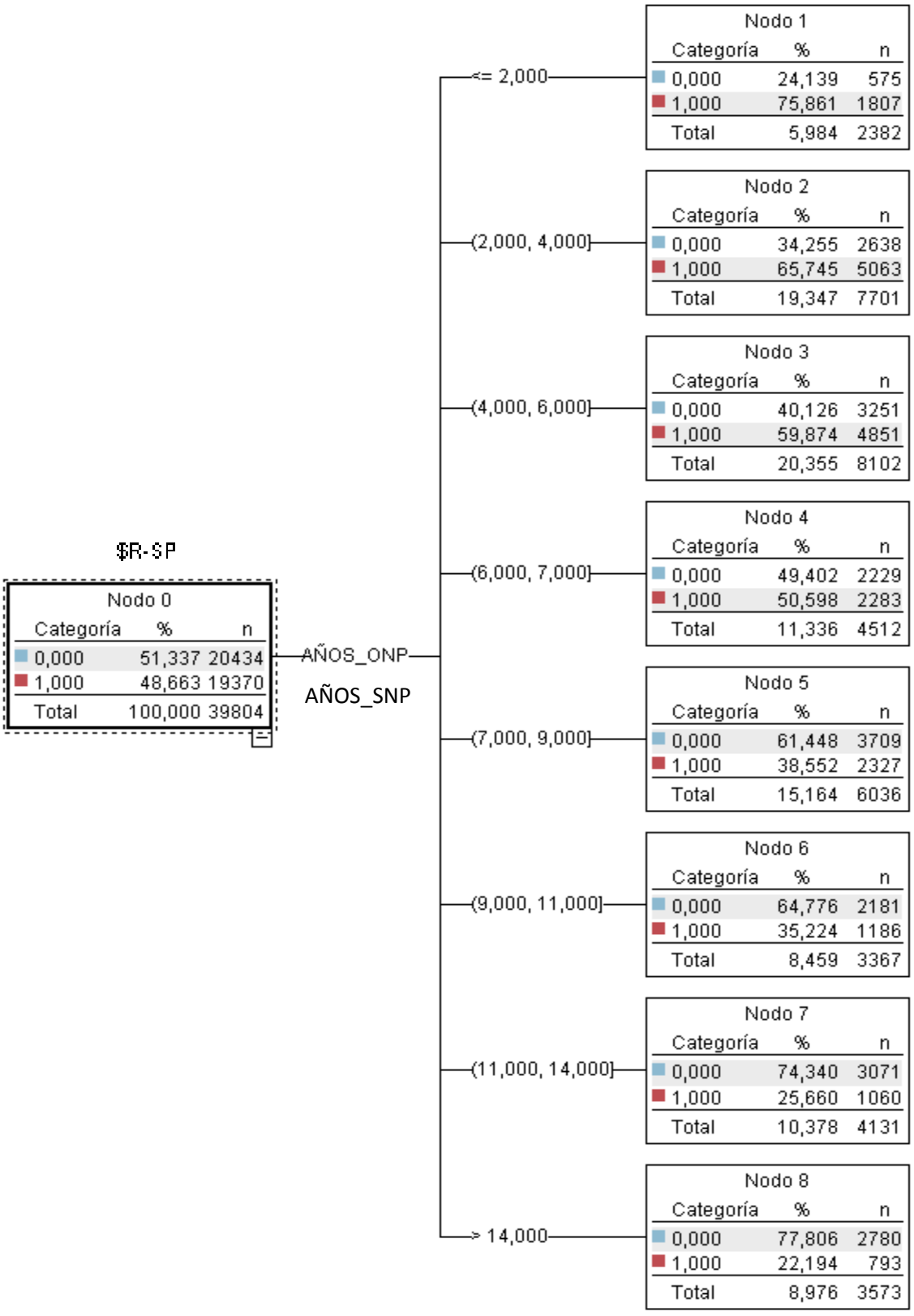
a.) Recategorización de variable “EDAD”



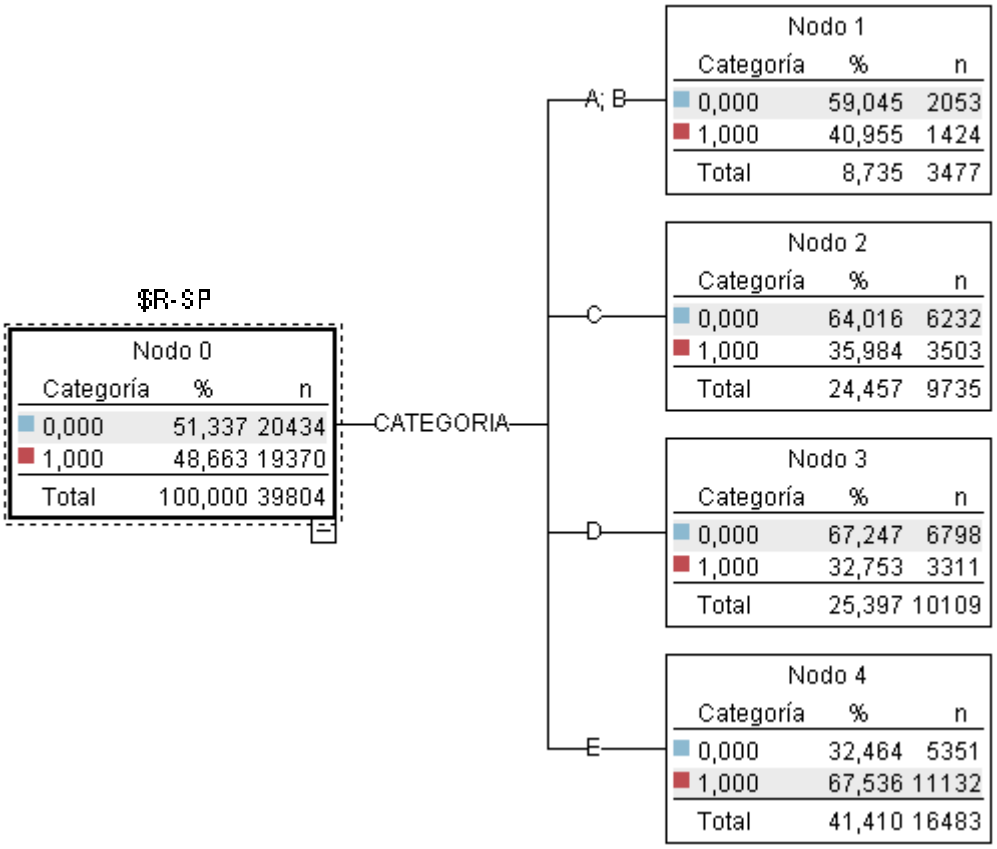
b.) Recategorización de variable “SECCION”



