UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

ANÁLISIS DE DATOS Y BÚSQUEDA DE PATRONES EN APLICACIONES MÉDICAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

ARNOL DAVID GARCÍA UBILLA

PROFESOR GUÍA:
JAIME ORTEGA PALMA

MIEMBROS DE LA COMISIÓN:
TAKESHI ASAHI KODAMA
JORGE AMAYA ARRIAGADA
SUSANA MORALES SILVA

SANTIAGO DE CHILE
JULIO 2015

Resumen

El suicidio en Chile se ha convertido en uno de los problemas más necesarios de hacer frente en salud pública, más aún, si consideramos que la enorme mayoría de las personas que mueren por suicidio presentan algún diagnóstico psiquiátrico y han consultado a un especialista los meses antes de cometer suicidio. Esto, motiva la creación de indicadores y alertas para detectar de forma eficaz y oportuna cuando una persona ingresa a una zona de riesgo suicida.

En el presente trabajo se aborda este problema, definiendo una zona o espectro de riesgo suicida, y generando modelos matemáticos y estadísticos para la detección de pacientes en esta zona de riesgo. Para esto, se utiliza una base de datos de 707 pacientes, consultantes de salud mental, de tres centros de salud distintos de la región metropolitana. La base de datos a su vez contempla 343 variables, incluyendo tanto información sociodemográfica de cada paciente, como también sus respuestas en siete instrumentos clínicos utilizados habitualmente en salud mental (DEQ, STAXI, OQ, RFL, APGAR, PBI Madre y PBI Padre).

Inicialmente la base de datos es depurada eliminando aquellos campos y/o registros con gran porcentaje de valores nulos, mientras que la imputación de valores perdidos se realiza mediante técnicas tradicionales y en algunos casos según el criterio experto, donde se utiliza un método de imputación según valor de subescala para los distintos instrumentos clínicos. Posteriormente, se realiza una reducción de atributos mediante el uso de herramientas estadísticas y provenientes del machine learning. Con esta información, se generan cinco modelos utilizando distintas técnicas y herramientas del ámbito de la minería de datos y machine learning mediante aprendizaje supervisado. Los modelos son generados y calibrados usando el lenguaje estadístico `R`, y se comparan sus resultados mediante cuatro métricas distintas: precisión (o accuracy), sensibilidad, especificidad, y mediante su representación en el espacio ROC.

El modelo o clasificador finalmente propuesto corresponde a un modelo de support vector machine, que permite discriminar cuando un paciente se encuentra en una zona de riesgo suicida. El modelo fue entrenado utilizando un kernel de tipo RBF, y utiliza tan sólo 22 variables predictoras, entregando una precisión aproximada del 78 %, calculada mediante k-validación cruzada de n-folds con $k = 100$ y $n = 10$.

Agradecimientos

En primer lugar quisiera agradecer a quienes han sido mi pilar durante todo este tiempo, y decirles que si hoy estoy acá es gracias a ellos, a mi madre y a mi padre, que me han apoyado en todas las decisiones que he tomado, que siempre han confiado en mi y en quienes siempre he encontrado una palabra de aliento y animo para seguir adelante. Ha sido un proceso largo, pero gratificante, y agradezco haber podido contar con ustedes en todo este proceso. También quiero agradecer a mis hermanos, Michaell y Jocelyn, que también han estado ahí para mi siempre que lo he necesitado. Gracias hermanita por acompañarme y aguantarme, siendo muchas veces no sólo una hermana, sino también una gran amiga.

A mi polola, mi compañera y mi mejor amiga, sin ti Heidi nada hubiera sido igual, te agradezco por todo el apoyo, la paciencia y sobretodo por siempre animarme a dar lo mejor de mi y a superarme día a día, dándome fuerzas cuando más lo he necesitado. Todo este tiempo a tu lado ha sido maravilloso y espero que sigamos teniendo muchos más momentos felices como los vividos hasta ahora.

Quiero dar las gracias a todos los que han estado conmigo y me han apoyado durante este largo proceso, desde mis amigos y profesores de la infancia, que con su apoyo me hicieron creer que con esfuerzo y dedicación todo es posible, hasta aquellos amigos y profesores en la Universidad con quienes viví gratos momentos y que me enseñaron lo necesario para estar hoy empezando una nueva etapa de mi vida. A todos y cada uno de ellos les doy las gracias.

También quiero agradecer a mis profesores guía Jaime Ortega y co-guía Takeshi Asahi, por todo el apoyo entregado, por su preocupación y guía a lo largo de este trabajo. Y por último dar las gracias a Susana Morales, Jorge Barros y Orietta Echavarri, que junto al Centro de Modelamiento Matemático de la Universidad de Chile, me dieron la oportunidad de ser partícipe de este interesante trabajo, enmarcado en un proyecto con financiamiento otorgado por el Fondo Nacional de Desarrollo Científico y Tecnológico FONDECYT N° 11121390 y 1111012, y del Instituto Milenio para la Investigación en Depresión y Personalidad -MIDAP- Código Proyecto: IS130005. Junto con el Patrocinio del Depto. de Psiquiatría de la P. Universidad Católica de Chile. Gracias por su apoyo y orientación a lo largo de todo este trabajo, ha sido un verdadero placer haber trabajado en conjunto con personas de tal intelecto y por sobretodo, gran calidad humana.

Tabla de contenido

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

El objetivo general de esta memoria es desarrollar indicadores de riesgo para la clasificación del riesgo suicida en pacientes consultantes a salud mental. Dada la importancia que tiene este problema dentro de la salud pública, más aún, si se considera que la enorme mayoría de las personas que mueren por suicidio presentan algún diagnóstico psiquiátrico y han consultado a un especialista, resulta evidente la necesidad de optimizar los métodos de detección, prevención, evaluación e intervención con las personas en riesgo. Así, en este trabajo se aborda el problema de la detección del riesgo suicida, generando un modelo de minería de datos para la predicción del riesgo suicida. En los párrafos siguientes se expone en mayor detalle el alcance e impacto de este problema de salud, principalmente dentro de Chile.

En Chile, el suicidio se ha convertido en uno de los problemas más acuciantes y necesarios de hacer frente en salud pública. La tasa de suicidio en Chile ha alcanzado a 13,3 individuos por cada 100.000 habitantes (MINSAL, 2013 [40]). Esta tasa ha experimentado un crecimiento de un 90 % entre los años 1990 y 2011 (OECD, 2014 [42]), situando a Chile en el lugar número 13 entre los países miembros de la Organización para la Cooperación y Desarrollo (OCDE). En el caso de los adolescentes chilenos el suicidio representa un 60 % del total de muertes, y se estima que los intentos de suicidio son hasta 20 veces más frecuentes, lo que ha llevado a anticipar un aumento en años de vida perdidos por discapacidad y en la carga de enfermedad en los adultos de las próximas décadas. Ante esta alarmante situación, el MINSAL estableció la meta de disminuir en un 15 % las tasas de suicidio adolescente en la década 2011-2020 (MINSAL, 2013b [41]).

No solo a nivel nacional se hace frente a un relevante problema de Salud Pública, sino también alrededor del mundo, donde muere una persona cada 40 segundos producto de un acto suicida, esto es cerca de 804.000 personas al año, lo que se traduce en una tasa de 11,4 por 100.000 habitantes, siendo 15 por 100.000 habitantes en hombres y 8 por 100.000 habitantes en mujeres (World Health Organization, 2014b [51]).

Este problema se amplifica si consideramos que por cada suicidio consumado se estima que

existen hasta 25 intentos de suicidio (Goldsmith, Pellmar, Kleinman & Bunney, 2002 [26]), lo que nos habla del sufrimiento de muchas personas. Más aun considerando que por cada persona que intenta suicidio, o se suicida, seis se ven afectadas directamente y este número aumenta si ocurre en un lugar de trabajo o institución donde podría afectar incluso a cientos.

Según la OPS (2010) [16], más del 90 % de los suicidios consumados se asocian a desórdenes mentales como depresión y abuso de sustancias. Es por esto que la depresión y el intento de suicidio previo se consideran como los predictores más robustos asociados al suicidio (Appleby, Shaw, Amos, Mc Donnel, Harris, Mc Cann et al., 1999 [3]; Baader, Richter & Mundt, 2004 [5]; Beautrais, 2009 [6]; Harris & Barraclough, 1997 [29]; Isometsa & Lönqvist, 1998 [31]; Kaplan & Sadock, 1999 [33]; Maris, Berman & Silverman, 2000 [39]). Sin embargo, el suicidio se relaciona con múltiples y complejos factores socioculturales, sobre todo, ligados a situaciones críticas de ámbitos socioeconómicos, familiares e individuales. Por lo tanto, existe consenso que en la conducta suicida participan factores de diversa índole: neurobiológicos, biográficos, caracterológicos, vulnerabilidades psicológicas, trastornos psiquiátricos y de la personalidad, estresores psicosociales, desesperanza y acceso a medios de autodestrucción, entre otros (American Psychiatric Association, 2003 [4]; Gabbard, 2002).

Se trata de un problema donde participan múltiples factores, que se configuran en forma única y en un momento determinando en cada caso, lo que lleva a que no sea posible detectar el momento y circunstancias exactas en las que pueda ocurrir un intento de suicidio. Sin embargo, sí resulta posible detectar cuando un paciente se encuentra en una zona de riesgo suicida. Es decir, en una configuración determinada de variables, que lo ponen en riesgo de suicidio, lo cual haría posible realizar acciones preventivas.

También se cuenta con el antecedente que hasta el 80 % de las personas que mueren por suicidio dan algún tipo de aviso (Bobes, Giner & Saiz, 2011 [25]) y la mayoría de ellas han consultado a un servicio de salud entre 3 meses a 1 año antes de su muerte (Horowitz, Bridge, Teach, Ballard, Klima, Rosenstein, et al., 2012 [30]; Appleby, Shaw, Amos, McDonnel, Harris, McCann, et al., 1999) [3]. En Chile, el riesgo suicida es motivo de consulta cada vez más frecuente en hospitales generales (Florenzano, Labra, Fasani, San Juan, Reynal & Quevedo, 2007 [20]).

Estas alarmantes cifras exigen a ir más allá de la búsqueda de factores asociados. Se cuenta a la fecha con abundante investigación, que ha buscado disminuir la incidencia del suicidio, aportando al conocimiento de factores protectores y de riesgo, información descriptiva y epidemiológica (Roberts, Roberts, & Xing, 2010 [44]; Salvo, Melipillán, & Castro, 2009 [47]). Aun existiendo estos importantes logros de investigación; todavía es un desafío realizar predicciones con mayor certeza del momento en que el paciente ingresa en una zona de riesgo suicida. Esto se relaciona con el hecho que nos encontramos frente a un complejo fenómeno, que si bien puede reunir características comunes, es multifactorial y afecta en forma única e individual a las personas (Ryan, Nielssen, Paton, & Large, 2010 [46]).

Con estos antecedentes resulta ineludible la magnitud y gravedad del problema del suicidio y al mismo tiempo queda en evidencia que la forma en que se está enfrentando, y los intentos de prevenirlo, no están dando los resultados esperados.

1.1. Objetivos

A lo largo de este trabajo se busca generar modelos de detección del riesgo suicida utilizando herramientas estadísticas y de minería de datos. Para esto se define el riesgo suicida como:

Definición 1.1 (Riesgo suicida) *Se define como la posibilidad de realizar un intento de suicidio o de pensar activamente en hacerlo.*

A partir de los análisis de los instrumentos clínicos y variables demográficas con las técnicas de minería de datos, se propondrá una configuración de variables, que ubican al consultante en una "zona de riesgo suicida", basado en los siguientes grupos:

Definición 1.2 (Grupo con conducta suicida) *Consultantes a servicios de salud mental, por intento suicida de baja, alta severidad y/o ideación suicida.*

Definición 1.3 (Grupo sin conducta suicida) *Consultantes a servicios de salud mental, sin conducta ni ideación suicida.*

Este modelo predictivo es de utilidad para la pesquisa del riesgo y posibilita la intervención oportuna. En los aspectos clínicos, el conocimiento de estos factores, puede permitir además de la pesquisa, realizar acciones preventivas de psicoeducación y fortalecimiento de los factores protectores.

1.1.1. Objetivo general

Así, el objetivo de este trabajo se resume como a continuación:

"Determinar un modelo predictivo de riesgo suicida en cuanto a variables individuales, familiares y sociales en consultantes a servicios de salud mental de la región metropolitana"

Para poder llevar a cabo esto, se cuenta con una base de datos de 707 pacientes y 343 campos, donde cada paciente tiene una etiqueta con la variable a clasificar, esto es, "grupo de comparación" o "grupo de riesgo suicida", basado en los grupos con/sin conducta suicida.

1.1.2. Objetivos específicos

Adicionalmente a la generación de este modelo de clasificación, se consideran los siguientes objetivos específicos del proyecto:

- Aplicación de técnicas de limpieza de datos.
- Seleccionar variables más relevantes para discriminar el riesgo suicida.
- Evaluar diferentes modelos de clasificación.
- Calcular medidas de error/precisión para el modelo finalmente seleccionado.

Todo esto, enmarcado en la utilización de herramientas de minería de datos como metodología, la cual es presentado en el capítulo 2.

La presente memoria se estructura en ocho Capítulos, el primero de ellos asociado a la introducción, donde además se definen los objetivos de esta memoria. Por su parte, el Capítulo 2 muestra el marco teórico asociado al trabajo realizado, en la Sección 2.1.1 se expone la metodología de trabajo KDD para proyectos de data mining, la cual se ocupa aquí, mientras que las Secciones 2.2 , 2.3 y 2.4 detallan las técnicas y herramientas utilizadas para la imputación de datos, modelamiento y validación de los modelos respectivamente.

Desde el Capítulo 3 hasta el Capítulo 7 se aborda el problema de generación de un modelo predictivo del riesgo suicida, visto como un problema clasificación utilizando aprendizaje supervisado. Y cada uno de estos Capítulos hace mención a uno de los pasos de la metodología KDD. Así por ejemplo, en el Capítulo 3 se revisan los datos usados para la modelación, y se da una breve explicación acerca de los instrumentos clínicos utilizados.

Por su parte en el Capítulo 4 se analiza la calidad de cada campo para su uso en el modelamiento, además se muestran diferentes aplicaciones de las herramientas de imputación para valores perdidos según lo descrito en el Sección 2.2.

En el Capítulo 5 se analizan de forma descriptiva los datos y se reclasifican los campos nominales en clases más aptas para el modelamiento. En esta Sección también se ilustran algunas transformaciones de los datos de instrumentos y se termina con una reducción de atributos en base a la correlación.

Por último, en los Capítulos 6 y 7 , se muestran los resultados obtenidos por los diferentes modelos, el ajuste de sus parámetros relevantes y la obtención de medidas error/precisión para una posterior comparación y selección del modelo más robusto.

Capítulo 2

Preliminares

2.1. ¿Qué es la minería de datos?

Hoy en día se vive en una era digital, donde grandes volúmenes de datos son recolectados y procesados diariamente. Datos de transacciones, clientes, información del genoma, incluso en astronomía, medicina o en otras áreas, los datos crecen a un paso agigantado. Sin embargo, la cantidad de información que es posible extraer desde estas fuentes de datos es aún sólo un pequeño porcentaje.

El desafío está ahí, es así como cada vez es más recurrente escuchar términos como minería de datos, *data mining*, *big data* o *machine learning* entre otros. Pues todos ellos se refieren a la manera de poder procesar grandes volúmenes de datos y obtener de ellos información valiosa, nueva y no trivial.

Así como el término *gold mining* se refiere a la extracción de oro a partir de las rocas, de forma análoga podemos considerar el término *data mining* como la extracción de conocimiento a partir de los datos. Si bien *knowledge mining*, pudiera ser más adecuado para esta analogía, esto no deja evidente el hecho de que el conocimiento se extrae de grandes volúmenes de datos, por lo que el uso de *data mining* se ha hecho más popular [28].

2.1.1. Metodología KDD

Para el desarrollo de este proyecto de *data mining* se utiliza la metodología de trabajo KDD (*Knowledge Discovery in Databases*). El proceso KDD es un proceso iterativo, en el que se analizan grandes volúmenes de datos para obtener relaciones y/o identificar patrones en estos. Este proceso extrae información a primera vista oculta y no trivial de los datos, pero de calidad y gran interés para obtener información y conclusiones muchas veces novedosas basadas en las relaciones o modelos dentro de estos.

Figura 2.1: La minería de datos extrae conocimientos desde los datos.

El proceso KDD se divide en 5 etapas [1], las que se muestran en el esquema de la figura 2.2.

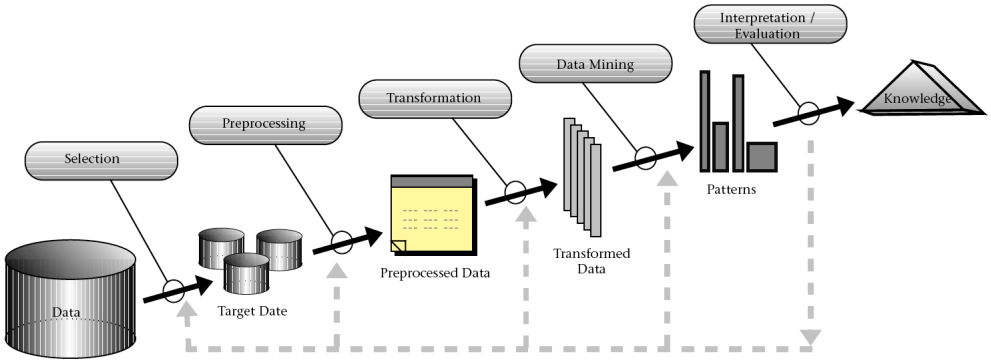

Figura 2.2: Etapas del KDD.

El detalle de cada una de estas fases del proceso KDD se muestra a continuación:

- **Selección de datos**: Durante esta etapa se determinan las fuentes de información y los tipos de datos a utilizar, una vez ya levantada toda la información considerada relevante se procede a la extracción de ésta desde las fuentes de datos disponibles.
- **Preprocesamiento**: Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.
- **Transformación**: Consiste en el tratamiento preliminar de los datos, en la transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización / estandarización, consolidando los datos de una forma necesaria para la fase siguiente.

- **Data mining**: Es la fase de modelamiento propiamente tal, en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, pero potencialmente útiles y comprensibles que están contenidos u ocultos en los datos.
- **Interpretación y Evaluación**: Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realiza una evaluación de los resultados obtenidos.

Si bien las diferentes fases antes mencionadas presentan una determinada estructura, como se mencionó anteriormente, este no es un proceso lineal, sino más bien iterativo, como se muestra en la figura 2.2.

En los Capítulos 3 a 7 se detalla el trabajo realizado usando la metodología KDD, mostrando el trabajo llevado a cabo en cada una de estas fases, para abordar el problema de la clasificación de riesgo suicida.