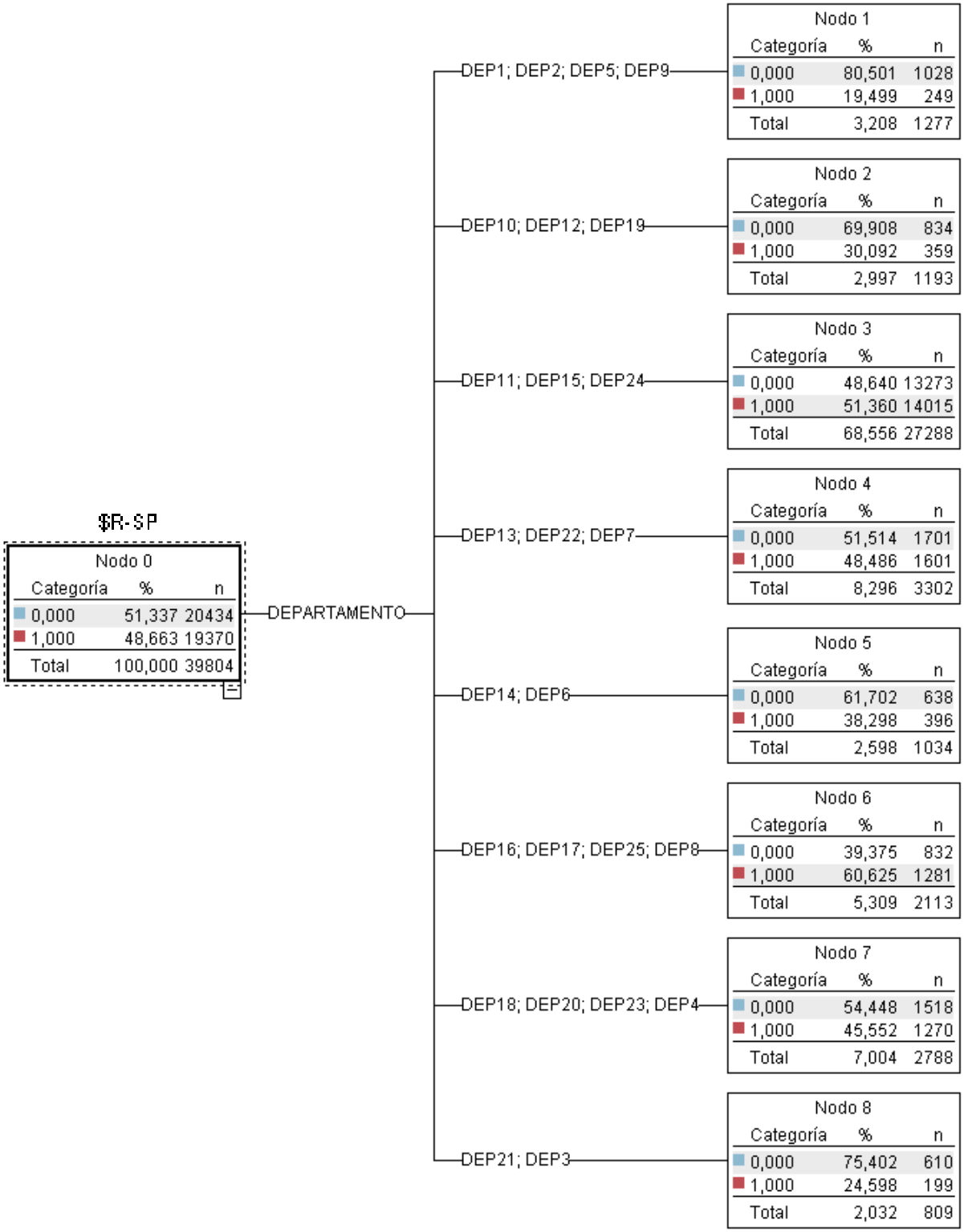
c.) Recategorización de variable “AÑOS_SNP”



d.) Recategorización de variable “CATEGORIA”



e.) Recategorización de variable “DEPARTAMENTO”



ANEXO F

CLASIFICACIÓN INTERNACIONAL INDUSTRIAL UNIFORME

La Clasificación Internacional Industrial Uniforme (siglas: CIIU) o, en inglés, *International Standard Industrial Classification of All Economic Activities* (abreviada como ISIC), es la clasificación sistemática de todas las actividades económicas cuya finalidad es la de establecer su codificación armonizada a nivel mundial. Es utilizada para conocer niveles de desarrollo, requerimientos, normalización, políticas económicas e industriales, entre otras utilidades.

Cada país tiene, por lo general, una clasificación industrial propia, en la forma más adecuada para responder a sus circunstancias individuales y al grado de desarrollo de su economía. Puesto que las necesidades de clasificación industrial varían, ya sea para los análisis nacionales o para fines de comparación internacional. La Clasificación Internacional Industrial Uniforme de todas las Actividades Económicas (CIIU) permite que los países produzcan datos de acuerdo con categorías comparables a escala internacional.

La CIIU desempeña un papel importante al proporcionar el tipo de desglose por actividad necesario para la compilación de las cuentas nacionales desde el punto de vista de la producción.

La última versión de la clasificación, la ISIC Rev.4, fue lanzada oficialmente el 11 de agosto de 2008.

Propósitos:

- Su propósito principal es ofrecer un conjunto de categorías de actividades que se pueda utilizar cuando se diferencian las estadísticas de acuerdo con esas actividades.

- El propósito secundario de la CIIU es presentar ese conjunto de categorías de actividad de modo tal que las entidades se puedan clasificar según la actividad económica que realizan.

Estructura:

El nivel superior de la clasificación está compuesto por las siguientes secciones:

- A - Agricultura, silvicultura y pesca
- B - Explotación de minas y canteras
- C - Industrias manufactureras
- D - Suministro de electricidad, gas, vapor y aire acondicionado
- E - Suministro de agua; alcantarillado, gestión de desechos y actividades de saneamiento
- F - Construcción

- G - Comercio al por mayor y al por menor; reparación de los vehículos de motor y de las motocicletas
- H - Transporte y comunicación
- I - Alojamiento y servicios de comida
- J - Información y comunicación
- K - Actividades financieras y de seguros.
- L - Actividades inmobiliarias
- M - Actividades profesionales, científicas y técnicas
- N - Actividades administrativas y servicios de apoyo
- O - Administración pública y defensa; planes de seguridad social de afiliación obligatoria
- P - Enseñanza
- Q - Servicios sociales y relacionados con la salud humana.
- R - Artes, entretenimiento y recreación
- S - Otras actividades de servicio
- T - Actividades de los hogares en calidad de empleadores, actividades indiferenciadas de producción de bienes y servicios de los hogares para uso propio.
- U - Actividades de organizaciones y órganos extraterritoriales

Las correspondencias entre la clasificación por sectores económicos y la de CIIU:

TEORÍA	**PRÁCTICA**
SECTOR	SECCIÓN
SUBSECTOR	DIVISIÓN
RAMA DE ACTIVIDAD	GRUPO
ACTIVIDAD	CLASE

Fuente:

Naciones Unidas (2005) Clasificación Industrial Internacional Uniforme de todas las actividades económicas (CIIU) Serie M, No 4, Revisión 3.1.Recuperado de: http://unstats.un.org/unsd/publication/SeriesM/seriesm_4rev3_1s.pdf

ANEXO G

CURVA ROC Y MATRIZ DE CONFUSION

Terminología y sus derivados
a partir de una matriz de confusión valign=top

Verdaderos Positivos (VP)

o también éxitos

Verdaderos Negativos (VN)

o también rechazos correctos

Falsos Positivos (FP)

o también falsas alarmas o Error tipo I

Falsos Negativos (FN)

o también, Error de tipo II

sensibilidad o Razón de Verdaderos Positivos (VPR)

o también razón de éxitos y, recuerdo en recuperación de información,

$$VPR = VP/V = VP/(VP + FN)$$

Ratio o Razón de Falsos Positivos (FPR)

o también razón de falsas alarmas o fall-out en recuperación de información

$$FPR = FP/N = FP/(FP + VN)$$

Precisión o Exactitud (*accuracy*) (ACC)

$$ACC = (VP + VN)/(P + N)$$

Especificidad (SPC) o Razón de Verdaderos Negativos

$$SPC = VN/N = VN/(FP + VN) = 1 - FPR$$

Valor Predictivo Positivo (PPV)

o también "precision" en recuperación de información

$$PPV = VP/(VP + FP)$$

Valor Predictivo Negativo (NPV)

$$NPV = VN/(VN + FN)$$

Ratio o Razón de Falsos Descubrimientos (FDR)

$$FDR = FP/(FP + VP)$$

En la Teoría de detección de señales una curva ROC (acrónimo de Receiver Operating Characteristic, o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a (1 – especificidad) para un sistema clasificador binario según se varía el umbral de discriminación. Otra interpretación de este gráfico es la representación de la razón o ratio de verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón o ratio de falsos positivos (FPR = Razón de Falsos Positivos) también según se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo). ROC también puede significar Relative Operating Characteristic (Característica Operativa Relativa) porque es una comparación de dos características operativas (VPR y FPR) según cambiamos el umbral para la decisión (3). En español es preferible mantener el acrónimo inglés, aunque es posible encontrar el equivalente español COR. No se suele utilizar ROC aislado, debemos decir "curva ROC" o "análisis ROC". Sobre la historia del acrónimo ROC consultar Swets (1996).

El análisis de la curva ROC, o simplemente análisis ROC, proporciona herramientas para seleccionar los modelos posiblemente óptimos y descartar modelos subóptimos independientemente de (y antes de especificar) el coste de la distribución de las dos clases sobre las que se decide. La curva ROC es también independiente de la distribución de las clases en la población (en diagnóstico, la prevalencia de una enfermedad en la población). El análisis ROC se relaciona de forma directa y natural con el análisis de coste/beneficio en toma de decisiones diagnósticas.

La curva ROC se desarrolló por ingenieros eléctricos para medir la eficacia en la detección de objetos enemigos en campos de batalla mediante pantallas de radar, a partir de lo cual se desarrolla la Teoría de Detección de Señales (TDS). El análisis ROC se aplicó posteriormente en medicina, radiología, psicología y otras áreas durante varias décadas. Sólo recientemente ha encontrado aplicación en áreas como aprendizaje automático (o machine learning en inglés), y minería de datos (data mining en inglés).

Un modelo de clasificación (clasificador o Clasificadores (matemático) o Diagnóstico) es una función que permite decidir cuáles de un conjunto de instancias están relacionadas o no por pertenecer a un mismo tipo o clase. El resultado del clasificador o del diagnóstico puede ser un número real (valor continuo), en cuyo caso el límite del clasificador entre cada clase debe determinarse por un valor umbral (por ejemplo para determinar si una persona tiene hipertensión basándonos en una medida de presión arterial), o puede ser un resultado discreto que indica directamente una de las clases.