2.1.2. Herramientas de la minería de datos.

Las herramientas de *data mining* sirven para resolver una variada gama de problemas, en general cualquiera que cuente con grandes cantidades de datos y de los cuales se quiera extraer información valiosa y no evidente. A continuación se resumen los principales tipos de problemas que se pueden resolver usando minería de datos:

- Caracterización y Discriminación.
- Asociación y Correlaciones.
- Clasificación y Regresión para análisis predictivo.
- Análisis de Clúster.
- Análisis de Outliers.

Dada la naturaleza interdisciplinaria de la minería de datos, es que ésta ha incorporado muchas técnicas desde otras áreas, tales como estadística, machine learning, reconocimiento de patrones, bases de datos, computación de alto desempeño y muchas otras [28, pág. 23-27], lo que se ilustra en la figura 2.3.

En las secciones siguientes se muestran ejemplos de algunas de las disciplinas que tienen una gran influencia en el desarrollo de la minería de datos, y en las cuales se sustentan los modelos utilizados en esta memoria.

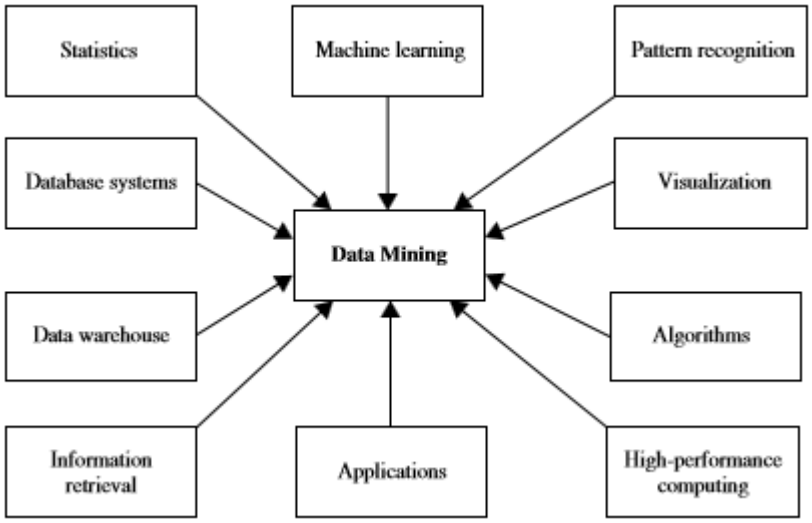

Figura 2.3: Contribución de diferentes áreas al desarrollo de la minería de datos.

Herramientas de la estadística en la minería de datos

La estadística estudia el análisis, interpretación, explicación y representación de los datos, de ahí la conexión inmediata con la minería de datos.

Un modelo estadístico es un conjunto de funciones matemáticas que sirven para describir el comportamiento de objetos y/o fenómenos, en términos de variables aleatorias y sus distribuciones de probabilidad. Los modelos estadísticos son ampliamente utilizados para el modelamiento de datos. Por ejemplo, en tareas de minería de datos tales como caracterización y clasificación, se pueden construir modelos estadísticos para caracterizar los datos o clasificarlos de acuerdo a una variable objetivo. Esto es, los modelos estadísticos pueden ser el resultado de una tarea de minería de datos.

Análogamente, puede ser que un modelo estadístico pueda ser usado para la aplicación posterior de técnicas de data mining, por ejemplo, podemos utilizar un modelo estadístico para modelar ruido e imputar valores perdidos para luego utilizar otras técnicas de *data mining* para resolver un problema diferente usando estos datos.

Por otra parte, métodos estadísticos pueden ser usados para resumir o describir un conjunto de datos. La estadística es útil para obtener patrones en los datos, como también para entender los mecanismos y/o fenómenos que generan y afectan dichos patrones.

Los métodos estadísticos también pueden ser usados para verificar los resultados de una tarea de minería de datos. Por ejemplo, luego de hacer un modelo de clasificación o predicción, el modelo debiera ser verificado mediante un test de hipótesis, usando las técnicas clásicas de estadística.

La aplicación de métodos estadísticos en minería de datos es lejos de ser un proceso trivial, muchas veces un desafío importante es como escalar un método estadístico a un gran conjunto de datos. Muchos métodos estadísticos son de gran complejidad computacional. Luego cuando tales métodos son llevados sobre grandes conjuntos de datos, los que también pueden estar distribuidos en diversas fuentes, deben ser cuidadosamente diseñados e implementados para reducir el costo computacional.

Algunas de las técnicas más comunes aplicadas en minería de datos que provienen de la estadística son regresión lineal, regresión logística y ANOVA entre otras.

Machine Learning

Machine learning se asocia a como los computadores pueden aprender a partir de los datos. Un área de investigación importante es programar una computadora para aprender de forma automática a reconocer complejos patrones en los datos, y tomar decisiones basadas en estos. Por ejemplo, un problema típico de *machine learning* es programar un computador para reconocer automáticamente códigos postales escritos en letra manuscrita en cartas, esto se logra al "enseñar" al computador a través de un conjunto de ejemplos. *Machine learning* es una disciplina en rápido crecimiento en la actualidad.

Entre los algoritmos de esta disciplina existen distintos tipos, como se muestra a continuación:

- **Aprendizaje supervisado**: Es básicamente un sinónimo para clasificación. La supervisión en el aprendizaje viene de la variable objetivo (la clasificación que se desea lograr con el modelo).
- **Aprendizaje no supervisado**: Es esencialmente un sinónimo para *clustering*. El aprendizaje no es supervisado ya que los datos de entrenamiento no poseen una clasificación a priori, usualmente se usa aprendizaje no supervisado para descubrir clases (no conocidas) en los datos.
- **Aprendizaje semi-supervisado**: Es una clase de técnica de *machine learning* que hace uso tanto de datos etiquetados (clasificación a priori) como no etiquetados en el entrenamiento del modelo. Los datos etiquetados sirven para definir las clases del modelo, mientras que los datos no etiquetados se utilizan para refinar los bordes de estas clases.
- **Aprendizaje activo**: Es un enfoque de *machine learning* que permite al usuario jugar un rol activo en el proceso de entrenamiento del modelo. En este enfoque se le puede estar preguntando al usuario experto que etiquete un conjunto de datos, el objetivo de este enfoque es optimizar la calidad del modelo mediante la adquisición activa de conocimiento desde los usuarios.

Dado lo anterior se puede observar que hay muchas similitudes entre minería de datos y

machine learning. Para tareas de clasificación y *clustering*, *machine learning* se enfoca más en la precisión del modelo, mientras que minería de datos se refiere más bien al proceso global, si bien la precisión de los modelos es importante también se debe dar énfasis en la eficiencia y escalabilidad de los métodos de *data mining* sobre grandes volúmenes de datos, así también como en el manejo de complejos tipos de datos y la exploración de nuevos y alternativos métodos. Entre las técnicas más utilizadas del *Machine Learning* se encuentran árboles de decisión, *support vector machine* (SVM) y algoritmos genéticos entre otros.

2.2. Técnicas de imputación de datos tradicionales

La mayoría de los paquetes estadísticos asumen que se trabaja con datos completos e incorporan opciones, no siempre las más adecuadas, para imputar observaciones sin que el usuario se dé cuenta de ello.

Está ampliamente documentado que la aplicación de procedimientos inapropiados de imputación en los datos, puede introducir sesgos y reducir el poder explicativo de los métodos estadísticos, quitándole eficiencia a la fase de inferencia, pudiendo incluso invalidar las conclusiones del estudio. De aquí la relevancia de definir un apropiado método de imputación para aquellos valores perdidos.

A modo de ejemplo, si se trabaja con una base de datos donde la variable objetivo cuenta con un 25 % de valores perdidos, se debe tener presente que imputar la respuesta en una de cada cuatro observaciones puedes ser adecuado dentro de un marco académico, sin embargo, se considera poco útil desde un punto de vista práctico. En especial cuando los resultados del análisis se desean utilizar para apoyar el desarrollo y/o evaluación de por ejemplo políticas públicas [23]

La ocurrencia de valores perdidos en lo datos, puede asociarse a diversas causas, por ejemplo:

- Fatiga del paciente y/o entrevistador.
- Desconocimiento de la información solicitada.
- Rechazo del paciente a responder lo solicitado.
- Problemas asociados a la calidad del marco de muestreo.
- Errores en la transcripción de los datos físicos (papel) a un sistema digital.
- En el caso del estudio que se presenta, se puede también asociar al impacto del paciente ante la pregunta o que la pregunta no aplique para su caso (por ejemplo el campo asociado a motivo de hospitalización no aplica para pacientes ambulatorios).

Existen diferencias entre la falta de respuesta total (faltan todas las preguntas del paciente), y la no respuesta parcial, donde no se obtiene respuesta sólo en algunas de las preguntas de los cuestionarios. Generalmente la falta de respuesta total se corrige eliminando aquellos registros de la base de datos y ajustando los factores de expansión, de manera que los registros

que permanecen en la muestra estimen, sin sesgos, el total poblacional. Por su parte, la no respuesta parcial se suele corregir con el método de imputación *hot-deck*, u otros métodos de imputación que se describirán más adelante.

Antes de proceder con la imputación de los valores perdidos es necesario determinar el patrón de los valores faltantes. Se afirma que un proceso de datos perdidos sigue un patrón MAR (*Missing at Random*), si la distribución de los valores observados no depende del patrón de comportamiento de los registros sin información.

Generalmente en la práctica se asume un patrón del tipo MCAR (*Missing Completely at Random*). El supuesto de que los datos faltantes siguen un patrón MCAR fue introducido por Rubin '77 [45] y Little - Rubin '87 [37], y como se menciona es el supuesto que se asume en la mayoría de los algoritmos de imputación. Sin embargo, en la práctica esto puede no ser así, y los datos podrían seguir otro patrón como MNAR (Missing not at Random), donde los valores perdidos si dependen de los mismos datos.

Cuando los datos provienen de una muestra aleatoria, los patrones de datos omitidos MAR, MCAR y MNAR pueden interpretarse a partir de X (atributos) e Y (objetivo). MCAR significa que la probabilidad de que falte Y no depende de los valores de X e Y (ni de los propios ni de las otras observaciones), MAR asume que la probabilidad de que falte Y puede depender de X pero no de Y, en tanto que MNAR se interpreta en el sentido de que la probabilidad de que existan datos omitidos depende de Y.

En lo que sigue se detallan algunos de los procedimientos de imputación más comunes.

2.2.1. Listwise deletion

El procedimiento de *listwise deletion* es el que se utiliza con mayor frecuencia en la práctica. Éste asume un patrón de datos MCAR, y consiste en eliminar aquellos casos con información perdida, trabajando sólo con aquellos registros que poseen información completa para todas las variables.

Cuando los datos analizados provienen de una muestra probabilística, eliminar observaciones no es correcto, ya que se debe tener presente que el tamaño y composición de la muestra fueron elegidas con un procedimiento aleatorio y con probabilidad de selección, conocida y distinta de cero, que no puede ser ignorada en el tratamiento de los datos ni en el cálculo de los estimadores y sus errores.

En muchas herramientas y paquetes estadísticos es habitual que se trabaje sólo con la in-

formación completa por defecto, a pesar de que se reconoce que esta práctica no es la más adecuada ya que genera sesgos en los coeficientes de asociación y correlación [32]. Sin embargo la utilización de esta u otra técnica de imputación se debe analizar caso a caso.

2.2.2. Imputación por medias no condicionadas

La imputación de datos mediante la media es uno de los métodos más antiguos y utilizados entre los investigadores de diversas disciplinas, a pesar de que por sus limitaciones teóricas no se considera un procedimiento adecuado. Esto debido básicamente a su fácil implementación respecto a otros métodos, que si bien pueden ser más adecuados son más complejos de implementar.

En la aplicación de este método se asume que los datos perdidos siguen un patrón MCAR, y consiste en sustituir, como su nombre lo indica, aquellos valores perdidos con el correspondiente promedio o moda de la variable, y en algunas variaciones se utiliza la mediana en vez del promedio.

Pese al gran uso de esta técnica, se ha documentado ampliamente que su aplicación afecta la distribución de probabilidad de la variable imputada, atenúa la correlación con el resto de las variables y subestima la varianza de la variable imputada, entre otras cosas. Por la manera en que se realiza la sustitución de los datos omitidos, la suma de cuadrados de las desviaciones de las observaciones respecto de la media permanece inalterada pero se incrementa el tamaño de muestra, lo cual origina que la varianza de la variable disminuya y se generen, en forma artificial, intervalos de confianza más estrechos. Un ejemplo de la alteración de la distribución se puede ver gráficamente en la figura 2.4.

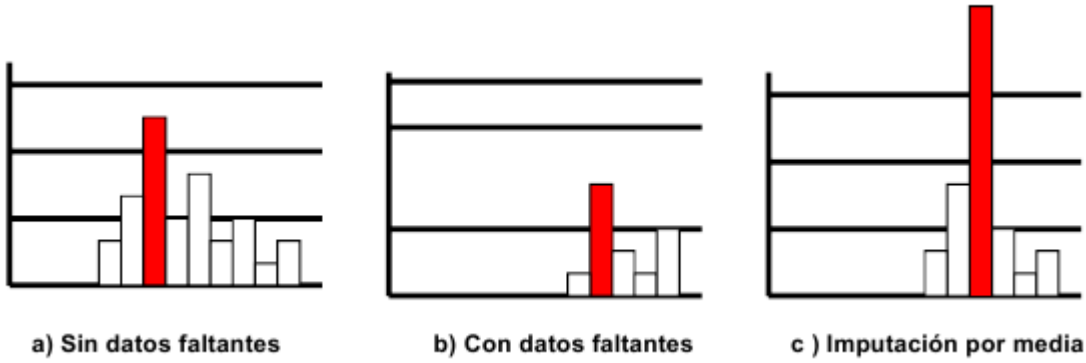

Figura 2.4: Ejemplo de distribución de datos modificada al utilizar imputación por medias no condicionadas.

Adicionalmente, en caso de que las variables imputadas se utilicen en análisis secundario de

datos, se demuestra, por ejemplo, que en los modelos de regresión se alteran los valores de los parámetros estimados, así como su significancia estadística.

2.2.3. Imputación por medias condicionadas

Una variante del procedimiento anterior consiste en formar categorías a partir de covariables correlacionadas con la variable de interés, e imputar los datos omitidos con observaciones provenientes de la submuestra que comparte características comunes [2].

Al igual que en el caso anterior (medias no condicionadas), se asume que los datos perdidos siguen un patrón MCAR, sin embargo, en este caso existirán tantos promedios como categorías se hayan definido. Si bien esto ayuda a disminuir los sesgos de la media no condicionada, en ningún caso los elimina.

2.2.4. Imputación mediante Random Hot-Deck (RHD)

Los estadísticos de encuestas crearon el procedimiento de imputación no paramétrico llamado hot-deck [38], cuyo fin es mantener invariante la distribución de probabilidad de las variables con datos incompletos.

El método asume un patrón de datos perdidos MCAR, y tiene como objetivo sustituir los valores perdidos (receptores) con información de una muestra "similar" con información completa (donantes). Los datos faltantes son sustituidos a partir de una selección aleatoria de valores observados en una categoría similar a la de los registros sin información. Así, de esta manera la imputación no introduce sesgos en la varianza del estimador.

Para esto, es fundamental generar agrupaciones que garanticen que la imputación se llevará a cabo entre observaciones con características comunes, y la selección de los donantes se realiza en forma aleatoria (con reemplazo dentro de la categoría) evitando que se introduzcan sesgos en el estimador de la varianza.

Su aplicación supone que la falta de respuesta se distribuye en forma aleatoria en cada una de las categorías, pero en caso de que la falta de respuesta se concentre en un estrato con pocas observaciones, es posible que se generen estimadores sesgados en la medida que el procedimiento seleccione varias veces el mismo donante.

Existen algunas variantes del procedimiento *hot-deck*. El "algoritmo secuencial", parte de un proceso de ordenación de los datos en cada subgrupo y selecciona donantes de acuerdo al or-

den definido en estos. Por su parte, el "método aleatorio" (RHD) sustituye aquellos registros sin datos mediante una elección estocástica del donante. También existe la posibilidad de que el donante sea el "vecino más cercano" al registro sin datos, en cuyo caso se debe definir una noción de distancia para identificar los vecinos.

El *hot-deck* y las variantes que se han comentado se consideran mejores opciones que el procedimiento *listwise deletion*, y es superior a los métodos de medias condicionadas y no condicionadas, ya que no introduce sesgos en el estimador y su error estándar.

Además, si se desea preservar la distribución de probabilidad de las variables imputadas, conforme a la opinión de algunos autores, se considera que el procedimiento *hot-deck* es más eficiente que el algoritmo de imputación múltiple y la regresión paramétrica [19].

2.2.5. Criterio experto

Según el problema que se esté analizando y de acuerdo al conocimiento que se tenga del fenómeno y de los datos, es posible realizar imputaciones utilizando un criterio experto.

2.2.6. Otras técnicas de imputación de datos

Adicionalmente a las técnicas de imputación antes descritas, que son las que se utilizan en esta memoria, a continuación se resumen otras que cabe la pena señalar:

- **Imputación por regresión**: Dado un patrón de datos faltantes MCAR, se especifica un modelo, en donde las covariables están altamente correlacionadas con la variable a imputar.
- **Estimación por máxima verosimilitud**: Proceso iterativo donde se calculan los estimadores del modelo usando sólo los datos completos. Una vez obtenidos se estiman los datos faltantes en función del estimador de máxima verosimilitud y esto se itera hasta encontrar la convergencia del método.
- **Imputación múltiple**: Utiliza métodos de simulación de Monte Carlo, que en conjunto con un modelo que relacione la variable a imputar y sus covariables (las que deben tener una alta correlación) permiten estimar los valores perdidos

Si bien la imputación múltiple (IM) y la EMV son métodos bastante aceptados y utilizados en la actualidad, no hay una regla inequívoca sobre cuál técnica es mejor que las otras por lo que se debe ser cauto y elegir caso a caso la mejor técnica adecuada al conjunto de datos que se está estudiando.

2.3. Técnicas de Clasificación

En esta sección se muestran los diferentes tipos de técnicas de clasificación utilizadas durante el desarrollo de esta memoria.

2.3.1. Árboles de decisión

Son modelos que tienen estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos. Básicamente, los árboles de decisión, son representaciones gráficas de la lógica de las probabilidades aplicada a las alternativas de decisión. El tronco del árbol es el punto de partida de la decisión. Las ramas de éste comienzan con la probabilidad del primer acontecimiento. La probabilidad de cada acontecimiento produce dos o más efectos posibles, algunos de los cuales conducen a otros acontecimientos de probabilidad y a puntos de decisión subconsecuentes. Los valores en los que se cifran las ramas del árbol, provienen de un análisis muy cuidadoso que se basa en el establecimiento de un criterio para la toma de decisión.