Consideremos un problema de predicción de clases binario, en la que los resultados se etiquetan positivos (p) o negativos (n). Hay cuatro posibles resultados a partir de un clasificador binario como el propuesto. Si el resultado de una exploración es p y el valor dado es también p, entonces se conoce como un Verdadero Positivo (VP); sin embargo si el valor real es n entonces se conoce como un Falso Positivo (FP). De igual modo, tenemos un Verdadero Negativo (VN) cuando tanto la exploración como el valor dado son n, y un Falso Negativo (FN) cuando el resultado de la predicción es n pero el valor real es p. Un ejemplo aproximado de un problema real es el siguiente: consideremos una prueba diagnóstica que persiga determinar si una persona tiene una cierta enfermedad. Un falso positivo en este caso ocurre cuando la prueba predice que el resultado es positivo, cuando la persona no tiene realmente la enfermedad. Un falso negativo, por el contrario, ocurre cuando el resultado de la prueba es negativo, sugiriendo que no tiene la enfermedad cuando realmente sí la tiene.

Definamos un experimento a partir de P instancias positivas y N negativas. Los cuatro posibles resultados se pueden formular en una Tabla de contingencia (o Matriz de confusión) 2x2 como sigue:

		Valor en la realidad		
		p	***n***	**total**
Predicción outcome	***p'***	Verdaderos Positivos	Falsos Positivos	P'
	n'	Falsos Negativos	Verdaderos Negativos	N'
	total	P	N	

Tomas Fawcett (2004) ROC Graphs: Notes and Practical Considerations for Researchers, Recuperado de: http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf

El espacio ROC

La tabla de contingencia puede proporcionar varias medidas de evaluación (ver caja de terminología). Para dibujar una curva ROC sólo son necesarias las razones de Verdaderos Positivos (VPR) y de falsos positivos (FPR). La VPR mide hasta qué punto un clasificador o prueba diagnóstica es capaz de detectar o clasificar los casos positivos correctamente, de

entre todos los casos positivos disponibles durante la prueba. La FPR define cuántos resultados positivos son incorrectos de entre todos los casos negativos disponibles durante la prueba.

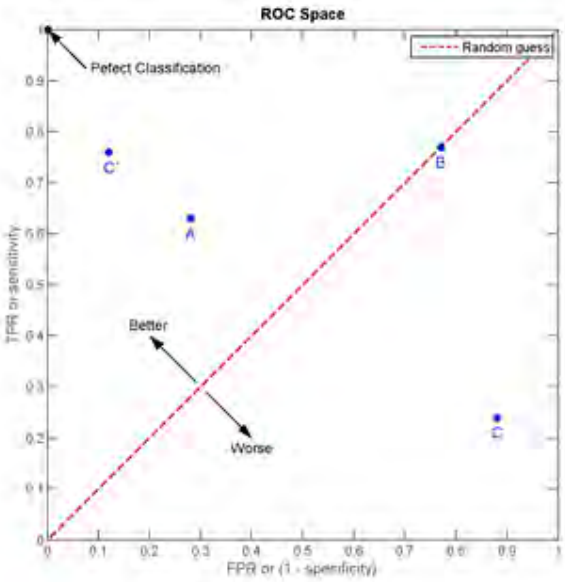
Un espacio ROC se define por FPR y VPR como ejes x e y respectivamente, y representa los intercambios entre verdaderos positivos (en principio, beneficios) y falsos positivos (en principio, costes). Dado que VPR es equivalente a sensibilidad y FPR es igual a 1-especificidad, el gráfico ROC también es conocido como la representación de sensibilidad frente a (1-especificidad). Cada resultado de predicción o instancia de la matriz de confusión representa un punto en el espacio ROC.

El mejor método posible de predicción se situaría en un punto en la esquina superior izquierda, o coordenada (0,1) del espacio ROC, representando un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún falso positivo). A este punto (0,1) también se le llama una *clasificación perfecta*. Por el contrario, una clasificación totalmente aleatoria (o adivinación aleatoria) daría un punto a lo largo de la línea diagnoal, que se llama también *línea de no-discriminación*, desde el extremo inferior izquierdo hasta la esquina superior derecha (independientemente de los tipos de base positiva y negativa). Un ejemplo típico de adivinación aleatoria sería decidir a partir de los resultados de lanzar una moneda al aire, a medida que el tamaño de la muestra aumenta, el punto de un clasificador aleatorio de ROC se desplazará hacia la posición (0.5, 0.5).

La diagonal divide el espacio ROC. Los puntos por encima de la diagonal representan los buenos resultados de clasificación (mejor que el azar), puntos por debajo de la línea de los resultados pobres (peor que al azar). Nótese que la salida de un predictor consistentemente pobre simplemente podría ser invertida para obtener un buen predictor.

Considérense los siguientes cuatros resultados de 100 instancias positivas y otras 100 negativas:

A			B			C			C'		
VP=63	FP=28	91	VP=77	FP=77	154	VP=24	FP=88	112	VP=76	FP=12	88
FN=37	VN=72	109	FN=23	VN=23	46	FN=76	VN=12	88	FN=24	VN=88	112
100	100	200	100	100	200	100	100	200	100	100	200
VPR = 0.63			VPR = 0.77			VPR = 0.24			VPR = 0.76		
FPR = 0.28			FPR = 0.77			FPR = 0.88			FPR = 0.12		
ACC = 0.68			ACC = 0.50			ACC = 0.18			ACC = 0.82		



Tomas Fawcett (2004) ROC Graphs: Notes and Practical Considerations for Researchers, Recuperado de: http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf

El espacio ROC y las parcelas de los cuatro ejemplos de predicción A, B, C y C'.

En la figura se muestran los puntos que los cuatro ejemplos anteriores en el espacio ROC. El resultado del método **A** muestra claramente ser el mejor de entre los métodos **A**, **B** Y**C**. El resultado de **B** se encuentra sobre la línea de estimación aleatoria (diagonal); en la tabla se puede ver que la precisión (**ACC**) de este método es del 50%. El método **C** aparece como el peor de los tres, con un resultado muy pobre.

Sin embargo, consideremos ahora la construcción de un cuarto método de predicción **C'** que simplemente invierte los resultados predichos por el método **C**. Este nuevo método mostrará una tabla de contingencia opuesta a la de **C** y su punto en el espacio ROC estará ahora por encima de la diagonal, y más próximo al punto de *clasificación perfecta* que el método **A**. Mientras **C** presentaba un pobre poder de predicción, a partir de él se ha construido un predictor mejor que todos los demás. Cuando el método **C** predice 'n' o 'p', el método **C'** predice 'p' o 'n' respectivamente. Siempre que un método presente un punto en el espacio ROC por debajo de la diagonal habrá que invertir sus predicciones para aprovechar su capacidad de predicción.