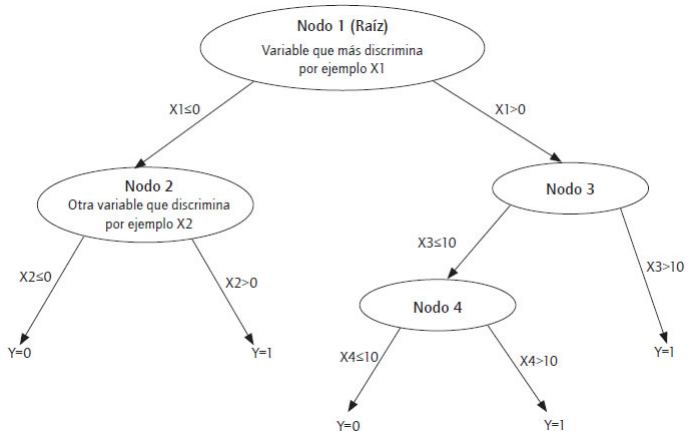

Figura 2.5: Ejemplo de árbol binario.

Hay varios tipos de árboles de decisión, los más conocidos son:

- ID3 (Iterative Dichotomiser 3)
- C4.5 (successor of ID3)
- CART (Classification And Regression Tree)
- CHAID (CHi-squared Automatic Interaction Detector).

Se explicará sólo el árbol de decisión del tipo CART, ya que éste es el que se utiliza en esta memoria. Una revisión más completa de ID3, CART y C4.5 se puede ver en [49].

CART (Classification And Regression Tree)

CART es un tipo de árbol de decisión binario, esto es, en cada nodo sólo se pueden dividir los datos en dos grupos. CART es capaz de manejar tanto datos numéricos como categóricos, e incluso posee sofisticados métodos para tratar con valores perdidos, como por ejemplo el uso de variables sustitutas o *surrogate variables* como se denomina en inglés. Otra de las características de CART es que se puede utilizar tanto en problemas de clasificación como de regresión.

A continuación se detallan las etapas en la construcción de CART.

Construcción del árbol

La construcción del árbol comienza en el nodo raíz, donde se incluyen todos los casos del conjunto de entrenamiento. El algoritmo de CART busca la mejor variable para dividir los datos. Para hacer esto, el algoritmo revisa todas las posibles variables, y dentro de ellas, todas las posibles divisiones.

Para elegir la mejor división, el algoritmo busca maximizar una medida de pureza de los nodos hijos resultantes de la división. Dependiendo del software que se utilice, algunas de las medidas de pureza / impureza utilizada pueden ser [49]:

- **Entropía**: Mide la impureza de un nodo. Se define, para una variable binaria, de la siguiente manera:

$$Entropía(t) = -\sum_{\mathrm{i}} p(\mathrm{i}/t) log_2 p(\mathrm{i}/t) \tag{2.1}$$

  Donde $p(\mathrm{i}/t)$ corresponde a la proporción de casos pertenecientes a la clase i en el nodo t.
- **Índice de Gini**: Otra medida de impureza, que mide la diferencia de las distribuciones de probabilidad de los valores de la variable objetivo en cada nodo hijo.

$$Indice\ de\ Gini = 1 - \sum (p(\mathrm{i}/t))^2 \tag{2.2}$$

- **Error de Clasificación**: Mide la proporción de casos mal clasificados. Se calcula de la siguiente manera:

$$Error\ de\ Clasificación(t) = 1 - \mathrm{máx}(p(\mathrm{i}/t)) \tag{2.3}$$

- **Ganancia de Información**: Es una medida de impureza basada en la entropía. Corresponde a la diferencia de entropía de un nodo antes de dividir (nodo padre) y después de hacerlo (nodo hijo).

$$Ganancia\ de\ Información(\nabla) = Entropía(nodo\ padre) - Entropía(nodo\ hijo) \tag{2.4}$$

- **Ratio de Ganancia**: El Ratio de ganancia normaliza el valor de la Ganancia de Información.

$$Ratio \text{ de } Ganancia = \frac{Ganancia \text{ de } Información(\nabla)}{Entropía} \tag{2.5}$$

- **Criterio de Twoing**: El criterio de Twoing o también llamado criterio binario, busca dividir los datos de manera más pareja que el indice de Gini, separando los datos en 2 grupos, cada uno con el 50 % de los datos aproximadamente. Se define como sigue:

$$Criterio \text{ de } Twoing(t) = \frac{P_L P_R}{4}\Big(\sum |p(\mathrm{i}/t_L) - p(\mathrm{i}/t_R)|\Big)^2 \tag{2.6}$$

  Donde L y R se refieren a los lados izquerdos y derechos del split.

La construcción del árbol continúa en cada nuevo nodo de forma recursiva hasta que es imposible seguir.

Asignación de clases a los nodos

Cada nodo, incluyendo el nodo raíz es asignado a una clase objetivo. La clase asignada puede depender de los siguientes 3 factores:

- La probabilidad a priori de cada clase, que se asume en cada nuevo conjunto de datos.
- La matriz de costos.
- La proporción de registros de cada clase calculada con los datos de entrenamiento que hay en cada nodo.

Así, estos tres factores son necesarios para asignar el valor de la clase en cada nodo, donde la fórmula para calcular la clase resultante viene dada por la ecuación 2.7

$$\frac{C(j|\mathrm{i})\pi(\mathrm{i})N_{\mathrm{i}}(t)}{C(\mathrm{i}|j)\pi(j)N_j(t)} > \frac{N_{\mathrm{i}}}{N_j}, \forall j \tag{2.7}$$

Donde

$C(\mathrm{i}|j)$ *Costo* de *clasificar* i *como* j

$\pi(\mathrm{i})$ *probabilidad a priori* de i

N_{i} *Número* de *elementos* de *la clase* i *en el conjunto* de *datos*

$N_{\mathrm{i}}(t)$ *Número* de *elementos* de *la clase* i *en el nodo* t

Valores perdidos

Para cada nodo, la "división primaria" es la variable que mejor divide el conjunto de datos, maximizando la pureza de los nodos hijos. Sin embargo, cuando la variable asociada a la división primaria posee valores perdidos, CART busca una "variable sustituta" para usar en su

lugar. Una variable sustituta es aquella cuyo patrón relativo a la variable objetivo es similar a la de la división primaria.

La utilización de variables sustitutas permiten utilizar todas las observaciones cuando la calidad de los conjuntos de datos es razonable. Esto significa una mejora respecto a otros árboles de decisión o incluso otras técnicas de clasificación que frente a valores perdidos eliminan esta información.

Detención de la construcción del árbol

El proceso de construcción del árbol continúa hasta que es imposible seguir, lo cual se puede deber a alguna de las siguientes razones:

- Hay sólo una observación en cada nodo hijo.
- Todas las observaciones en cada nodo hijo pertenecen a la misma clase objetivo.
- Un límite externo en el número de niveles del árbol ha sido definido por el usuario.

Poda del árbol

Para generar un árbol más simple, CART utiliza un parámetro de complejidad, que sirve para podar aquellos nodos hijos si la mejora en el error de clasificación del nuevo árbol no es sustancialmente importante respecto al aumento en la complejidad del árbol al incorporar dicha división. En otras palabras, este parámetro es una medida de cuanta más precisión debe aportar una división al árbol entero para garantizar una complejidad mayor. Luego, mientras mayor sea este parámetro, más nodos son podados resultando un árbol mucho más simple.

Selección del árbol óptimo

Un árbol máximal, esto es, uno con todos sus nodos e hijos, y que no ha sido podado, siempre se ajustará con gran precisión al conjunto de datos de entrenamiento, sin embargo, esta precisión sobreestima la real precisión sobre un nuevo conjunto de datos que sigue la misma distribución que la de entrenamiento. Esto sucede ya que el árbol máximal se ajusta a la peculiaridad y al ruido de los datos de entrenamiento (sobreajuste). Así, la selección del árbol óptimo se refiere a la búsqueda del parámetro de complejidad, que hace que un árbol aprenda de los datos pero que no se sobreajuste y sea capaz de generalizar los resultados ante un nuevo conjunto de datos. Para lograr esto, es necesario un conjunto de datos de validación, como se explicará en el capítulo 2.4. En la figura 2.6 se muestra un esquema de la selección del parámetro de complejidad en función del error de clasificación cometido en los datos de validación.

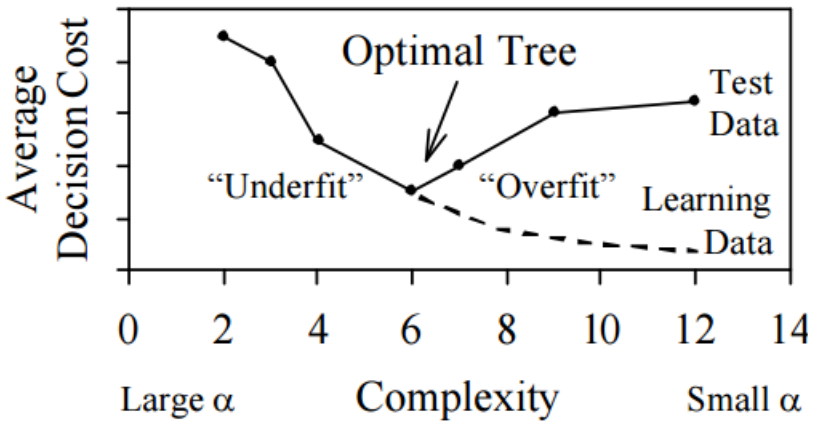

Figura 2.6: Ajuste del parámetro de complejidad para evitar el sobreajuste.

2.3.2. Support vector machine (SVM)

Las máquinas de soporte vectorial o máquinas de vectores de soporte (*Support Vector Machines*, SVMs) son un conjunto de algoritmos de aprendizaje supervisado desarrollados por Vladimir Vapnik y su equipo en los laboratorios AT&T ([8] , [15]).

Estos métodos están propiamente relacionados con problemas de clasificación y regresión [27]. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, un SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio posible. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas a una u otra clase.

Más formalmente, un SVM construye un hiperplano o un conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá una clasificación correcta, luego el algoritmo busca el o los hiperplanos que poseen un mayor "margen" entre las clases, como se ilustra en la figura 2.7.

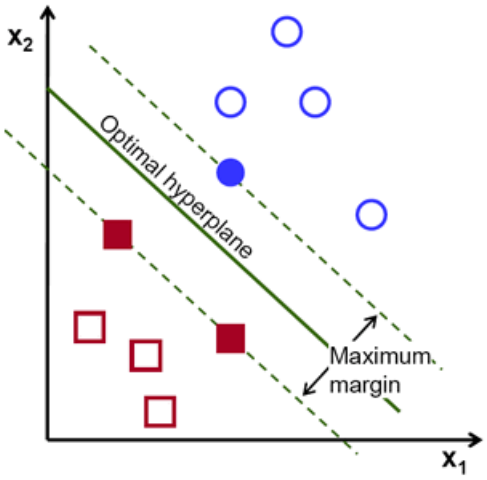

Figura 2.7: Esquema del hiperplano separador generado por una SVM.

Tipos de Kernels

Para poder llevar las observaciones a espacios de dimensión mayor, las SVM hacen uso de funciones de enlace o transformaciones, las que se conocen con el nombre de *Kernels* o núcleo de la función. Entre los tipos de *Kernels* más utilizados destacan:

- Lineal: $K(x_{\mathrm{i}}, x_j) = x_{\mathrm{i}}^T x_j$
- Polinomial: $K(x_{\mathrm{i}}, x_j) = (\gamma x_{\mathrm{i}}^T x_j + r)^{\mathrm{d}}, \gamma > 0$
- Función de base radial (RBF): $K(x_{\mathrm{i}}, x_j) = exp(-\gamma \|x_{\mathrm{i}} - x_j\|^2), \gamma > 0$
- Sigmoidea: $K(x_{\mathrm{i}}, x_j) = tanh(\gamma x_{\mathrm{i}}^T x_j + r)$

Donde γ, r y d son parámetros exclusivos de cada Kernel.

Ventajas del SVM

Algunas de las ventajas del SVM respecto a otras técnicas de clasificación son:

- La función Kernel permite transformar a espacios de dimensión muy superior (incluso infinita).
- El espacio de búsqueda tiene sólo un mínimo global.
- El entrenamiento es muy eficiente.
- La clasificación es muy eficiente.
- Se diseña sólo la función de Kernel y el parámetro de costo C.
- Muy buen funcionamiento en problemas típicos.

- Extremadamente robusto para generalización; menos necesidad de heurísticas para entrenamiento.

2.3.3. K- Nearest Neighbors (KNN)

K-nearest neighbors es un algoritmo de clasificación simple [24], que almacena todas las observaciones disponibles y clasifica nuevos casos basado en una medida de similaridad. Así, al usar KNN una nueva observación es clasificada por votación entre sus K vecinos más cercanos. KNN has sido utilizado en reconocimiento de patrones desde 1970 como una técnica de clasificación no paramétrica.

Algunas de las medidas de similaridad más utilizadas son:

- Distancia Euclídea: $\mathrm{d}(x, y) = \sqrt{\sum_{\mathrm{i}} |x_{\mathrm{i}} - y_{\mathrm{i}}|^2}$
- Distancia de Manhattan: $\mathrm{d}(x, y) = \sum_{\mathrm{i}} |x_{\mathrm{i}} - y_{\mathrm{i}}|$
- Distancia de Minkowski: $\mathrm{d}(x, y) = (\sum_{\mathrm{i}} |x_{\mathrm{i}} - y_{\mathrm{i}}|^p)^{\frac{1}{p}}$

Dada las medidas de similaridad utilizadas, es necesario normalizar las variables en una escala común antes de la aplicación del método, de lo contrario podría verse afectado por la escala en que se miden los atributos. Otro punto a considerar es que las medidas definidas anteriormente solo se aplican para atributos numéricos. Para poder evaluar variables categóricas se deben usar métricas como la de Hamming, definida por la ecuación 2.8

$$\mathrm{d}(x, y) = \sum \delta(x_{\mathrm{i}}, y_{\mathrm{i}}) \tag{2.8}$$

Donde

$$\delta(x_{\mathrm{i}}, y_{\mathrm{i}}) = \begin{cases} 0 & si \quad x_{\mathrm{i}} \neq y_{\mathrm{i}} \\ 1 & si \quad x_{\mathrm{i}} = y_{\mathrm{i}} \end{cases} \tag{2.9}$$

Por otra parte, el parámetro K (Cantidad de vecinos más cercanos) debe ser definido para el ajuste del modelo, una forma de hacerlo es por inspección en la data. Generalmente un K más grande es más preciso ya que reduce el ruido, aunque no hay garantía de ello por lo que debe ser evaluado caso a caso. Otra forma de estimar el K óptimo es mediante validación cruzada, donde se observan los resultados para distintos valores de K y se elige aquel que tenga un menor error promedio. Usualmente el K óptimo se encuentra entre 3 y 10 en la mayoría de los conjuntos de datos.

2.4. Validación de modelos

Para la evaluación del desempeño de un determinado modelo de minería de datos se deben considerar varios puntos a evaluar. Algunos de ellos son:

- Estrategia para el entrenamiento y prueba del modelo.
- Tipos de error y medidas de evaluación.
- Medidas de "performance" y aplicabilidad.

Antes de detallar cada uno de estos puntos se revisarán algunas definiciones básicas a tener en cuenta.

Definición 2.1 (Conjunto de datos) *Se distinguen dos tipos, el conjunto de entrenamiento y el conjunto de prueba. Para obtener estos, dividimos los datos muestrales en dos partes; una parte se utiliza como conjunto de entrenamiento para determinar los parámetros del clasificador y la otra parte, llamada conjunto de prueba (validación, test ó conjunto de generalización) se utiliza para estimar el error de generalización ya que el objetivo final es que el clasificador consiga un error de generalización pequeño evitando el sobreajuste (ó sobreentrenamiento), que consiste en una sobrevaloración de la capacidad predictiva de los modelos obtenidos.*

Definición 2.2 (Modelo) *o clasificador, es una relación entre las variables que son dadas y las que se van a predecir. Usualmente las variables que se van a predecir se denominan dependientes, y aquellas que se utilizan para predecir se llaman independientes.*

Definición 2.3 (Error del modelo) *En modelos de clasificación, se asocia el error del modelo a la proporción de observaciones mal clasificadas sobre observaciones totales.*

$$Error\ \text{del}\ modelo\ = \frac{\#\ \text{de}\ observaciones\ mal\ clasificadas}{\#\ \text{de}\ observaciones\ totales} \tag{2.10}$$

Luego, como se verá más adelante la idea es obtener un modelo cuyo error en el conjunto de validación sea el menor posible, para la estimación del error en el conjunto de validación existen más de una estrategias, entre ellas por ejemplo la metodología *holdout* y la validación cruzada. En la sección siguiente se revisará esto con mayor detalle.

2.4.1. Estrategia para el entrenamiento y prueba del modelo

Al contar con una sola base de datos se debe diseñar un método de medición independiente, luego, la idea natural es dividir la base de datos en dos conjuntos, el primero, asociado al conjunto de entrenamiento del modelo, y un segundo conjunto asociado a la validación de éste.

El principal problema de esto es que ambas partes deben ser representativas, es decir, todas las clases deben estar bien representadas en cada conjunto. Al seleccionar de forma aleatoria estos conjuntos existe un *trade-off* entre la cantidad de datos seleccionados para el entrenamiento y la validación. Por una parte, es necesario contar con un conjunto de entrenamiento grande para poder ajustar el modelo, sin embargo, también es necesario contar con un conjunto de prueba mayor para tener una buena estimación del error. Generalmente se considera un 70-80 % de los datos como parte del conjunto de entrenamiento y el restante 20-30 % de los datos como parte del conjunto de validación.

A esta forma de separar el conjunto de datos y entrenar con una y validar con la otra se le denomina método *holdout*. El modelo se ajusta usando el conjunto de entrenamiento y luego se evalúa el error del modelo en el conjunto de validación. El error sobre este conjunto es el que se utiliza finalmente como indicador de precisión del modelo.

La ventaja de este método es su fácil implementación y la rapidez computacional, sin embargo, su evaluación puede tener una gran varianza, ya que la evaluación puede depender fuertemente de los conjuntos de datos utilizados. Luego, los resultados obtenidos pueden variar significativamente dependiendo de cómo se seleccionan ambas particiones.

Una mejora del método anterior es lo que se denomina validación cruzada de "n-folds". En este caso, el conjunto de datos se divide en n subconjuntos disjuntos, el modelo se entrena con $n-1$ conjuntos y se evalúa el modelo obtenido en el conjunto no utilizado para el entrenamiento. Lo anterior se realiza n veces (generándose n modelos), y dejando cada vez uno de los n conjuntos como conjunto de validación, como se muestra en la figura 2.8.

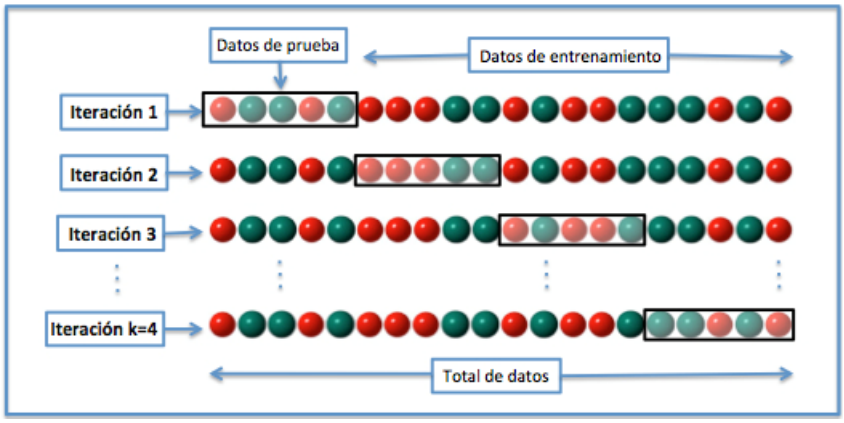

Figura 2.8: Esquema de división del conjunto de datos para la validación cruzada del modelo.

Una vez realizada las n iteraciones, el error del modelo se calcula como el error promedio. Usualmente en la práctica el caso más común es usar $n = 10$, aunque en ciertos casos, cuando la cantidad de observaciones es reducida, también se suele ajustar este parámetro n de manera de reducir el error de clasificación.

La idea detrás de la validación cruzada y sobre la cual se basa, es que el desempeño promedio de los n modelos generados en cada una de las n iteraciones es un excelente estimador del desempeño del modelo original (el cual se genera usando todos los datos) al evaluarlo en un futuro conjunto independiente de pacientes.

La desventaja de este método es que el algoritmo para entrenar el modelo debe ser ejecutado n veces, lo que significa que se incremeneta el costo computacional del modelo, ya que demora n veces más en obtener una evaluación en comparación al método *holdout*.

La ventaja de este método es que, a diferencia del método *holdout*, la forma como los datos se particionan tiene menor importancia. Pues, cada caso está en el conjunto de validación una vez, y está en el conjunto de entrenamiento $n-1$ veces. Por otro lado la varianza del error resultante disminuye a medida que n aumenta. En resumen, si bien la validación cruzada de "n-folds" es un método computacionalmente intensivo para validar el proceso de construcción del modelo, la ventaja de este método es que evita el requisito de un nuevo o independiente conjunto de validación, y su estimación del error es más robusta que al usar el método *holdout*.

Por último, una variante de la validación cruzada de *n-folds* es la K-validación cruzada de *n-folds*, la que corresponde a iterar K veces la evaluación cruzada de *n-folds*, y su intención es minimizar (promediando) el ruido incorporado. Así, si los experimentos tienden a infinito, entonces el promedio del error obtenido tiende al error sistemático del modelo. Si bien este método es muy costoso computacionalmente, es el método estándar en los experimentos en papers, donde usualmente se utiliza $K = 100$ o 1000 y $n = 10$.

2.4.2. Tipos de error y medidas de evaluación

Además del error de clasificación antes mencionado, existen otras medidas de error utilizadas para la evaluación de modelos de clasificación. Muchas de estas medidas se calculan en función de la matriz de confusión asociada al modelo, la que se define a continuación:

Definición 2.4 (Matriz de confusión) *Una matriz de confusión [34] contiene información acerca de la clasificación real y la predicha por un sistema de clasificación. La figura 2.9 muestra la matriz de confusión para un clasificador binario.*

		Valor Observado	
		Clase 1	Clase 2
Predicción	Clase 1	VP	FP
	Clase 2	FN	VN

Figura 2.9: Matriz de confusión para un clasificador binario.

Donde las entradas en la matriz de confusión tienen el siguiente significado:

- ***Verdaderos positivos (VP)**: Casos que pertenecen a la clase 1 y el clasificador los definió como 1.*
- ***Falsos positivos (FP)**: Casos que pertenecen a la clase 1 y el clasificador los definió como 2.*

- ***Falsos negativos (FN)**: Casos que pertenecen a la clase 2 y el clasificador los definió como 1.*
- ***Verdaderos negativos (VN)**: Casos que pertenecen a la clase 2 y el clasificador los definió como 2.*

Adicionalmente a la definición anterior, en estudios clínicos es necesario conocer los siguientes indicadores de precisión de un modelo de clasificación:

Definición 2.5 (Sensibilidad de una prueba diagnóstica) *Probabilidad de obtener un resultado positivo cuando el individuo tiene la enfermedad. Mide su capacidad para detectar la enfermedad o condición de riesgo cuando está presente.*

$$Sensibilidad \;=\; \frac{enfermos\ positivos}{total\ enfermos} \;=\; \frac{VP}{VP+FN} \tag{2.11}$$

Definición 2.6 (Especificidad de una prueba) *Probabilidad de obtener un resultado negativo cuando el individuo no tiene la enfermedad. Mide su capacidad para descartar la enfermedad o condición de riesgo cuando ésta no está presente.*

$$Especificidad \;=\; \frac{sanos\ negativos}{total\ sanos} \;=\; \frac{VN}{VN+FP} \tag{2.12}$$

La sensibilidad también suele llamarse como VPR (Razón de verdaderos positivos), mientras que al término $1-especificidad$ se le llama FPR (Razón de falsos positivos).

Adicionalmente a la sensibilidad y especificidad del modelo se pueden considerar las siguientes medidas de error/precisión:

- Error de clasificación $= \dfrac{FN+FP}{N}$
- Precisión o Exactitud $= \dfrac{VP+VN}{N} = 1 - Error\ de\ clasificación$
- Valor predictivo positivo $= PPV = \dfrac{VP}{VP+FP}$
- Valor predictivo negativo $= NPV = \dfrac{VN}{VN+FN}$

2.4.3. Curva Receiver Operating Characteristic (Curva ROC)

Otra herramienta útil para evaluar el desempeño de modelos de clasificación es mediante la utilización de una curva ROC. En la Teoría de detección de señales, una curva ROC (acrónimo de *Receiver Operating Characteristic*, o Característica Operativa del Receptor en español) es una representación gráfica de la sensibilidad (o VPR) frente al valor FPR (equivalente a $1-especificidad$) para un sistema clasificador binario. Donde al variar alguno de los parámetros del modelo, es posible obtener una curva en el espacio ROC.

Así, 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, costos). Donde cada resultado de predicción o instancia de la matriz de confusión representa un punto en el espacio ROC.

En la figura 2.10 se muestra el espacio ROC y diferentes puntos dentro de este a modo de ejemplo.

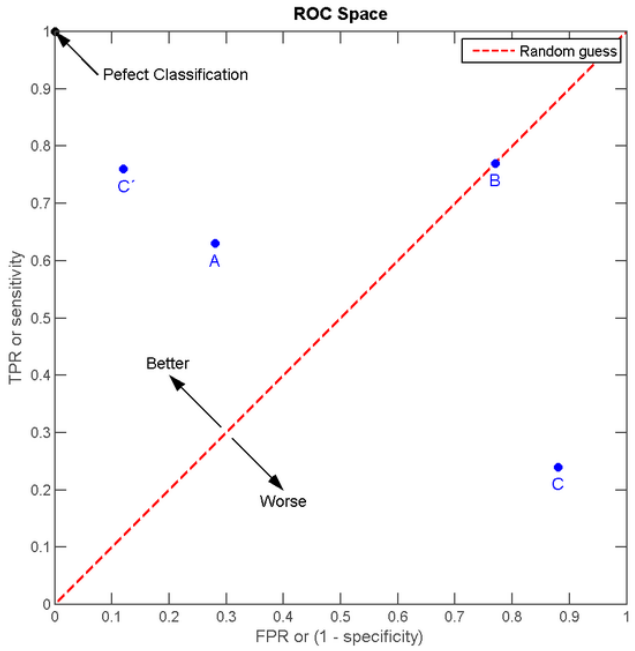

Figura 2.10: Espacio ROC.

Así, el mejor método posible de predicción se situaría en un punto en la esquina superior izquierda, de coordenadas (0,1) del espacio ROC, lo cual representaría un 100 % de sensibilidad (ningún falso negativo) y un 100 % también de especificidad (ningún falso positivo). A este punto (0,1) también se le denomina como clasificación perfecta. Por el contrario, una clasificación totalmente aleatoria (o adivinación aleatoria) daría un punto a lo largo de la línea diagonal (B), desde el extremo inferior izquierdo hasta la esquina superior derecha, que se llama también línea de no-discriminación. 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 (A) representan los buenos resultados de clasificación (mejor que el azar), y los puntos por debajo de la línea representan los resultados pobres (peor que al azar). Nótese que la salida de un predictor consistentemente pobre (C) simplemente podría ser invertida para obtener un buen predictor

(C').

De acuerdo con [11], el uso de las curvas ROC en la evaluación de pruebas diagnósticas presenta las siguientes ventajas:

1. Son una representación fácilmente comprensible de la capacidad de discriminación de la prueba en todo el rango de puntos de corte.
2. Son simples, gráficas y fáciles de interpretar visualmente.
3. No requieren un nivel de decisión particular porque está incluido todo el espectro de puntos de corte.
4. Son independientes de la prevalencia, ya que la sensibilidad y la especificidad se obtienen en distintos subgrupos. Dado esto, no es necesario tener cuidado para obtener muestras con prevalencia representativa de la población. De hecho, es preferible generalmente tener igual número de individuos en ambos subgrupos.
5. Proporcionan una comparación visual directa entre pruebas en una escala común, mientras que otro tipo de gráficos, como los diagramas de puntos o los histogramas de frecuencias, requieren diferentes gráficos cuando difieren las escalas.
6. La especificidad y la sensibilidad son accesibles en el gráfico, en contraste con los diagramas de puntos y los histogramas.

2.4.4. Medidas de performance y aplicabilidad.

Las medidas de performance y aplicabilidad están relacionadas básicamente a la implementación del modelo en algún sistema. Algunos de los criterios de evaluación a considerar en este sentido son:

- Tiempo de entrenamiento del modelo.
- Tiempo de evaluación de nuevas observaciones en el modelo.
- Interpretabilidad de los resultados obtenidos por el modelo.
- El campo de aplicación del modelo obtenido.
- Costo de obtener los datos adecuados para el modelo.
- Costo de actualización del modelo.

Para este trabajo en particular, dado el tamaño de la base de datos (solo 707 casos), es que estas consideraciones no son de interés, salvo la interpretabilidad de los resultados, motivo por el cual, junto a los modelos desarrollados se entrega información descriptiva de los análisis realizados.

Capítulo 3

Selección de datos

La información a utilizar en este estudio, se puede clasificar en 3 grandes grupos:

- **Información personal y sociodemográfica**: Corresponde a información descriptiva del paciente, su identificación y variables de carácter demográfico y social, como por ejemplo sexo, edad y ocupación entre otras.
- **Información clínica y de diagnóstico**: Corresponde a información asociada al ingreso del paciente en el servicio de salud y a su historia clínica (desde la perspectiva de salud mental), aquí se cuenta por ejemplo con el tipo de ingreso, diagnóstico, y en el caso de que la persona haya tenido un intento de suicidio se incluye información del intento y de la escala de riesgo - rescate, entre otra información.
- **Instrumentos clínicos**: Corresponden a cuestionarios que se le realizan a los pacientes con el objetivo de entender y crear un perfil clínico del paciente, cada cuestionario entrega una clasificación con el estado del paciente en dicho instrumento, según un puntaje calculado en base a sus respuestas.

Así, con estos 3 grandes grupos de información se genera la base de datos inicial del proyecto. Esta es proporcionada mediante un archivo Excel, directamente por los especialistas en salud mental con quienes se realiza el presente trabajo.

Antes de ahondar en cada una de estas fuentes de información, se entregarán algunos aspectos generales de la base de datos.

3.1. Acerca de la Base de Datos

La base de datos del proyecto contiene la información de 707 pacientes consultantes sobre salud mental, esto es clave para entender los datos y el trabajo que se ha de desarrollar, pues la población en estudio no es cualquier población, sino una bien particular, son personas consultantes de salud mental, no necesariamente asociada a un riesgo suicida, pero que dista de ser una población general, de ahí el cuidado que se debe tener al querer extender los

resultados de este trabajo a población general.

La información de los pacientes fue recopilada durante un periodo de 5 años, entre marzo del 2010 y enero del 2015. Si bien para el análisis del riesgo suicida, el factor temporal no afecta mayormente los análisis, la gran ventana de tiempo asociado a la recolección de información sí afecta la metodología de recolección, como se verá en el capítulo de pre procesamiento, ya que dado el gran periodo de recolección de datos, hay información sociodemográfica y de instrumentos que fue agregada tardíamente, lo que produce una baja completitud para algunos de los campos de la base de datos.

Otro aspecto a considerar es que los datos son recolectados de tres centros distintos de salud, todos ellos de la región metropolitana, si bien en los análisis esto no es relevante, esto es importante para entender los datos y es ocupado en la imputación de estos. Cada uno de los centros puede ser identificado con un grupo socioeconómico distinto, por lo que la demografía de cada sector puede variar de un centro a otro. Los centros de salud donde se recopiló información son:

- Complejo asistencial Dr. Sótero del Río Servicio de Salud Metropolitano Sur oriente.
- Clínica UC San Carlos de Apoquindo Red Salud UC Christus.
- Centro Médico San Joaquín Red Salud UC Christus.

Salvo algunas posibles diferencias demográficas entre los tres centros, el centro de salud por sí mismo no es en esencia relevante para el modelamiento, ya que el modelo final debe ser una herramienta transversal, y debe estar basado en las preguntas de los cuestionarios que resulten relevantes y no del centro donde es atendido el paciente.

Adicionalmente para mantener la confidencialidad de la información (recordar que se está trabajando con información personal de cientos de pacientes) la información de identificación se ha anonimizado y se utiliza un ID correlativo para la diferenciación de cada caso. Notemos que para la aplicación de las técnicas de data mining este campo es sólo un identificador del registro y no tiene una mayor importancia en los análisis que se realizan.

Así, la base de datos inicial cuenta con 707 casos y 343 campos, distribuidos entre información del paciente, demográfica, clínica y de los instrumentos. En las secciones siguientes se revisarán los contenidos y las escalas de medición de cada componente de la base de datos.

3.2. Variable Objetivo

La variable Objetivo en la base de datos corresponde al campo denominado Grupo P1, donde en base a diferentes instrumentos clínicos y al conocimiento del mismo especialista durante la evaluación clínica, se le asigna un nivel de riesgo a cada paciente. Los posibles valores

presentes en el campo corresponden a:

- Riesgo suicida alto
- Riesgo suicida bajo
- Ideación suicida
- Grupo de comparación

Para el análisis y estudio presente, y como se mencionó en el capitulo 1 se utiliza una recodificación del campo anterior, agrupándose los primeros 3 valores (riesgo suicida alto, riesgo suicida bajo e ideación suicida) en un único valor, que se ha denominado Grupo de riesgo suicida. Esta clasificación está basada en el criterio experto de los psicólogos y doctores con quienes se ha desarrollado esta memoria. Así, los nuevos valores de la variable objetivo son:

- Grupo de riesgo suicida
- Grupo de comparación o control

Se debe considerar que el grupo de comparación es una muestra de la población que si bien consulta por salud mental, no lo hace por algún intento o ideación suicida. Luego, este grupo servirá para poder discriminar y encontrar aquellos patrones en los datos que caracterizan a las personas en la zona de riesgo suicida y las diferencian de aquellos en el grupo de control.

3.3. Información demográfica, clínica y de diagnóstico

La información demográfica está asociada a 11 variables descriptivas de los pacientes, estas son: Sexo, edad, escolaridad, con quien vive, ocupación, ocupación agrupada, estado civil, número de hijos, uso de drogas, consumo de alcohol y tasa de consumo.

Por su parte las variables de carácter clínico y/o de diagnóstico son 29, asociadas a: Centro de salud, servicio, días de hospitalización, motivo de la hospitalización, motivo de la consulta ambulatoria, ideación suicida en último mes, intento suicida en último mes, método del intento suicida, fecha del intento suicida, diagnóstico, diagnóstico según ejes 2-5, puntaje de riesgo-rescate, gravedad del intento, nivel de intención suicida y riesgo suicida general.

Si bien el campo asociado al centro de salud está asociado a la consulta/hospitalización del paciente, también se puede considerar a éste como una variable de carácter descriptivo. Lo anterior, dado que cada uno de los 3 centros de salud está asociado a un sector socioeconómico y demográfico distinto.

En la tabla 3.1 se muestran los campos antes descritos, con su correspondiente descripción.

Campo	Descripción
CSALUD	Centro de Salud del paciente
SERVICIO	Tipo de servicio, si es ambulatorio u hospitalización
SEXO	Sexo del paciente
EDAD	Edad del paciente
ESCOLARIDAD	Nivel de estudios del paciente
VIVECON	Indica con quien vive
OVUP	Ocupación
OCUPAG	Ocupación agrupada.
ECIVIL	Estado civil
HIJOS	Indica si tiene hijos y cuantos
USODROGAS	Indica si usa drogas y que tipo
CONSALCOHOL	Indica si consumo alcohol y que tipo
TCONSUMO	Tasa de consumo
DHOSP	Dias de hopitalizacion (por cualquier motivo, no necesariamente ideacion/intento suicida)
MOTHOSP	Motivo de la hospitalización (por cualquier motivo, no necesariamente ideacion/intento suicida)
MC_ACT_AMB	Motivo de consulta actual (paciente ambulatorio y/o hospitalizado).
IDSUI_UMES	Indica si tuvo ideación suicida en el último mes
INSUI_UMES	Indica si tuvo intento suicida en el ultimo mes
MET_INSUI	Metodo utilizado si intentó suicidarse en el ultimo mes
MET_INSUI_AG	Metodo agrupado
CDTA_ANTERIOR	Indica si tiene conducta suicida anterior al último mes
NVECES_CDTA_ANT	Número de veces que ha tenido algún tipo de conducta suicida (intento, autoagresión y/o ideación)
CONDUCTA	Tipo de conducta, intento, ideación o autoagresión
MET_CDTA_AG	Metodo para autoagredirse (el anterior)
GATILLANTE	Factor gatillante de la conducta anterior
F_INSUI	Fecha del intento de suicidio
MC_MED_TRAT	Motivo de consulta
DIAG	Diagnostico
DIAG_AG	Diagnostico agrupado
EJE2	Diagnóstico del eje2: Tipo de trastorno de la personalidad
EJE3	Diagnóstico del eje3: Enfermedades médicas
EJE4	Diagnóstico del eje4: Problemas psicosociales y ambientales
EJE5	Evaluación actividad global
PTJE_RIESGO	puntaje de riesgo
PTJE_RESCATE	puntaje de rescate
PTJE_RIESGO_RESCATE	puntaje riesgo-rescate
GRAVGRALINTENTO	Gravedad general del intento
NCA	puntaje de intencion suicida
NIVELINTENCIÓNSUICIDA	Nivel de la intencion suicida
RIESGO_SUIGRAL	Riesgo suicida general
INTENCIÓN DEL INTENTO	Intención del suicidio
¿POR QUÉ QUEDÓ VIVO?	Indica porque quedó vivo el paciente depués de un intento suicida
I.S. ACTUAL- GATILLANTE	Factor gatillante de la conducta actual

Tabla 3.1: Diccionario de campos asociados a información demográfica, clínica y de diagnóstico

3.4. Descripción de los instrumentos

Dentro de las herramientas con que los especialistas en salud mental cuentan, se encuentran una serie de cuestionarios, los que respondidos por los pacientes sirven para orientar al profesional respecto al estado de ánimo, rabia, perfil, funcionamiento familiar, etc. Los profesionales usan esta información para generar un perfil de cada paciente, detectar donde se observan carencias para trabajar sobre estos puntos y reforzar aquellos aspectos protectores para cada paciente. Estos instrumentos no son exclusivos (al menos no todos) de realizar sobre pacientes en riesgo suicida, sin embargo, su aplicación sobre esta población en riesgo ayuda de mejor manera a su entendimiento, identificación, y con el presente trabajo, su pronóstico sobre el nivel de riesgo para cada paciente consultante de salud mental.