Cuanto más cerca esté un método de la esquina superior izquierda (clasificación perfecta) mejor será, pero lo que en realidad marca el poder predictivo de un método es la distancia de este a la línea de estimación aleatoria, da igual si por arriba o por abajo.

Los clasificadores discretos, como los Árbol de decisión o los sistemas de reglas, dan como resultados a valores numéricos una etiqueta binaria. Cuando se usan estos clasificadores con un conjunto concreto de instancias para clasificar o predecir, el rendimiento del clasificador proporciona un único punto en el espacio ROC. Para otros clasificadores, como un Clasificador bayesianoo una Red neuronal artificial, la salida son valores de probabilidad que representan hasta qué punto una instancia pertenece a una de las dos clases.

Para estos métodos se debe fijar un valor umbral que determinará un punto en el espacio ROC. Por ejemplo, si ante una determinada magnitud fijamos ese umbral en 0.8, la probabilidad de las instancias iguales o superiores serán predichas como positivas, y los valores por debajo serán predichos como negativos. Por tanto podremos calcular una tabla de contingencia (o matriz de confusión) para ese umbral de 0.8, y encontrar el punto correspondiente en el espacio ROC. Según vamos variando el umbral (por ejemplo, en pasos de 0.1) tendríamos una tabla de contingencia y un nuevo punto en el espacio ROC. Dibujar la curva ROC consiste en poner juntos todos los puntos correspondientes a todos los umbrales o puntos de corte, de tal modo que ese conjunto de puntos se parecerá más o menos a una curva en el espacio cuadrado entre (0,0) y (1,1). Dependiendo del tipo de modelo la curva se parecerá más a una escalera (métodos no paramétricos) o una verdadera curva (métodos paramétricos).[3] A medida que desplazamos ese valor umbral, en realidad estamos alterando las tasas de verdaderos positivos (VP) y falsos positivos (FP).

La curva ROC se puede usar para generar estadísticos que resumen el rendimiento (o la efectividad, en su más amplio sentido) del clasificador. A continuación se proporcionan algunos:

- El punto de inserción de la curva ROC con la línea convexa a la línea de discriminación.
- El área entre la curva ROC y la línea de convexo-paralela discriminación.
- El área bajo la curva ROC, llamada comúnmente AUC (*Área Bajo la Curva*). También se puede encontrar denominada A' ("a-prima"),[4] o el estadístico 'c' (*c-statistic*).[5]
- Índice de sensibilidad o d' (*d-prima*, por cierto siempre minúscula). Es la distancia entre la media de la distribución de actividad en el sistema bajo condiciones de sólo ruido y su distribución bajo condiciones de sólo señal, dividido por su desviación típica, bajo el supuesto de que ambas distribuciones son normales con la misma desviación típica. Bajo estos supuestos, se puede probar que la forma de la curva ROC sólo depende de este parámetro d'.

El indicador más utilizado en muchos contextos es el área bajo la curva ROC o AUC. Este índice se puede interpretar como la probabilidad de que un clasificador ordenará o puntuará una instancia positiva elegida aleatoriamente más alta que una negativa. Se puede demostrar que el área bajo la curva ROC es equivalente a la Prueba de Mann-Whitney, una prueba no paramétrica aplicada a dos muestras independientes, cuyos datos han sido medidos al menos en una escala de nivel ordinal. Se trata de una prueba estadística virtualmente idéntica a la realización de una prueba paramétrica ordinaria T de dos muestras en los datos después de haber ordenado las muestras combinadas. Es también equivalente a la Prueba de los signos de Wilcoxon. También se ha demostrado la relación del área bajo la curva ROC con el Coeficiente de Gini, con la siguiente fórmula G1 + 1 = 2xAUC, donde:

$$G_1 = 1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1})$$

Otra forma básica de calcular AUC es usando un promedio de una serie de aproximaciones trapezoidales.

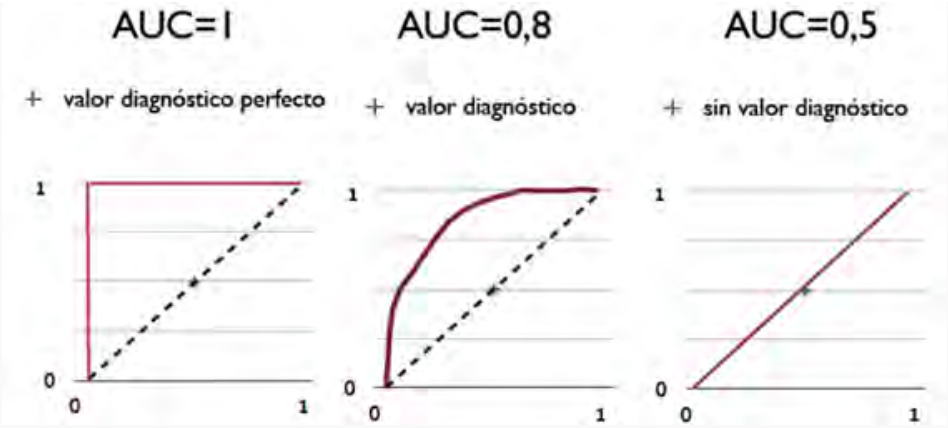
Sin embargo, se ha comentado que este indicador, en general, reducir la curva ROC en varios metros, hace perder información sobre el patrón de intercambios del algoritmo discriminador en cuestión.

La comunidad de aprendizaje automático utiliza el estadístico AUC para la comparación de modelos. En otras áreas de ingeniería se prefiere la medida del área entre la curva ROC y la línea de no-discriminación. Finalmente en Psicofísica se utiliza preferentemente d'.

En resumen: se trata de una medida pura de la eficacia o capacidad predictiva del sistema, independientemente del punto de corte que se utilice, de las reglas de las personas que usen los sistemas predictivos y también, y muy importante, de las tasas de verdaderos positivos en la población (o Prevalencia en contextos de diagnóstico médico).