Dentro de la base de datos, 298 campos están asociados a estos instrumentos, ya sea como preguntas individuales o campos resumen para cada uno de los cuestionarios, como se observa en la tabla 3.2

Instrumento	**Descripción corta**	**# de preguntas**	**# de preguntas resumen**
APGAR	Cuestionario de funcionamiento familiar	5	2
OQ	Outcome Questionare	45	8
DEQ	Cuestionario de vivencias depresivas	66	2
STAXI	Estado, característica, control y expresión de la rabia	44	10
PBI_MA	Apego percibido de la madre	25	5
PBI_PA	Apego percibido del padre	25	5
RFL	Razones para vivir	50	6

Tabla 3.2: Resumen de los instrumentos clínicos incluidos en el presente estudio

Si bien cada instrumento tiene sus puntajes y niveles asociados, estos no son considerados para los análisis, centrándose el estudio sobre las preguntas únicamente. En lo que sigue se describen a modo general cada uno de estos cuestionarios.

3.4.1. APGAR

El APGAR es un instrumento de cinco preguntas que evidencia cómo percibe el paciente el nivel de funcionamiento de la unidad familiar de forma global, en un momento dado.

Cada una de las preguntas valoran un aspecto diferente de la dinámica familiar en las áreas de adaptación, vida en común, crecimiento, afecto y resolución. El paciente responde a cada una de las preguntas con los valores 0 (Nunca), 1 (Algunas veces) o 2 (Siempre), respecto a la respuesta que mejor refleje la frecuencia con que el paciente está de acuerdo en cada una

de las preguntas/afirmaciones. En la figura 3.1 se muestran las preguntas del instrumento APGAR.

Cuestionario de funcionamiento familiar

(Smilkstein, 1978)

Por favor indique con un X el espacio que refleje mejor la frecuencia con que está de acuerdo con las siguientes afirmaciones respecto a su familia

	Nunca 0	Algunas veces 1	Siempre 2
Me satisface la ayuda que recibo de mi familia cuando tengo algún problema y/o necesidad			
Me satisface la forma como mi familia habla de las cosas y comparte los problemas conmigo			
Me satisface como mi familia acepta y apoya mis deseos de emprender nuevas actividades			
Me satisface como mi familia expresa afecto y responde a mis emociones como rabia, tristeza o amor			
Me satisface cómo compartimos en familia el tiempo de estar juntos, los espacios en la casa o el dinero			

Figura 3.1: Instrumento APGAR.

Posteriormente, se obtiene información sobre la satisfacción familiar del paciente con cada uno de los componentes funcionales de la dinámica familiar. La información obtenida proporciona datos básicos sobre el nivel de dinámica familiar, dando al especialista de la salud (enfermera, psicólogo, etc.) una idea de qué áreas necesitan una valoración e intervención más detallada y de las fuerzas familiares que pueden utilizarse para solucionar otros problemas familiares.

Su aplicación no es exclusiva de pacientes de salud mental, y es usado de manera amplia en situaciones como embarazos, depresión postparto, condiciones alérgicas, hipertensión arterial, entre otros. Sin embargo, su utilización en pacientes consultantes de salud mental es un aporte notable e importante para determinar factores de riesgo suicida.

3.4.2. OQ (Outcome Questionnaire)

El instrumento OQ u *Outcome Questionnaire*[1] es un cuestionario de auto-evaluación de 45 preguntas, creado por Lambert en 1996 [12]. Este instrumento sirve para obtener información acerca del estado general de la persona, es decir, cómo se siente. Además permite medir progresos en psicoterapia a través de mediciones sucesivas, como se muestra en [14]. La evaluación se realiza en tres áreas o subescalas:

- Sintomatología ansiosa y depresiva (25 preguntas).
- Satisfacción con relaciones interpersonales (11 preguntas).
- Satisfacción con el rol social (9 preguntas).

Cada una de las preguntas del OQ corresponde a una afirmación, donde el paciente debe responder en una escala de 0 a 4 de acuerdo a lo que mejor describa como se ha sentido en los últimos siete días. En la figura 3.2 se muestra un extracto con las primeras 10 preguntas del OQ.

	Nunca	Casi nunca	A veces	Con frecuencia	Casi siempre
1. Me llevo bien con otros	4	3	2	1	0
2. Me canso rápidamente.	0	1	2	3	4
3. Nada me interesa	0	1	2	3	4
4. Me siento presionado (estresado) en el trabajo/escuela/dueña de casa	0	1	2	3	4
5. Me siento culpable.	0	1	2	3	4
6. Me siento irritado, molesto.	0	1	2	3	4
7. Me siento contento con mi matrimonio/pareja.	4	3	2	1	0
8. Pienso en quitarme la vida.	0	1	2	3	4
9. Me siento débil.	0	1	2	3	4
10. Me siento atemorizado.	0	1	2	3	4

Figura 3.2: Primeras 10 preguntas del instrumento OQ.

Notar que si bien la escala de respuesta $(0-4)$ está invertida en algunos casos, de acuerdo a la formulación de la pregunta, siempre un valor de respuesta mayor está asociado a un mayor riesgo o una condición clínica de cuidado.

3.4.3. DEQ (Depressive Experiences Questionnaire)

El cuestionario de vivencias depresivas o DEQ por sus siglas en inglés, es un cuestionario de auto-evaluación compuesto de 66 preguntas. El DEQ está diseñado para diferenciar entre tres tipos de estilos de personalidad en relación a la vivencia depresiva: eficaz, dependiente y autocrítico. Donde tanto el perfil dependiente como el autocrítico están asociados con un

[1]Para el presente estudio se utiliza el OQ 45.2

mayor riesgo por depresión y por psicopatologías en general.

A diferencia del APGAR y el OQ, el DEQ no tiene subescalas definidas, pues todas las preguntas contribuyen en mayor o menor medida a cada uno de los 3 perfiles, luego el grado de prevalencia de cada perfil depende de las 66 preguntas del instrumento.

En cada pregunta, el paciente completa con el valor con el cual se siente más representado, usando una escala numérica entre 1 y 6, donde 1 significa "en total desacuerdo" y 6 significa "en total acuerdo".

Se entrega una muestra de las preguntas del instrumento en la figura 3.3.

CUESTIONARIO DE EXPERIENCIAS DEPRESIVAS
(DEQ)

Nombre:

Sidney J. Blatt, Ph.D.; Carrie E. Schaffer, Ph.D.; Susan A. Bers, Ph.D.; Donald M. Quinlan, Ph.D., 1989.

Traducción y Adaptación al Idioma Español

Por Ps. Susana Morales Silva. Doctora en Psicoterapia P. Universidad Católica de Chile-U de Chile sumorales@med.puc.cl

Tomado de la primera adaptación al español de la escala para adolescentes, de Humberto L. Persano MD, Ph.D., 2003., Mental Health Department, School of Medicine, University of Buenos Aires.

Se le solicita que lea cada una de las siguientes frases y decide que tan bien te describen. Luego haz un círculo alrededor del número más apropiado para cada ítem mencionado, basándote en la escala que se presenta a continuación

En total desacuerdo 1	2	3	4	5	6	En total acuerdo 7

1. Establezco mis objetivos a un nivel muy alto.

1	2	3	4	5	6	7

2. Sin el apoyo de los que están cerca de mí me encontraría desamparado (a).

1	2	3	4	5	6	7

3. En general, me siento más conforme con mis planes y metas, que intentando alcanzar objetivos más elevados.

1	2	3	4	5	6	7

Figura 3.3: Primeras 3 preguntas del cuestionario de experiencias depresivas (DEQ).

3.4.4. STAXI (State-Trait Anger Expression Inventory)

En líneas generales, el STAXI permite evaluar la experiencia y expresión del enojo, en sus dos dimensiones (estado y rasgo) y en sus tres direcciones (expresión, supresión y control del enojo), para un total de 5 subescalas dentro del mismo instrumento.

El cuestionario consta de 44 preguntas, separadas en 3 partes, la primera (10 preguntas) evalúa la dimensión del estado de la rabia, la segunda (10 preguntas) evalúa la dimensión asociada al rasgo de la rabia. Mientras que la tercera parte (24 preguntas), evalúa su tres

direcciones, expresión (8 preguntas), supresión o guardado (8 preguntas) y control de la rabia (8 preguntas).

El paciente, al igual que en el resto de los cuestionarios, responde a cada pregunta utilizando una escala numérica de acuerdo a la opción que mejor lo identifica. La escala toma los valores 1, 2 y 3, donde 1 significa "en total desacuerdo", 2 significa "medianamente de acuerdo" y 3 corresponde a "muy de acuerdo".

3.4.5. PBI (Parental Bonding Madre y Padre)

Este cuestionario consta de 25 afirmaciones, cada una de las cuales se refiere a cómo recuerda el paciente a su padre/madre en su infancia, esto es, hasta los 16 años. Ambos instrumentos (el del padre y la madre) son iguales, de 25 preguntas cada uno, y sólo cambia a quien va dirigido (si es al padre o la madre).

Cada afirmación es seguida por una escala de valor, donde el paciente escoge la que más lo representa. La escala toma los siguientes valores:

- Muy en desacuerdo.
- Moderadamente en desacuerdo.
- Moderadamente en acuerdo.
- Muy en acuerdo.

Al igual que lo que sucede con el instrumento OQ, en este caso el correspondiente valor numérico que va desde 0 a 4, puede estar invertido, es decir, para algunas preguntas un valor de 0 puede significar "muy de acuerdo", mientras que en otros casos el valor 0 corresponde a "muy en desacuerdo". Dado lo anterior es importante que el paciente siempre responda respecto al grado con que está de acuerdo en la afirmación y es el especialista quien se encarga de transcribir el correspondiente valor numérico de cada pregunta y calcular los puntajes del cuestionario.

Nivel de cuidado	**Nivel de sobreprotección**	**Tipo de vínculo**
Bajo	Baja	Vínculo ausente
Bajo	Alta	Control sin afecto
Alto	Alta	Constricción cariñosa
Alto	Baja	Vínculo óptimo
Promedio	Promedio	Vínculo promedio
Alto	Promedio	Vínculo promedio
Bajo	Promedio	Vínculo promedio
Promedio	Alta	Vínculo promedio
Promedio	Baja	Vínculo promedio

Tabla 3.3: Tipos de vinculo con el padre/madre analizados por el instrumento PBI.

El instrumento mide el apego percibido en 2 subescalas, cuidado y sobreprotección, generando 5 perfiles o tipos de vínculos distintos según se aprecia en la tabla 3.3.

3.4.6. RFL (Reasons for living)

El RFL es un instrumento utilizado para determinar las razones por las cuales una persona podría verse disuadida de atentar contra la propia vida. El instrumento consta de 60 preguntas, cada una con una afirmación que dan las personas para no suicidarse, y el objetivo del paciente es calificar cada afirmación de acuerdo al grado de importancia que tiene para él para no cometer suicidio. La escala de valores es según se indica a continuación.

1. No es importante.
2. Muy poco importante.
3. Poco importante.
4. Importante
5. Muy importante.
6. Extremadamente importante.

Así, el paciente califica cada afirmación con la opción que mejor lo representa respecto a la afirmación en cuestión.

El RFL se puede dividir en 6 subescalas, las que se indican a continuación:

- Objeciones morales.
- Creencias en capacidad de supervivencia y afrontamiento.
- Miedo a la muerte y desaprobación social.
- Responsabilidad con la familia.
- Preocupación por los hijos.
- Percepción de incapacidad para suicidarse.

Adicionalmente a las 6 subescalas antes mencionadas, hay 2 preguntas dentro del instrumento que no están incluidas en ninguna de las subescalas, esto ya que son preguntas de una versión anterior del instrumento.

Capítulo 4

Preprocesamiento de la información

En este caso, la base de datos ya se encuentra consolidada desde su origen, lo cual fue realizado por el área de salud mental que trabaja en este proyecto, por lo que no es necesario hacer un trabajo para unir las diferentes fuentes de datos. Sin embargo, al considerarse datos de tres centros de salud distintos, y durante un periodo de tiempo de 5 años, se requiere un proceso de limpieza y estandarización de la base de datos recibida.

Durante el depurado y limpieza de la base de datos se espera poder gestionar los valores perdidos y/o nulos y manejar las inconsistencias presentes, producto del largo periodo en la toma de datos (diferentes metodologías a lo largo del tiempo, lo que se traduce en instrumentos y/o campos incompletos durante algunos periodos de tiempo en la recolección).

4.1. Recodificación y corrección de errores en la base de datos

La primera tarea dentro del preprocesamiento de los datos consiste en estandarizar los nombres de los campos (eliminando espacios, acentos y signos de puntuación entre otras cosas) y recodificar todos los valores numéricos de la base de datos en su correspondiente valor categórico o nominal. Esto se hace ya que cada campo viene codificado en una escala numérica diferente, por ejemplo el campo VIVECON viene codificado con los valores 1, 2 y 3, donde un 1 corresponde a "sólo(a)", el valor 2 corresponde a "pareja", y el valor 3 se refiere a "familia".

Lo anterior se realiza en una primera instancia para poder realizar los análisis y transformación de los datos de forma más eficiente, ya que de esta manera es más claro saber a qué se refiere cada uno de los valores en la base de datos sin necesidad de revisarlos en el diccionario de la base de datos. Sin embargo, para la etapa de modelamiento, aquellas variables relevantes son nuevamente transformadas, creando dummies para cada categoría.

Por su parte las respuestas de las preguntas son valores numéricos en una escala prestablecida (distinta para cada instrumento), pero que poseen un "orden", ya que cada pregunta intenta determinar el grado de aceptación/ocurrencia de lo enunciado. Luego estos valores se consideran como numéricos u ordinales, sin ser recodificados como las variables descriptivas.

Durante el preprocesamiento también se manejan aquellos valores fuera de rango (*outliers*), erróneos y nulos. Éstos son analizados con ayuda de los especialistas, e incluso en algunos casos volviendo directamente a la fuente, revisando nuevamente los cuestionarios que los pacientes respondieron. Un error común asociado a la base de datos, ocurre en el instrumento RFL, donde si ben las preguntas del cuestionario tienen un rango válido de 1 a 6, existen varios casos con valores respondidos con un 7 e incluso 8.

Estos casos son tratados caso a caso. Así, por ejemplo, en casilleros, donde dice 7 y el límite es 6, se revisa que en la subescala (es decir, las preguntas pertenecientes a esa sub dimensión clínica o de la personalidad), estuviera poniendo 6, corrigiendo entonces a este mismo valor.

Finalmente, dentro de la recodificación se procede a reclasificar la variable objetivo en dos clases a partir de la variable GRUPO_P1, como se expuso en el capítulo 3.2. Esto obedece al enfoque que se le desea dar al modelo generado, más que distinguir un nivel de riesgo suicida, se desea clasificar aquellos que están en riesgo de los que no lo están.

4.2. Selección de campos

4.2.1. Eliminación de campos de acuerdo a su relevancia para el proyecto

Una vez estandarizados y corregidos los valores de la base de datos se procede a su depurado. El primer paso corresponde a un filtro preliminar de los campos a utilizar, considerando aspectos de relevancia de la información como aporte al modelo a desarrollar. El trabajo en este ámbito se divide en los siguientes pasos:

1. Eliminación de campos asociados a puntajes o nivel (resumen) de los instrumentos: Uno de los objetivos de este trabajo es encontrar un número acotado de variables que ayuden a identificar el riesgo suicida de un paciente, por lo que no se utilizarán estos campos, ya que se tendría que considerar todas las preguntas del instrumento correspondiente para poder calcular estos campos.
2. Eliminación de campos asociados a la escala riesgo-rescate, escala de intención suicida de Pierce y de la variable Grupo_P1: La escala de intención suicida, y la escala de riesgo-rescate son utilizadas para la definición del riesgo suicida, por lo que estos campos no pueden ser utilizados en el modelo predictivo (serían redundantes).
3. Eliminación de los campos asociados a la conducta anterior: El que un paciente tenga

antecedentes de intentos suicidas ya es un fuerte indicador del nivel de riesgo de ese paciente y por tanto sirve para definir su nivel de riesgo suicida. Sin embargo, el análisis que se desea lograr con este trabajo está enfocado en la utilización de las preguntas de los instrumentos para la detección del riesgo por lo que esta variable no resulta útil para el objetivo de este trabajo.

4.2.2. Eliminación de campos con gran porcentaje de valores perdidos

Una vez ya eliminado aquellos campos no relevantes para el desarrollo del modelo, se procede al manejo de los valores perdidos, lo cual se realiza en dos frentes: El primero, eliminar aquellos campos de mala calidad, es decir, que posean un gran porcentaje de registros sin información, y el segundo, eliminar/imputar aquellos registros con un rango manejable de valores perdidos.

Dado el reducido tamaño de la base datos, en cuanto a cantidad de registros, el manejo de valores perdidos se debe realizar con cuidado. Pues sólo se cuenta con información de 707 pacientes, mientras que la cantidad de campos asciende a 343. Así, para no caer en el sobreajuste del modelo, se favorece la eliminación de campos (y de esta manera también reducir la dimensionalidad del problema) por sobre la eliminación de registros, eliminando sólo una pequeña parte de los casos.

Así, el primer paso consiste en eliminar aquellos campos que tengan bajo un 95 % de completitud. Esto, para las variables demográficas, mientras que para los instrumentos se considerará no válidos si en promedio (al considerar todas las preguntas del instrumento) tienen menos de un 95 % de completitud.

En la tabla 4.1 se puede ver el porcentaje de completitud para cada uno de los campos descriptivos en la base de datos, de donde dado la regla anterior se eliminan los campos: MOTHOSP, USADROGAS, CONSALCOHOL, TCONSUMO, EJE2, EJE3, EJE4, EJE5, OVUP, MC_ACT_AMB.

Por su parte, al calcular el porcentaje de completitud promedio por Instrumento, se observa de la tabla 4.2 que los instrumentos Parental Bonding Madre y Parental Bonding Padre poseen una completitud inferior a la mínima requerida, por lo que son eliminados.

Así, luego de esta selección de campos, la nueva base de datos baja de 343 a 224 campos, esto es, 1 variable objetivo y 223 atributos distribuidos en las siguientes categorías, según se aprecia en la tabla 4.3.

Campo	% Completo	Registros válidos
CSALUD	100	707
SERVICIO	100	707
SEXO	100	707
OCUPAG	100	707
HIJOS	100	707
EDAD	99,859	706
ECIVIL	99,859	706
ESCOLARIDAD	98,727	698
DIAG_AG	97,171	687
DHOSP	97,03	686
DIAG	97,03	686
VIVECON	95,898	678
MOTHOSP	91,231	645
USODROGAS	85,007	601
CONSALCOHOL	85,007	601
TCONSUMO	62,801	444
EJE4	42,857	303
EJE2	39,18	277
EJE5	31,259	221
EJE3	29,137	206
OVUP	0	0
MC_ACT_AMB	0	0

Tabla 4.1: Porcentaje de registros válidos para los campos descriptivos.

Instrumento	% Completo (promedio)	Registros válidos (promedio)
APGAR	99,43	703
DEQ	99,14	701
STAXI	98,87	699
RFL	96,85	685
OQ	95,31	674
PBI_PRE_MA	75,79	536
PBI_PRE_PA	72,67	514

Tabla 4.2: Porcentaje de completitud para cada instrumento. Se considera el valor de completitud promedio entre todas las preguntas del instrumento.

Categoría	Nro. de campos
Objetivo	1
ID	1
Variables descriptivas	12
APGAR	5
OQ	45
DEQ	66
STAXI	44
RFL	50

Tabla 4.3: Cantidad de variables según categoría después de realizar la selección de campos inicial.

4.3. Imputación de valores perdidos

Ahora para la imputación de los valores perdidos, se trabaja sobre cada uno de los campos/instrumentos incompletas. En la tabla 4.4 se muestra la cantidad de registros incompletos para cada uno de los campos descriptivos y el promedio de campos válidos para cada instrumento.

Campo	% Completo	Registros válidos	Valores nulos
ID	100,00 %	707	0
OBJETIVO	100,00 %	707	0
CSALUD	100,00 %	707	0
SERVICIO	100,00 %	707	0
SEXO	100,00 %	707	0
EDAD	99,86 %	706	1
ESCOLARIDAD	98,73 %	698	9
VIVECON	95,90 %	678	29
OCUPAG	100,00 %	707	0
ECIVIL	99,86 %	706	1
HIJOS	100,00 %	707	0
DHOSP	97,03 %	686	21
DIAG	97,03 %	686	21
DIAG_AG	97,17 %	687	20
APGAR (promedio del instrumento)	99,43 %	703	4
DEQ (promedio del instrumento)	99,14 %	701	6
STAXI (promedio del instrumento)	98,87 %	699	8
RFL (promedio del instrumento)	96,85 %	685	22
OQ (promedio del instrumento)	95,31 %	674	33
TOTAL DE CAMPOS	76,94 %	544	163

Tabla 4.4: Porcentaje de completitud para los campos descriptivos e instrumentos despues de la selección de campos.

El medio más común para tratar valores perdidos en los datos es mediante listwise deletion, es decir, eliminando todos aquellos casos que poseen algún valor perdido. Si los datos perdidos siguen un patrón MCAR, como se asume en este caso, el procedimiento no agrega sesgos a los datos. Sin embargo, disminuye el poder del análisis al disminuir el tamaño efectivo de la muestra. Dado que en este caso el número de registros con valores nulos es alto (sobre el 23 % de los registros posee al menos un valor perdido), la técnica de listwise deletion no resulta ser la más adecuada. Así, dado que la base de datos posee un número reducido de registros, se privilegian otros métodos de imputación en vez del listwise deletion, para así poder mantener un tamaño adecuado de la muestra en estudio.

Finalmente, la imputación de los datos pasa por la combinación de varias técnicas de imputación, siendo personalizada y de acuerdo a cada uno de los campos a imputar. Las técnicas utilizadas son:

- Criterio experto.
- Medias condicionadas.
- Distribución condicionada (RHD).

Donde además de las técnicas asociadas a los métodos clásicos en estadística, en este trabajo también se hace uso de una regla de imputación "experta", es decir, basada en el criterio experto de los especialistas en salud mental. Dicha técnica es llamada para efectos de este trabajo como "imputación por valor de subescala" (IVS).

4.3.1. Imputación por valor de subescala (IVS)

Esta regla consiste en imputar aquellos valores perdidos "aislados" usando el promedio de las respuestas del mismo paciente dentro de la subescala asociada a la pregunta sin información. El término valor perdido "aislado" es utilizado para referirse a aquellas respuestas perdidas de un paciente dentro de un instrumento, pero donde se cuenta con información del resto de las preguntas dentro de la misma subescala (esto es válido para los instrumentos STAXI, OQ y RFL).

La motivación del método consiste en notar que para cada paciente, las preguntas dentro de una subescala de un instrumento particular están altamente correlacionadas. Lo anterior sucede por construcción misma de los instrumentos, para que al calcular los puntajes por subescala se tenga una cierta robustez en el dato, y así también tener una cierta consistencia en las respuestas del paciente. A modo de ejemplo para clarificar aún más lo anterior, al considerar 2 de las subescalas del RFL se tiene lo siguiente:

Subescala: Responsabilidad con la familia.

- **Pregunta 1**: Soy responsable y estoy comprometido(a) con mi familia.
- **Pregunta 7**: Mi familia podría pensar que no los quise.
- **Pregunta 9**: Mi familia depende de mí y me necesita.
- **Pregunta 16**: Quiero y disfruto mucho de mi familia, no la querría dejar.
- **Pregunta 30**: Le dolería mucho a mi familia y no quiero que sufran.
- **Pregunta 47**: No querría que mi familia se sintiera culpable después.
- **Pregunta 48**: No querría que mi familia pensara que soy egoísta ni cobarde.

Subescala: Objeciones morales.

- **Pregunta 5**: Pienso que sólo Dios tiene el derecho a terminar una vida.
- **Pregunta 23**: Tengo miedo de ir al infierno.
- **Pregunta 27**: Mis creencias religiosas lo prohíben.
- **Pregunta 34**: Lo considero "malo" moralmente.

De donde se observa claramente una relación entre las preguntas de una misma subescala, no así entre distintas subescalas o entre pacientes distintos.

Así, este criterio o regla experta considera la valoración promedio de las preguntas dentro de una misma subescala para cada paciente. Es decir, en cuanto a una sub dimensión clínica o

de la personalidad, el paciente se consideró de cierta manera, lo que es replicable en los datos faltantes que pertenezcan a la misma sub dimensión o aspecto clínico o de la personalidad. Luego la imputación por valor de subescala parece ser el método más adecuado para completar estos valores perdidos aislados.

En lo que sigue se analiza por separado cada uno de los campos que requiere imputación de algún tipo.

4.3.2. Campo Edad

De la tabla 4.4 se observa que sólo un registro posee este campo nulo en la base de datos. Luego al identificarlo se tiene lo expuesto en la tabla 4.5:

ID	167
OBJETIVO	Grupo de riesgo suicida
CSALUD	Sotero
SERVICIO	Ambulatorio
SEXO	Mujer
EDAD	Null
ESCOLARIDAD	Universitaria incompleta
VIVECON	Familia
OCUPAG	Desempleado
ECIVIL	Casado
HIJOS	3
DHOSP	0
DIAG	Null
DIAG_AG	Null

Tabla 4.5: Identificación del registro con el campo EDAD sin información.

Para la imputación del valor de edad en este caso se utiliza medias condicionadas, utilizando el promedio de edad para las mujeres que tienen 3 o más hijos y son del centro de salud Sótero del Río. De la tabla 4.6 se observa que la edad promedio en este caso corresponde a 53 años.

SEXO	HIJOS_2	CSALUD	EDAD_Mean	EDAD_Min	EDAD_Max	EDAD_SDev	# de registros
Mujer	0	San Carlos	24,1	14	43	8,019	79
Mujer	0	San Joaquín	28,38	14	67	11,898	45
Mujer	0	Sótero	33,25	14	63	14,422	44
Mujer	1	San Carlos	36,25	23	53	10,154	12
Mujer	1	San Joaquín	41,96	19	68	14,995	23
Mujer	1	Sótero	45,16	19	74	10,97	121
Mujer	2	San Carlos	41	21	59	9,112	32
Mujer	2	San Joaquín	43,5	30	71	12,647	20
Mujer	2	Sótero	50,5	27	71	9,459	62
Mujer	3 o más	San Carlos	48,62	34	68	9,617	26
Mujer	3 o más	San Joaquín	45,91	32	73	12,31	11
Mujer	3 o más	Sótero	53,05	34	76	8,151	89

Tabla 4.6: Edad promedio para las mujeres según número de hijos y Centro de Salud.

4.3.3. Campo ECIVIL

De forma análoga para el campo ECIVIL (Estado civil) se identifica el registro sin información, el cual se detalla en la tabla 4.7.

ID	301
OBJETIVO	Grupo de comparación
CSALUD	Sotero
SERVICIO	Ambulatorio
SEXO	Hombre
EDAD	55
ESCOLARIDAD	Media completa
VIVECON	Familia
OCUPAG	Otra
ECIVIL	Null
HIJOS	1
DHOSP	0
DIAG	Episodio depresivo moderado
DIAG_AG	Episodio depresivo moderado

Tabla 4.7: Identificación del registro con el campo ECIVIL sin información.

Luego, para la imputación del atributo ECIVIL se utilizan los campos SEXO, VIVECON, HIJOS, y se asigna el valor más representativo (moda) para los hombres que viven con la familia y tiene al menos un hijo. De la tabla 4.8 se tiene que el valor a utilizar es "casado".

SEXO	**VIVECON**	**tieneHijo**	**ECIVIL**	**# de registros**
Hombre	Familia	Con_Hijo	Casado	27
Hombre	Familia	Con_Hijo	Soltero	12
Hombre	Familia	Con_Hijo	Separado o divorciado	2
Hombre	Familia	Con_Hijo	Unión libre o conviviente	2
Hombre	Familia	Con_Hijo	Null	1
Hombre	Familia	Con_Hijo	Viudo o separado	1

Tabla 4.8: Cantidad de pacientes según estado civil entre los hombres que viven con la familia y tiene al menos un hijo.

4.3.4. Campo VIVECON

Para este campo se tienen 29 valores perdidos, luego ya no es posible hacer una imputación caso a caso como en los casos anteriores. Para abordar los valores perdidos en este campo se caracterizan aquellos registros con la información de este campo faltante mediante los 6 grupos presentados en la tabla 4.9 y se imputan los datos mediante medias condicionadas (en este caso sería moda y no media), utilizando los campos SEXO, ECIVIL y Estudiante como variables de agrupación.

Grupos	SEXO	Estado civil	Estudiante	EDAD Promedio	# de registros sin información
A	Hombre	Casado o Unión libre o conviviente	No	33	1
B	Hombre	Soltero	No	35,8	4
C	Hombre	Soltero	Si	22,7	3
D	Mujer	Casado o Unión libre o conviviente	No	41,2	10
E	Mujer	Soltero	No	38,5	2
F	Mujer	Soltero	Si	22,3	9

Tabla 4.9: Agrupación de pacientes según sexo, estado civil y si es estudiante o no.

Para la imputación se utiliza la moda del campo VIVECON dentro de estos 6 grupos usando el resto de los datos de la base de datos. Así se obtienen las siguientes estimaciones para cada grupo, mostrados en la tabla 4.10.

Grupo	VIVECON (Moda)	% de la moda en el grupo	# de registros
A	Familia	81 %	30
B	Familia	83 %	33
C	Familia	96 %	47
D	Familia	74 %	174
E	Familia	81 %	81
F	Familia	81 %	81

Tabla 4.10: Moda para el campo VIVECON dentro de los grupos con valores perdidos.

Lo anterior se realiza para todos los grupos salvo para el grupo D, dado que la cantidad de registros a imputar en esta caso es de 10, y la moda solo representa al 74 % de los casos dentro del grupo, por lo que se opta por una imputación de tipo *random hot-deck* (RHD) en vez de usar la moda, y así mantener la distribución de los datos, la cual se muestra en la tabla 4.11. Luego, al aplicar RHD, la imputación de los 10 casos en el grupo D queda como se muestra en la tabla 4.12.

VIVECON	% dentro del grupo D	# de registros en el grupo D
Familia	73,7 %	174
Pareja	25 %	59
Solo	1,3 %	3

Tabla 4.11: Distribución del campo VIVECON dentro del grupo D.

4.3.5. Campo ESCOLARIDAD

Dentro de este campo se tienen 9 registros sin información, los que se caracterizan en la tabla 4.13. De aquí se destaca que todos estos pacientes son mujeres, y en general del centro de Salud de San Carlos. Para realizar la imputación se utilizarán los campos EDAD y OCUPAG.

Notar que dos de los casos son "estudiantes" y tienen entre 21 y 22 años, luego el valor de ESCOLARIDAD más idóneo sería Universitaria incompleta o Técnica incompleta, optándose por universitaria incompleta dado que el centro de salud San Carlos atiende generalmente a

Familia
Familia
Pareja
Familia
Familia
Familia
Familia
Familia
Familia
Familia

Tabla 4.12: Valores obtenidos tras la aplicación de RHD para imputar los valores perdidos en el grupo D.

ID	OBJETIVO	CSALUD	SEXO	EDAD	VIVECON	OCUPAG	ECIVIL	HIJOS
178	Grupo de riesgo suicida	Sotero	Mujer	53	Familia	Dueña de casa	Casado	3
87	Grupo de riesgo suicida	San Carlos	Mujer	21		Estudiante	Soltero	0
102	Grupo de riesgo suicida	San Carlos	Mujer	22		Estudiante	Soltero	0
107	Grupo de comparación	San Carlos	Mujer	53		Industria	Casado	2
86	Grupo de riesgo suicida	San Carlos	Mujer	34		Servicios	Casado	5
97	Grupo de riesgo suicida	San Carlos	Mujer	42		Servicios	Casado	4
105	Grupo de riesgo suicida	San Carlos	Mujer	37		Servicios	Soltero	0
85	Grupo de riesgo suicida	San Carlos	Mujer	46	Pareja	Servicios	Unión libre o conviviente	4
95	Grupo de comparación	San Carlos	Mujer	43		Servicios	Unión libre o conviviente	0

Tabla 4.13: Identificación de observaciones con el campo ESCOLARIDAD incompleto

un grupo socioeconómico de altos ingresos, donde los jóvenes generalmente optan por carreras universitarias y no técnicas.

Para los 7 registros restante, y al igual que en los casos anteriores se calcula la moda del campo ESCOLARIDAD separado por las distintas ocupaciones (campo OCUPAG). Aquí se utiliza el supuesto de que la Escolaridad está ligada en cierto modo a la ocupación de cada paciente y considerando sólo las mujeres (moda condicionada).

Los valores obtenidos se presentan en la tabla 4.14, de donde se observa que a diferencia de los casos anteriores donde la moda representaba el 70 % o más de los casos, aquí, no alcanza el 30 %, esto da cuenta de una mayor variabilidad del campo escolaridad incluso dentro de un mismo subgrupo común. Dado lo anterior, una imputación por medias (o modas en este caso) no es lo más adecuado y se opta por una imputación de tipo RHD para el campo ESCOLARIDAD dentro del subgrupo definido por las mujeres con ocupación agrupada definida como "servicio". Y ya que sólo un registro perdido corresponde a las ocupaciones "dueña de casa" e industria respectivamente, estos si son imputados por la moda.

Para la imputación de los registros con OCUPAG = "servicios", hay que notar que todos estos casos corresponden al centro de salud San Carlos, el cual como se ha dicho anteriormente atiende a un sector socioeconómico de mayores ingresos y por consiguiente de un nivel

OCUPAG	ESCOLARIDAD	% de la moda dentro de OCUPAG
Dueña de casa	Media completa	38,2 %
Industria	Universitaria completa	71,4 %
Servicios	Universitaria completa	29,8 %

Tabla 4.14: Moda del campo ESCOLARIDAD según ocupación agrupada.

educacional mayor. Así, esta variable es relevante para la estimación del campo escolaridad, pues como se aprecia en la tabla 4.15 la distribución de las distintas escolaridades en este grupo cambian.

CSALUD	ESCOLARIDAD	OCUPAG	# de registros	% dentro de la categoría
San Carlos	Universitaria completa	Servicios	25	55,6 %
San Carlos	Postgrado completo	Servicios	7	15,6 %
San Carlos	Técnica completa	Servicios	5	11,1 %
San Carlos	Universitaria incompleta	Servicios	4	8,9 %
San Carlos	Media completa	Servicios	3	6,7 %
San Carlos	Media incompleta	Servicios	1	2,2 %

Tabla 4.15: Distribución del campo ESCOLARIDAD entre las mujeres del Centro de Salud San Carlos y con ocupación agrupada "servicios"

Luego, para este nuevo subgrupo, los datos son imputados mediante RHD en esta categoría. Y así, el resumen de la aplicación de las técnicas de imputación antes descritas se muestran en la tabla 4.16.

ID	SEXO	OCUPAG	ESCOLARIDAD (valor imputado)
178	Mujer	Dueña de casa	Media Completa
87	Mujer	Estudiante	Universitaria Incompleta
102	Mujer	Estudiante	Universitaria Incompleta
107	Mujer	Industria	Universitaria Completa
86	Mujer	Servicios	Universitaria completa
97	Mujer	Servicios	Universitaria completa
105	Mujer	Servicios	Universitaria completa
85	Mujer	Servicios	Técnica completa
95	Mujer	Servicios	Media completa

Tabla 4.16: Resumen de los valores imputados por RHD para el campo ESCOLARIDAD.

4.3.6. Campos DHOSP, DIAG y DIAG_AG

Dado lo diverso del campo DIAG (posee muchas clases) es que este campo es eliminado en la etapa de transformación de los datos, al igual que el campo DHOSP, dado esto estos dos atributos no son imputados, ya que no son relevantes para los análisis posteriores.