En ocasiones puede ser más útil mirar a una región específica de la curva ROC más que a toda la curva. Es posible calcular áreas parciales bajo la curva, o AUC parciales. Por ejemplo, nos podríamos concentrar en la región de la curva con razones de falsos positivos más bajas, que es a menudo el interés principal de las pruebas de Detección precoz (o medicine screening) en la población.

Curvas ROC para pruebas diagnósticas



Tomas Fawcett (2004) ROC Graphs: Notes and Practical Considerations for Researchers, Recuperado de: http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf

Diferentes curvas ROC

Para la elección entre dos pruebas diagnósticas distintas, se recurre a las curvas ROC, ya que es una medida global e independiente del punto de corte. Por esto, en el ámbito sanitario, las curvas ROC también se denominan curvas de rendimiento diagnóstico.

La elección se realiza mediante la comparación del área bajo la curva (AUC) de ambas pruebas. Esta área posee un valor comprendido entre 0,5 y 1, donde 1 representa un valor diagnóstico perfecto y 0,5 es una prueba sin capacidad discriminatoria diagnóstica. Es decir, si AUC para una prueba diagnóstica es 0,8 significa que existe un 80% de probabilidad de que el diagnóstico realizado a un enfermo sea más correcto que el de una persona sana escogida al azar. Por esto, siempre se elige la prueba diagnóstica que presente una mayor área bajo la curva.

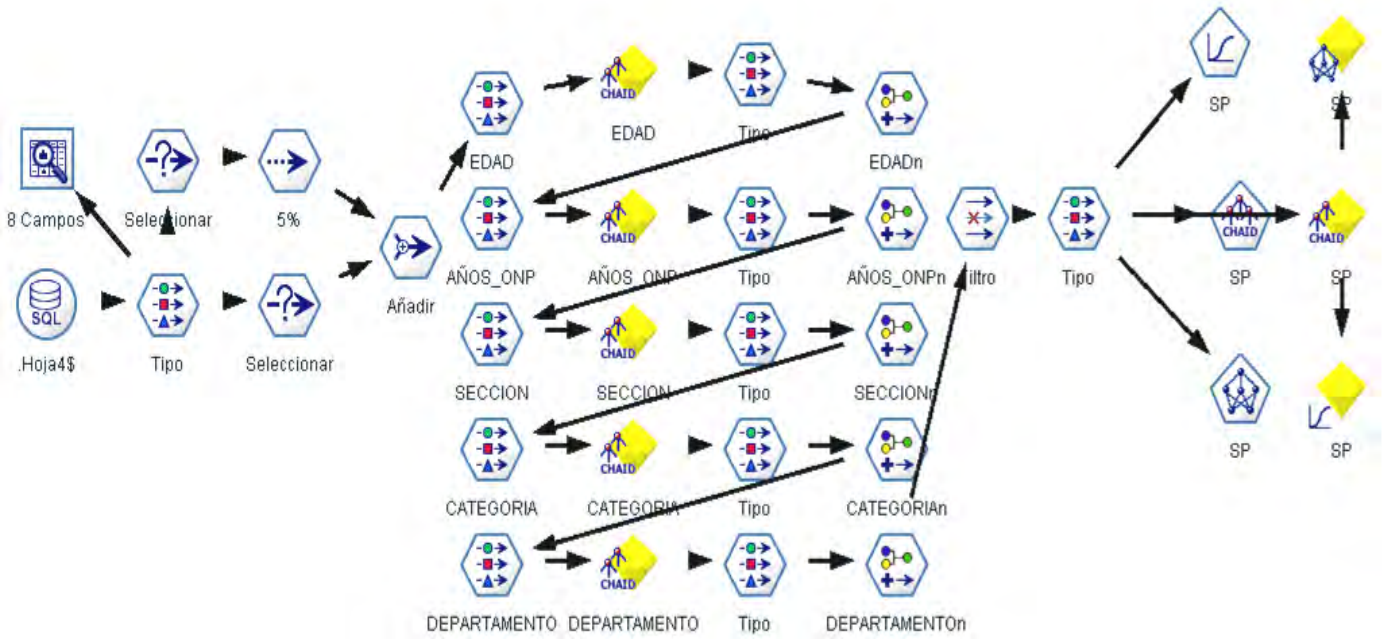
A modo de guía para interpretar las curvas ROC se han establecido los siguientes intervalos para los valores de AUC:

[0.5, 6): Test malo.
[0.6, 0.75): Test regular.
[0.75, 0.9): Test bueno.
[0.9, 0.97): Test muy bueno.
[0.97, 1): Test excelente.

Fuente:
Tomas Fawcett (2004) ROC Graphs: Notes and Practical Considerations for Researchers, Recuperado de: http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf

ANEXO H

ESQUEMA EN SPSS CLEMENTINE DE LOS MODELOS DE MINERÍA DE DATOS



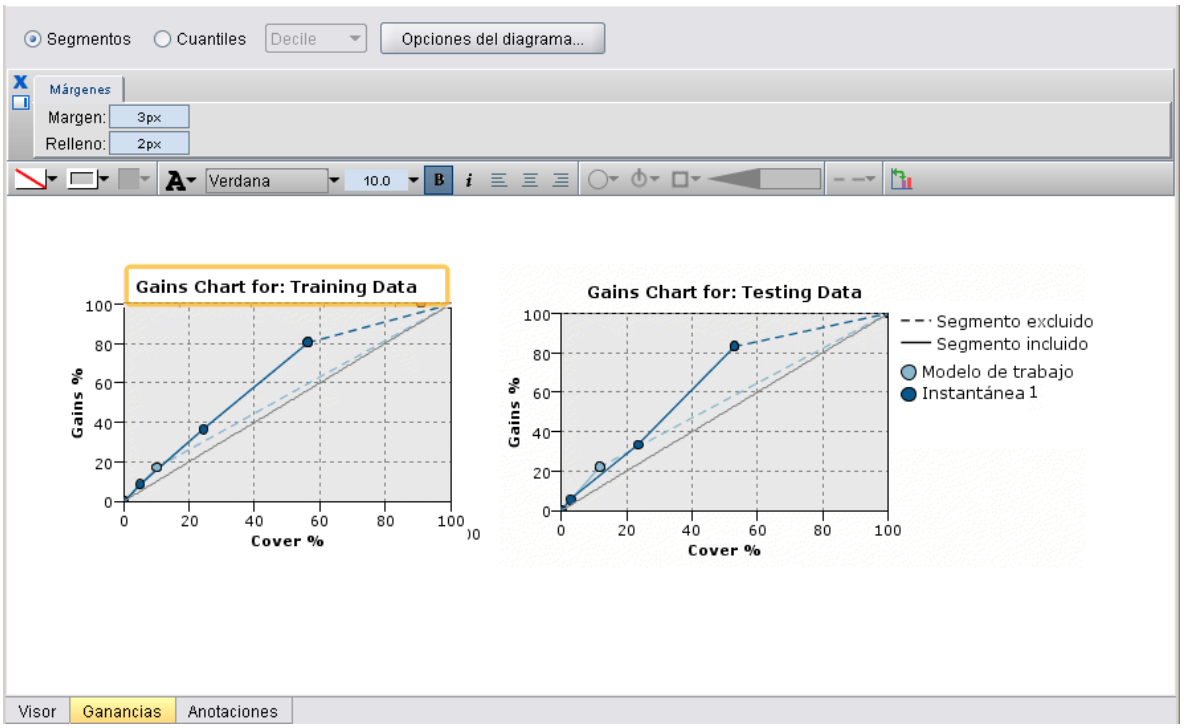
ANEXO I

GRÁFICO DE GANANCIAS

Los gráficos de ganancias representan los valores de la columna % ganancia en la tabla. Las ganancias se definen como la proporción de aciertos en cada uno de los incrementos en relación con el número total de aciertos en el árbol, y se obtienen mediante la ecuación:

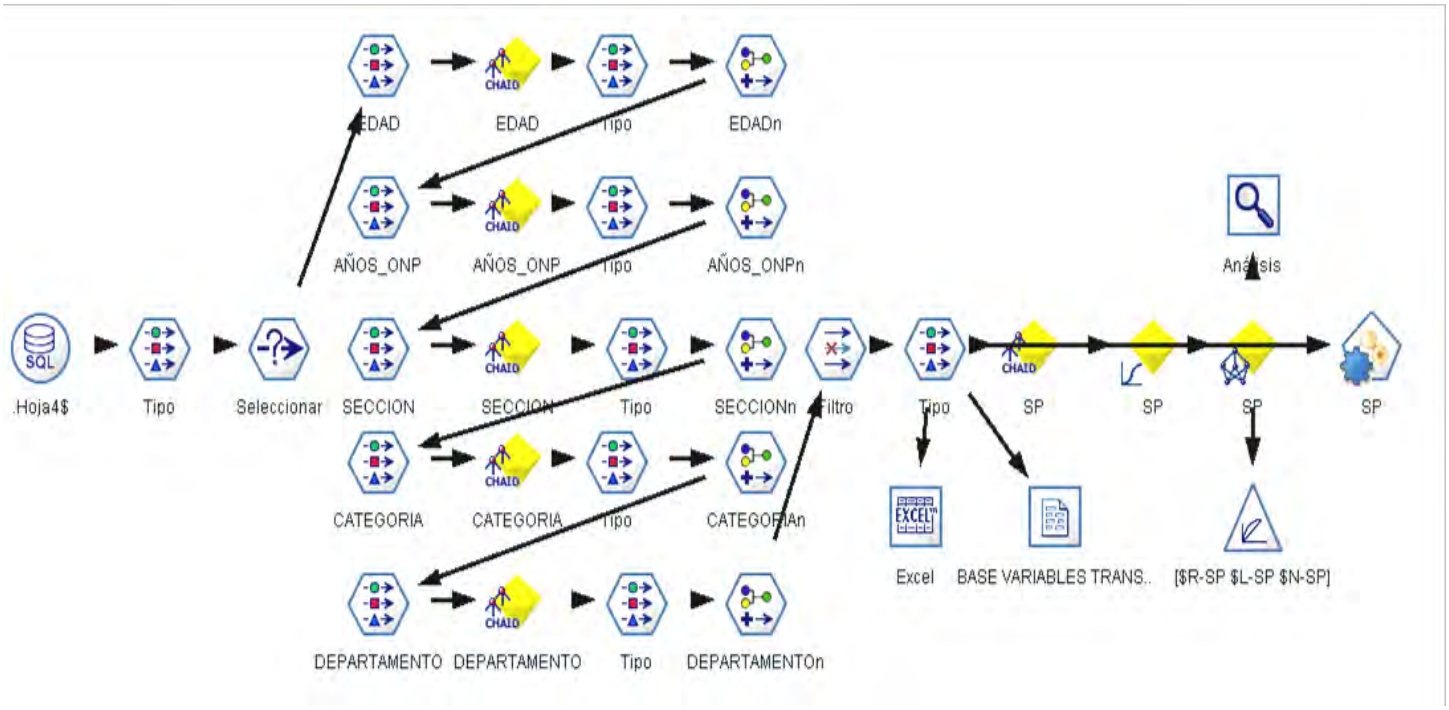
(aciertos del incremento / número total de aciertos) x 100%

El gráfico de ganancias ilustra de manera eficaz la difusión necesaria para una red cuando se desea capturar un porcentaje determinado de todos los aciertos del árbol. La línea diagonal representa la respuesta esperada para la muestra completa, si no se utilizase el modelo. En este caso la tasa de respuesta debería ser constante, ya que una persona tiene la misma probabilidad de responder que otra. Para duplicar los resultados deberá preguntar dos veces al mismo número de personas. La línea curvada indica hasta qué punto se puede mejorar la respuesta incluyendo únicamente elementos situados en los percentiles superiores en función de las ganancias. Por ejemplo, si incluye el 50% superior, obtendrá más del 70% de respuestas positivas. Cuanto más pronunciada es la curva, mayor es la ganancia.