Por otra parte uno de los registros con información perdida en el campo DIAG_AG (20 casos en total) es completada con el campo DIAG, mientras que los 19 casos restantes se

recodifican en la misma variable con una nueva clase que se define como "sin información".

Luego de aplicadas las primeras reglas de imputación de datos a las variables descriptivas, la cantidad de registros eliminados si se usara listwise deletion disminuye a 92 casos, esto es, un 13 % de los casos, lo que es una gran mejora considerando que solo se intervino 6 campos, y asociados a información descriptiva. Sin embargo, para extraer aún más la información proveniente de los datos y evitar la pérdida de información se sigue con el proceso de imputación para las variables asociadas a los instrumentos.

4.3.7. Imputación datos perdidos en Instrumentos

El primer paso consiste en imputar aquellos registros perdidos "aislados", es decir, preguntas individuales dentro de un instrumento que se encuentren sin información, pero donde el resto (o al menos un grupo) de preguntas contienen información dentro del mismo instrumento. Esto es válido para los instrumentos STAXI, OQ y RFL, ya que se pueden dividir en subescalas, luego una primera imputación de los datos perdidos se hace ocupando el método de imputación por valor de subescala, descrito al inicio de esta sección.

El paso siguiente en la imputación de los instrumentos es eliminar aquellos registros con gran porcentaje de nulos dentro de los instrumentos, la idea detrás de esto es que con la imputación de datos se quiere mantener la información entregada por los pacientes, aunque esta sea parcial, siempre y cuando la cantidad de valores perdidos a imputar sea razonable en comparación a lo que está completo.

Finalmente el último paso es imputar aquellos instrumentos que se encuentran perdidos por completo. Así el marco general para la imputación de los datos de instrumentos se resume a continuación, esto es válido para la imputación de valores perdidos para los 5 instrumentos analizados.

1. Identificar subescalas dentro de cada instrumento y calcular el promedio de las respuestas dentro de cada subescala para cada paciente (en caso que el instrumento no tenga subescalas esto se omite, como sucede con el DEQ).
2. Para valores perdidos aislados se imputa el valor perdido por el correspondiente valor promedio redondeado de la subescala (IVS), el redondeo se realiza para utilizar el mismo tipo de dato de las preguntas, el cual es discreto.
3. En el caso que el paciente tenga toda una subescala incompleta o el instrumento no tiene subescalas se realiza imputación por medias condicionadas usando el resto de los campos dentro del mismo instrumento y las variables descriptivas.
4. Eliminación de la base de datos de aquellos registros con alto porcentaje de valores perdidos.
5. Imputación de Instrumentos completos utilizando técnicas de RHD.

Para el paso 4, una vez ya imputados estos valores perdidos "aislados" en los instrumentos,

se procede a determinar la cantidad de registros que tiene sin información un instrumento completo. La cantidad de valores perdidos para cada instrumento se muestran en la tabla 4.17.

Instrumento	% de valores perdidos (promedio)	# de registros con valores perdidos (promedio)
APGAR	0,6 %	4
DEQ	0,7 %	5
OQ	4,7 %	33
RFL	2,8 %	20
STAXI	0,8 %	6

Tabla 4.17: Porcentaje de valores perdidos promedio según instrumento.

Así también, al analizar los patrones de datos perdidos se obtiene lo representado en la figura 4.1 y figura 4.2. Además de la tabla 4.18 se tiene que sólo 6 registros poseen más de un instrumento sin información simultáneamente.

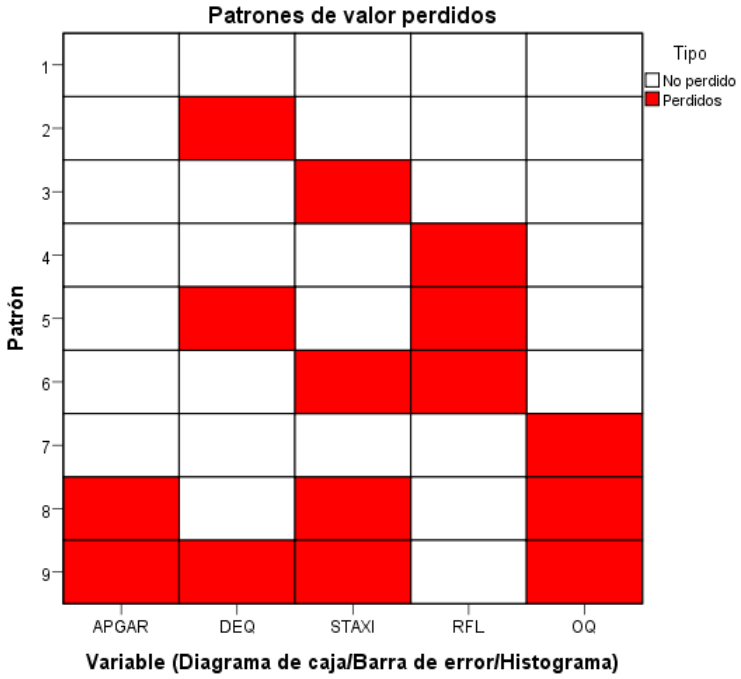

Figura 4.1: Patrones de valores perdidos.

Luego, se eliminan los registros asociados a los patrones 5, 6, 8 y 9 (con 2 o más instrumentos sin información), esto es, 6 registros que representan el 0,85 % de los datos.

Para los datos restantes, esto es, registros con a lo más un instrumento sin información, se procede a la imputación de los registros perdidos (patrones 2, 3, 4 y 7, correspondiente a un total de 49 registros), en este caso se procede como a continuación:

1. Imputar los valores perdidos mediante la técnica de imputación RHD. Usando como variables de agregación: SEXO, estudiante (si/no) y variable objetivo (grupo de riesgo/control).

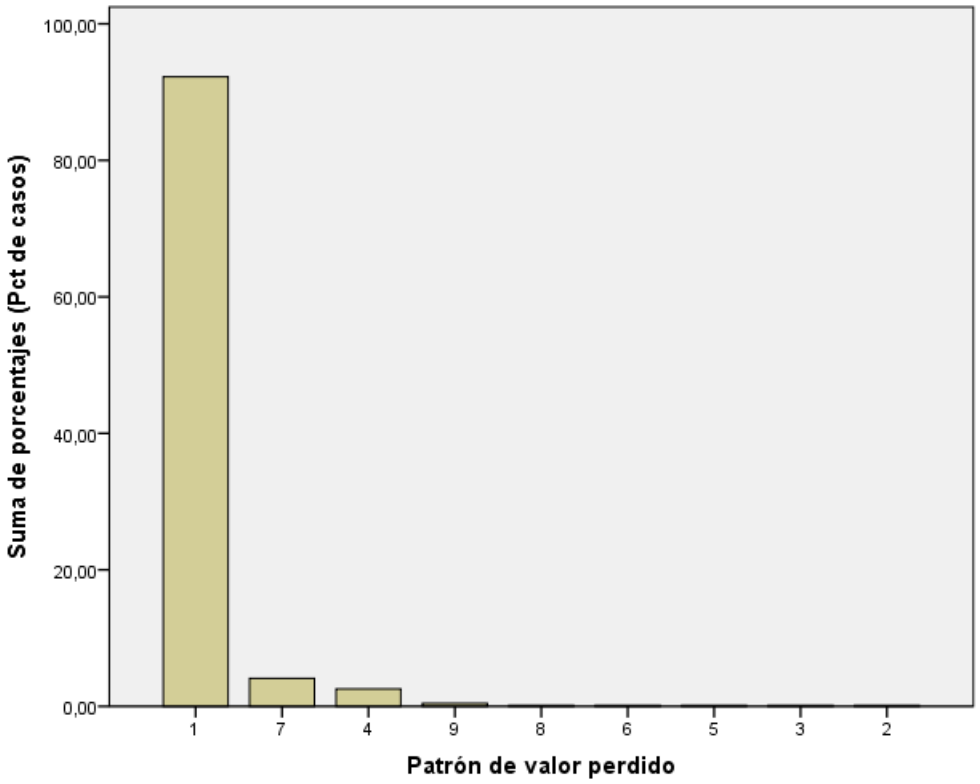

Figura 4.2: Porcentaje de registros según patron.

Patrón	**# de registros en el patrón**
1	652
2	1
3	1
4	18
7	29
5	1
6	1
8	1
9	3

Tabla 4.18: Cantidad de registros según patrón de datos.

2. La imputación se hace a nivel de registro y no de campo, es decir, por ejemplo si un paciente tiene el instrumento APGAR incompleto, se selecciona otro paciente al azar dentro del mismo grupo con información completa y se replican todas sus preguntas en este instrumento. De esa manera se mantiene la interrelación de las preguntas dentro de un instrumento para un mismo paciente, y no se crean nuevos patrones.

Así, luego de realizar todas las imputaciones antes descritas se obtiene una nueva base de datos, con 100 % de completitud, 701 registros y 224 campos.

Capítulo 5

Transformación y Reducción de variables

Luego de seleccionar los campos a utilizar y preprocesar la información se procede a la transformación de los datos, de acuerdo a la metodología de trabajo KDD.

Es en esta etapa donde se analizan los datos y se intenta obtener el mayor potencial de los datos antes del modelado mismo. Algunas de las tareas realizadas en esta etapa son: Reclasificación de categorías, normalización de campos numéricos, generación de variables dummies, creación de variables de tipo ratio, etc. Adicionalmente se realiza un segundo filtro de los campos a utilizar, eliminando aquellos campos "no relevantes" para el modelado o aquellos altamente correlacionados.

5.1. Reclasificación de campos descriptivos

En las secciones siguientes se analizan los campos descriptivos más relevantes.

5.1.1. Reclasificación campo Edad

Al analizar el campo edad se observa que este no posee una correlación clara con el riesgo suicida, sin embargo, se observa que hay edades con un mayor riesgo suicida. De la figura 5.1 se observa que entorno a los 20 y en torno a los 40, el riesgo suicida es mayor, mientras que a partir de los 60 este riesgo baja.

Luego, esta variable se reclasifica en una variable categórica con 6 clases, cada una asociada a un tramo etario. Los cortes de cada tramo son calculados de forma automática en base a los datos, usando un árbol CART que se entrena para clasificar el riesgo suicida usando solo la edad como variable, lo que da los tramos presentados en la figura 5.2. El porcentaje de pacientes del grupo de riesg suicido en cada categoría se presenta en la tabla 5.1, de donde

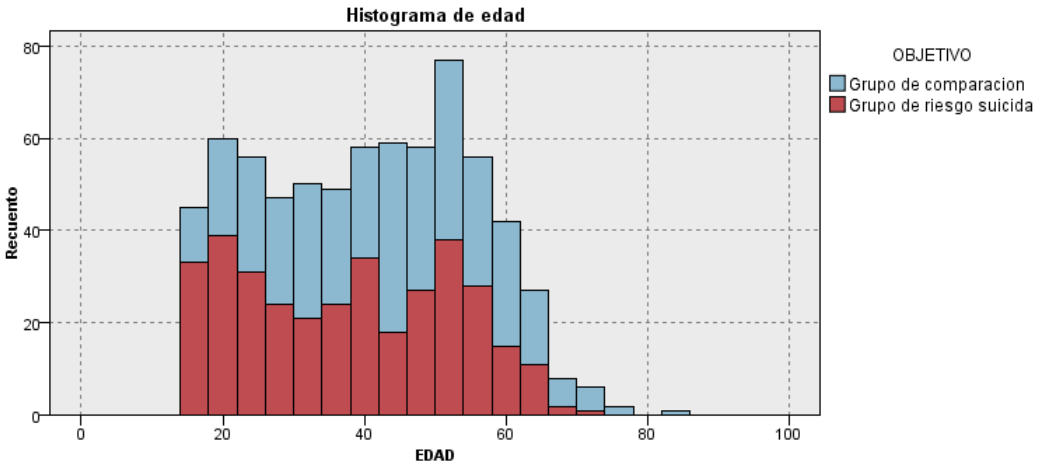

Figura 5.1: Histograma del campo EDAD, y su relación con la variable objetivo.

se observa que la clase "0-22" y "37-40" poseen un mayor porcentaje de pacientes en riesgo, mientras que el porcentaje de riesgo suicida disminuye en el tramo de "60 o más".

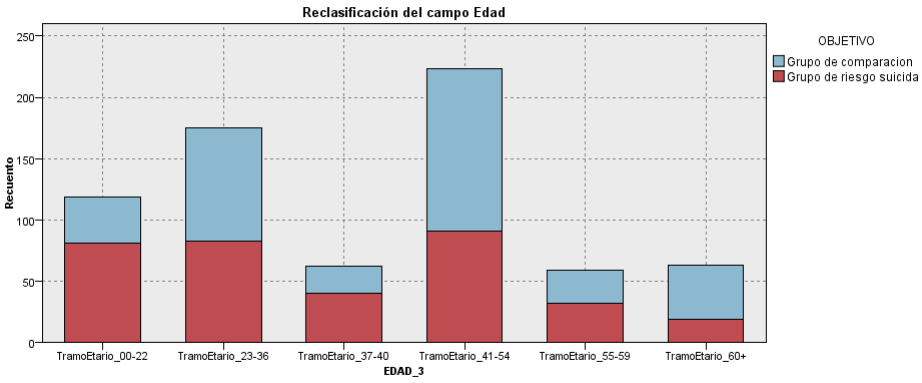

Figura 5.2: Distribución del campo EDAD reclasificado, y su relación con la variable objetivo

Tramos etarios	% de pacientes en grupo de riesgo suicida
00-22	68,1 %
23-36	47,4 %
37-40	64,5 %
41-54	40,8 %
55-59	54,2 %
60+	30,2 %

Tabla 5.1: Porcentaje de pacientes en el grupo de riesgo suicida en cada tramo etario

5.1.2. Reclasificación campo Hijos

El campo hijos contiene información del número de hijos del paciente, el cual varía en la base de datos entre 0 y 13. La reclasificación de este campo se hace de forma análoga al campo edad, utilizando un árbol de decisión para generar los cortes óptimos de las nuevas clases. Así, en este caso, la nueva variable de Hijos, posee 4 clases, definidas como sigue:

- Clase 1: 0 hijos
- Clase 2: 1 hijo
- Clase 3: 2 hijos
- Clase 3: 3 o más hijos

La distribución de pacientes según la nueva reclasificación se muestra en la figura 5.3

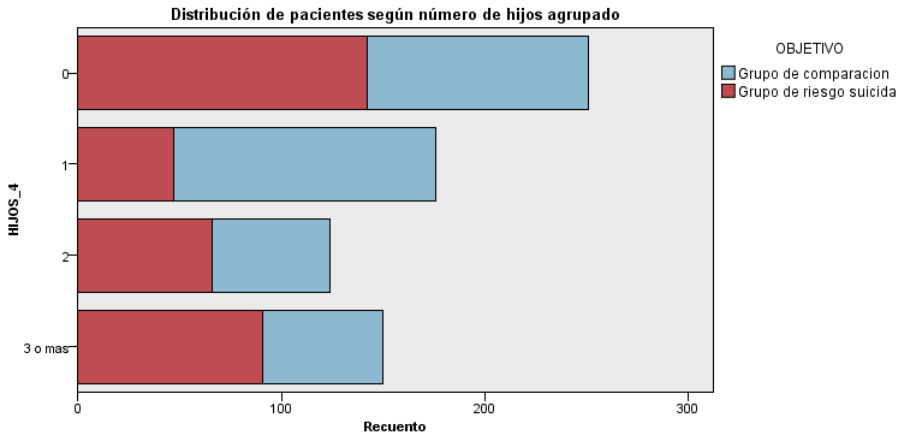

Figura 5.3: Distribución del campo reclasificado HIJOS, y su relación con la variable objetivo.

Algo a destacar de la figura anterior, es que el poseer 1 hijo es un factor protector ante el riesgo suicida, mientras que no tener hijos o tener 3 o más hijos puede ser un factor de riesgo. Obviamente todo esto dentro del contexto correspondiente, que es dentro de personas consultantes a salud mental.

5.1.3. Reclasificación campo Escolaridad y ratio Escolaridad-Edad

Al analizar el campo escolaridad se observa una gran cantidad de clases, e incluso algunas con solo un par de registros en ella, sin embargo, no es posible ver a simple vista alguna relación entre el nivel de escolaridad y el riesgo suicida como se muestra en la figura 5.4.

En este caso se prueban varias reclasificaciones sin mucho éxito, como se muestra en las figuras refrecEscol2 y refrecEscol3. Pero ninguna presenta una mejora en la discriminación del riesgo suicida.

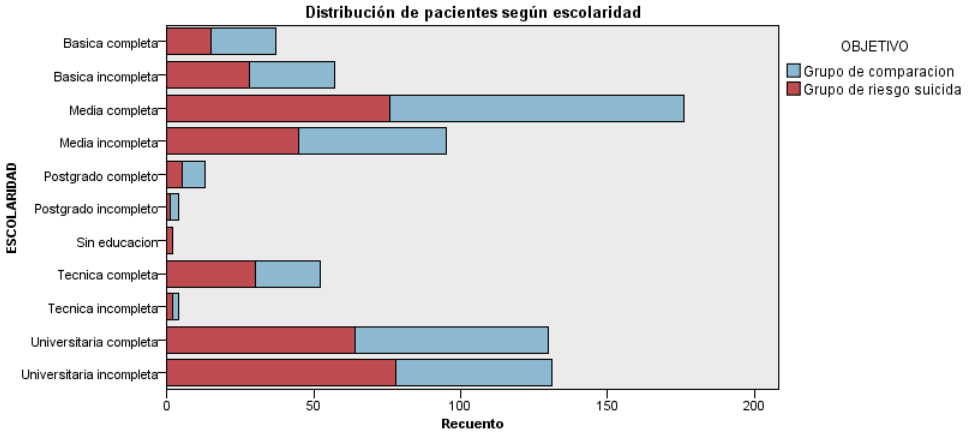

Figura 5.4: Distribución de pacientes según el campo ESCOLARIDAD y la variable objetivo.

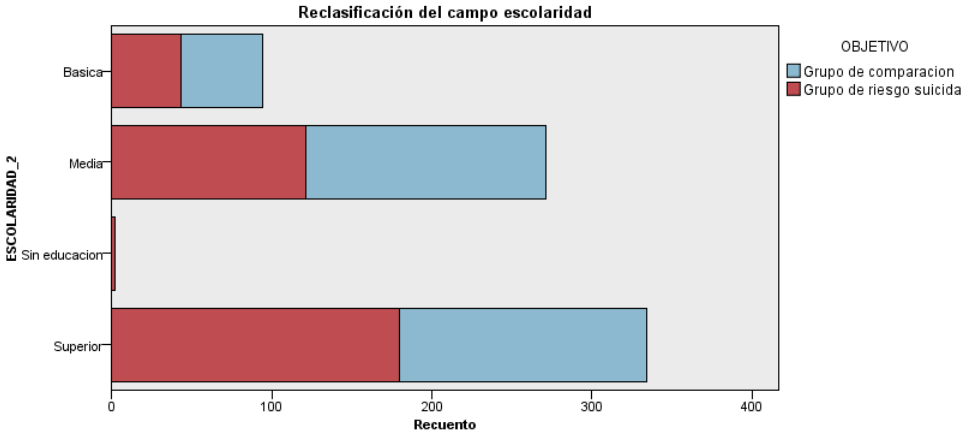

Figura 5.5: Reclasificación del campo ESCOLARIDAD usando 4 clases.