ANEXO J

ESQUEMA EN SPSS CLEMENTINE DEL MODELO FINAL APLICADO AL UNIVERSO DE DATOS



ANEXO K

AFILIACION MENSUAL SISTEMA PRIVADO DE PENSIONES

Fecha	Horizonte	Integra	Prima	Profuturo	Total SPP
oct-08	5,137	4,462	3,658	5,267	18,524
nov-08	4,323	3,504	3,151	3,984	14,962
dic-08	3,843	2,564	2,729	3,337	12,473
ene-09	5,419	3,917	3,217	5,619	18,172
feb-09	4,742	3,499	4,357	4,880	17,478
mar-09	4,747	3,010	3,946	4,622	16,325
abr-09	4,769	2,870	4,342	3,715	15,696
may-09	5,186	3,011	4,056	3,453	15,706
jun-09	4,866	3,269	3,345	3,494	14,974
jul-09	5,225	3,226	3,571	3,514	15,536
ago-09	6,627	3,879	3,721	3,488	17,715
sep-09	6,314	4,583	4,018	3,534	18,449
oct-09	5,878	4,667	3,918	3,196	17,659
nov-09	5,127	3,815	3,475	3,149	15,566
dic-09	4,409	3,691	3,219	2,458	13,777
ene-10	4,995	4,139	3,561	3,270	15,965
feb-10	5,274	3,720	4,221	2,839	16,054
mar-10	5,932	4,153	4,311	2,107	16,503
abr-10	5,559	3,708	4,197	2,286	15,750
may-10	5,492	3,854	4,495	2,323	16,164
jun-10	5,443	4,166	3,842	2,420	15,871
jul-10	5,462	3,973	3,783	2,640	15,858
ago-10	5,813	4,972	4,388	2,787	17,960
sep-10	7,260	6,276	5,421	3,337	22,294
oct-10	6,898	7,314	5,739	3,474	23,425
nov-10	7,459	7,452	6,007	3,785	24,703
dic-10	5,625	6,017	5,704	2,379	19,725
ene-11	6,784	6,900	6,860	3,754	24,298
feb-11	7,263	7,474	7,070	4,244	26,051
mar-11	7,758	7,789	7,280	4,513	27,340
abr-11	6,731	7,157	6,404	3,883	24,175
may-11	7,259	6,697	6,542	4,042	24,540
jun-11	7,405	7,452	6,942	3,751	25,550
jul-11	6,845	7,223	6,991	4,105	25,164
ago-11	7,706	7,748	7,493	4,316	27,263
sep-11	8,122	8,269	8,772	4,579	29,742

DOTACION MENSUAL SISTEMA PRIVADO DE PENSIONES

Fecha	Horizonte	Integra	Prima	Profuturo
oct-08	450	386	407	676
nov-08	455	385	405	697
dic-08	423	295	335	710
ene-09	304	296	271	682
feb-09	261	296	266	579
mar-09	253	249	259	468
abr-09	250	248	242	442
may-09	252	249	239	430
jun-09	251	245	242	434
jul-09	244	246	236	438
ago-09	242	245	228	429
sep-09	235	241	229	424
oct-09	233	241	227	411
nov-09	230	273	217	413
dic-09	236	267	216	409
ene-10	230	265	219	405
feb-10	228	265	226	405
mar-10	226	270	236	407
abr-10	228	298	232	407
may-10	226	281	231	407
jun-10	228	276	239	369
jul-10	231	276	238	333
ago-10	230	277	241	325
sep-10	224	289	233	316
oct-10	236	288	242	304
nov-10	240	291	252	303
dic-10	238	292	270	315
ene-11	232	296	263	302
feb-11	235	295	263	300
mar-11	237	294	270	296
abr-11	239	295	269	300
may-11	240	294	274	297
jun-11	240	294	274	297
jul-11	241	302	276	285
ago-11	239	305	266	267
sep-11	233	306	270	274

PRODUCTIVIDAD MENSUAL SISTEMA PRIVADO DE PENSIONES

Fecha	Horizonte	Integra	Prima	Profuturo
oct-08	11.4	11.6	9.0	7.8
nov-08	9.5	9.1	7.8	5.7
dic-08	9.1	8.7	8.1	4.7
ene-09	17.8	13.2	11.9	8.2
feb-09	18.2	11.8	16.4	8.4
mar-09	18.8	12.1	15.2	9.9
abr-09	19.1	11.6	17.9	8.4
may-09	20.6	12.1	17.0	8.0
jun-09	19.4	13.3	13.8	8.1
jul-09	21.4	13.1	15.1	8.0
ago-09	27.4	15.8	16.3	8.1
sep-09	26.9	19.0	17.5	8.3
oct-09	25.2	19.4	17.3	7.8
nov-09	22.3	14.0	16.0	7.6
dic-09	18.7	13.8	14.9	6.0
ene-10	21.7	15.6	16.3	8.1
feb-10	23.1	14.0	18.7	7.0
mar-10	26.2	15.4	18.3	5.2
abr-10	24.4	12.4	18.1	5.6
may-10	24.3	13.7	19.5	5.7
jun-10	23.9	15.1	16.1	6.6
jul-10	23.6	14.4	15.9	7.9
ago-10	25.3	17.9	18.2	8.6
sep-10	32.4	21.7	23.3	10.6
oct-10	29.2	25.4	23.7	11.4
nov-10	31.1	25.6	23.8	12.5
dic-10	23.6	20.6	21.1	7.6
ene-11	29.2	23.3	26.1	12.4
feb-11	30.9	25.3	26.9	14.1
mar-11	32.7	26.5	27.0	15.2
abr-11	28.2	24.3	23.8	12.9
may-11	30.2	22.8	23.9	13.6
jun-11	30.9	25.3	25.3	12.6
jul-11	28.4	23.9	25.3	14.4
ago-11	32.2	25.4	28.2	16.2
sep-11	34.9	27.0	32.5	16.7

Fuente:

Estadísticas Sistema Privado de Pensiones. (2011, Octubre20). Recuperado 20:00, Octubre 20, 2011, de SBS (Superintendencia de Banca y Seguros y AFP): http://www.sbs.gob.pe/0/modulos/JER/JER_Interna.aspx?ARE=0&PFL=