Finalmente se utiliza una variable de tipo ratio, entre la ESCOLARIDAD y la EDAD, para ello primero se utiliza una reclasificación del campo escolaridad en un nuevo campo ordinal de acuerdo a la regla en la tabla 5.2. Así, cada valor de escolaridad está representada por un valor numérico asociado a una edad correspondiente a dicho ciclo escolar.

Luego el ratio es de la forma

$$Ratio = \frac{\text{escolaridad ordinal}}{\text{mín}(\text{edad}, 30)} \tag{5.1}$$

La idea detrás de esta variable de tipo ratio es medir el efecto de la escolaridad de acuerdo a la edad, pues no es lo mismo considerar a alguien con media incompleta si tiene 16 años o si tiene 50. Adicionalmente, para eliminar el efecto de la edad en el índice para personas adultas, se usa una edad máxima de 30, esto se refiere a que para dos personas con la misma escolaridad, y con más de 30 años la medida del índice debiera ser la misma, no importa si

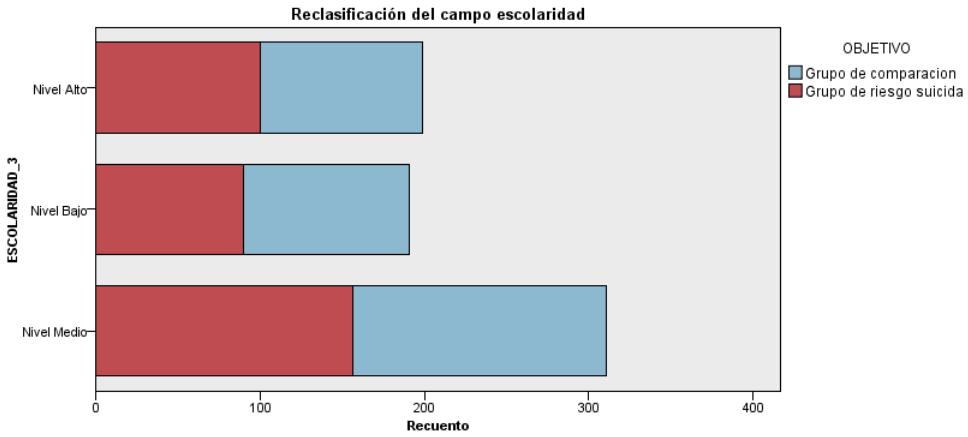

Figura 5.6: Reclasificación del campo ESCOLARIDAD utlizando 3 clases.

Escolaridad	**Escolaridad Ordinal**
Sin educación	0
Básica incompleta	11
Básica completa	15
Media incompleta	17
Media completa	19
Técnica incompleta	21
Universitaria incompleta	22
Técnica completa	24
Universitaria completa	26
Postgrado incompleto	28
Postgrado completo	30

Tabla 5.2: Escolaridad esperada al terminar cada ciclo escolar.

tiene 32 o 50 años, ya el efecto de la escolaridad no es relevante para su rango de edades.

Básicamente lo que se pretende rescatar con este índice, son aquellos pacientes que poseen una baja escolaridad respecto a lo que debieran tener, basados en su edad, o por otro lado aquellos que van incluso adelantados respecto a su edad, pues esto podría ser un factor de riesgo.

5.1.4. Reclasificación campo Estado Civil

Al analizar el campo ECIVIL (estado civil) lo primero que destaca, es que dado la conjunción de diferentes bases de datos para formar la base de datos final, las diferentes clases contenidas en este campo no fueron estandarizados. Como se aprecia en la figura 5.7, existen 3 clases no excluyentes entre sí, estas son:

- Separado / divorciado
- Viudo
- Viudo /separado

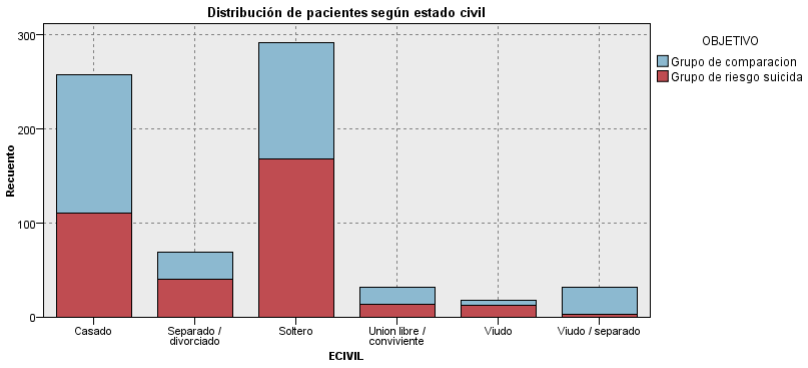

Figura 5.7: Distribución de pacientes según el estado civil y su relación con la variable objetivo.

Esto se puede deber a que una de las bases de datos originales contenía la clase viudo / separado de forma agrupada, mientras que en otro centro de salud se utilizaba separada. De acuerdo a los especialistas en salud mental lo correcto sería esto último, pues un paciente viudo, en teoría, es diferente a alguien separado y debieran tratarse de manera distinta. Sin embargo, esta separación no se puede hacer a posteriori, al menos no sin agregar ruido en los datos.

Es por esto, que se reclasifica la variable Estado civil en 3 clases, si bien, no era la primera opción, al menos esta nueva clasificación genera clases mutuamente excluyentes, además combina 3 clases similares, estas son, casado, unión libre, y conviviente. La nueva clasificación se muestra en la figura 5.8.

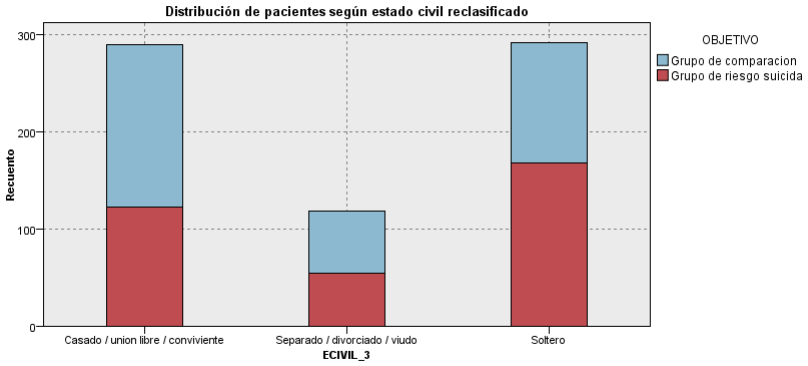

Figura 5.8: Reclasificación del campo ECIVIL en tres clases.

5.1.5. Reclasificación campo Ocupación

La figura 5.9 muestra la distribución de valores del campo ocupación, donde destacan algunas clases con pocos registros. Dado eso, se procede a reclasificar el campo en 5 clases según se muestra en la figura 5.10.

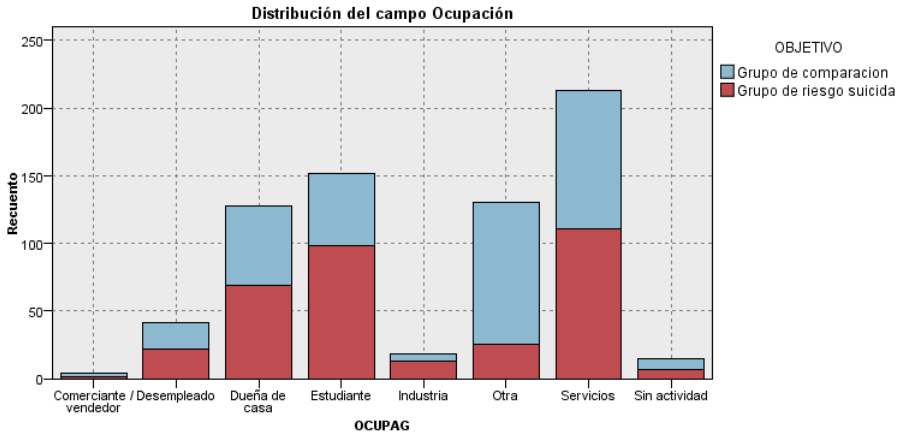

Figura 5.9: Distribución de pacientes según el campo OCUPAG.

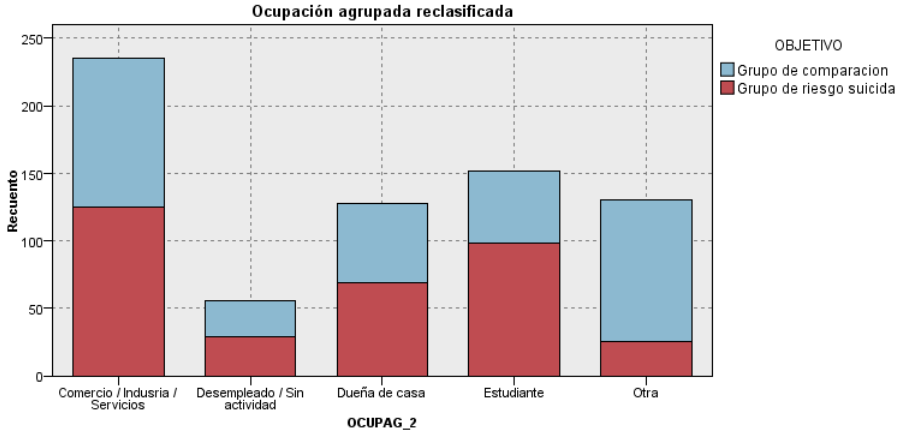

Figura 5.10: Distribución de pacientes utilizando la reclasificación del campo OCUPAG, esta reclasificación genera 5 clases.

5.1.6. Reclasificación campo Diagnóstico

La gran cantidad de clases distintas del campo diagnóstico, ver figura 5.11, hace necesaria su reclasificación. En la figura 5.12 se muestra una reclasificación de acuerdo a un criterio clínico, agrupando psicopatologías relacionadas según el tipo de trastorno. Sin embargo, aún existen varias clases con tan solo un par de registros por lo que se procede a generar una

nueva reclasificación, esta vez agrupando diagnósticos del tipo trastorno o episodio depresivo ya sean moderados o severos. La información asociada a esta nueva clasificación se muestra en la figura 5.13.

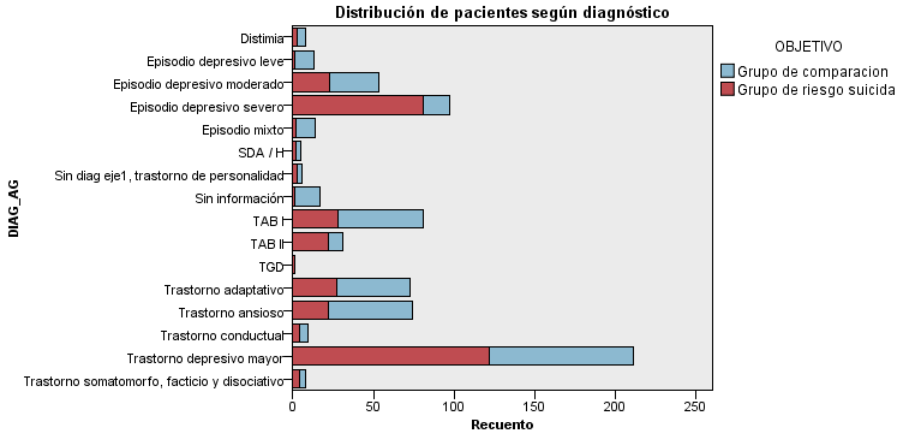

Figura 5.11: Distribución del campo DIAG_AG.

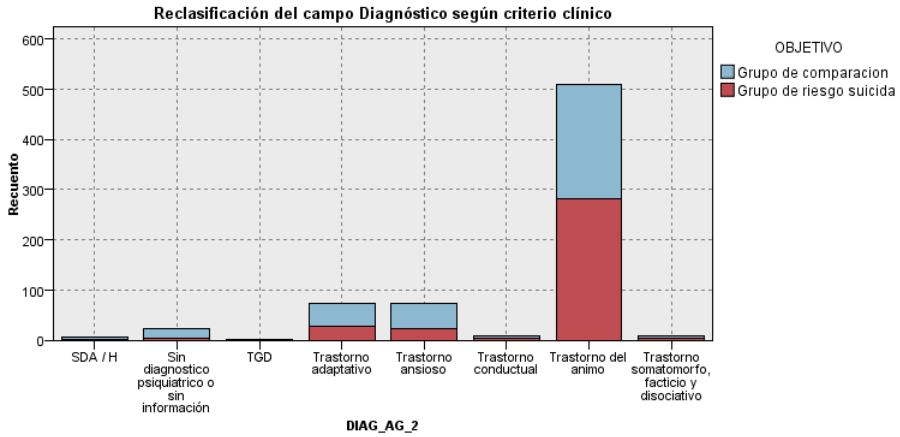

Figura 5.12: Reclasificación del campo DIAG_AG según criterio de agrupación clínico.

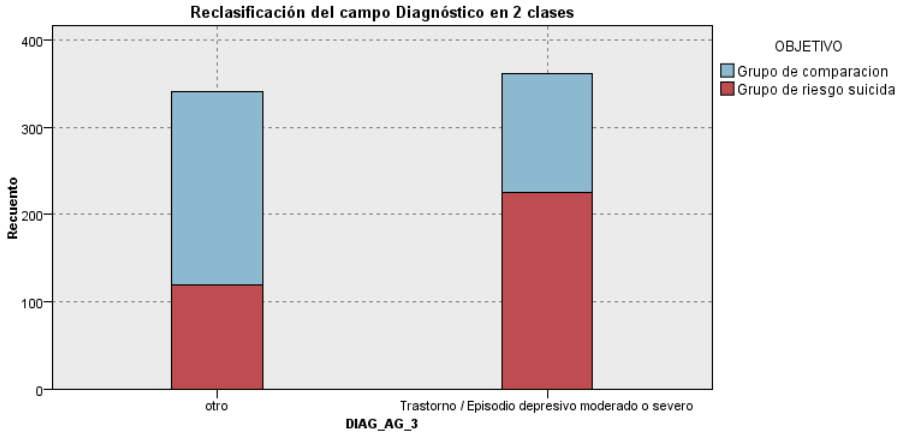

Figura 5.13: Reclasificación del campo DIAG_AG en dos clases, considerando el diagnóstico con mayor cantidad de pacientes.

5.2. Selección de campos descriptivos más relevantes

Adicionalmente a las transformaciones antes descritas, se generan más clasificaciones para luego utilizar la variable que mejor ayude al modelo.

Para seleccionar las variables / transformaciones que más aporten al modelado se un test chi-cuadrado de Pearson con la herramienta SPSS Modeler. El test calcula la importancia de cada variable con respecto a la variable objetivo, donde se analiza la independencia de las dos variables entre sí, mediante la presentación de los datos en tablas de contingencia.

La elección de la variable se realiza en base al $p - valor$ asociado a cada atributo, donde para cada concepto o grupo de variables se selecciona aquella que tenga el $p-valor$ más pequeño.

El resultado de la selección de características con los atributos relevantes se muestra en la tabla 5.3. Por su parte las variables no utilizadas son eliminadas de la base de datos.

Campo	**Descripción**	**Importancia (1 - p-valor)**
OCUPAG_2	Ocupación agrupada en 5 clases	1
DIAG_AG_3	Diagnóstico agrupado en 2 clases	1
HIJOS_4	Número de hijos agrupados en 4 clases	1
EDAD_3	Edad agrupada en 6 clases	1
EDAD	Edad como variable continua	0,99999
RATIO_ESC_EDAD_1	Ratio escolaridad vs edad	0,99993
ECIVIL_3	Estado civil agrupado en 3 clases	0,99903

Tabla 5.3: Campos descriptivos seleccionados como relevantes

Notar que tanto el campo EDAD, como la reclasificación de éste en categorías, EDAD_3,

permanecen en la base de datos a pesar de hacer referencia al mismo campo. Esto se hizo ya que el campo original es continuo mientras que la reclasificación es categórica, luego el efecto de cada variable puede ser distinto respecto a la técnica de modelación que se utilice.

5.3. Transformación de datos de instrumentos clínicos

Dadas las características de los datos asociados a los instrumentos es que estos pueden ser usados de diferente forma, como campos continuos, ordinales e incluso nominales. Luego, dadas las diferentes herramientas a utilizar se opta por usarlos como datos continuos, pues si bien cada instrumento tiene su escala (mínimo y máximo) particular, en cada uno de ellos y en cada pregunta existe una correlación con el grupo o factores de riesgo, siendo por ejemplo para algunos instrumentos un mayor valor asociado a un mayor riesgo suicida.

Así, la primera transformación a realizar consiste en normalizar las respuestas en una escala común. Esto es recomendable para poder aplicar algunos métodos sobre campos numéricos, como el k-nn u otros. La normalización utilizada es de la forma:

$$Nueva\ variable\ = \frac{variable - \text{mín}(variable)}{\text{máx}(variable) - \text{mín}(variable)} \tag{5.2}$$

Otras transformaciones realizadas dependen del criterio experto, dividiendo el espectro de respuestas de algunos instrumentos en 2 o 3 clases, por ejemplo el RFL, se transforma en variables binarias donde 1, 2 y 3 representan lo "no importante", y las respuestas 4, 5 y 6 se asocian a lo "muy importante".

Además de las anteriores se realizaron otras transformaciones de los datos de instrumentos, sin embargo, no se obtuvieron mejoras significativas por lo que finalmente no se utilizaron. Entre algunas de las transformaciones utilizadas se encuentran:

- **Transformación binaria** : Con los campos ya normalizados se definieron variables binarias $0-1$ según la siguiente regla:

$$T_{bin}(x) = \begin{cases} 0 \quad si \quad x < 0{,}5 \\ \\ 1 \quad si \quad x \geq 5 \end{cases} \tag{5.3}$$

- **Transformación binaria con corte óptimo** : Con la ayuda de un árbol de decisión es posible obtener punto de corte que maximicela ganancia de información alhacer un split con una variable dada, de esta manera se calcula tal división para todas las variables asociadas a los instrumentos y se define una variable binaria usando este punto de corte en vez del usual 0,5.

$$T_{binOpt}(x) = \begin{cases} 0 \quad si \quad x < c_{opt} \\ \\ 1 \quad si \quad x \geq c_{opt} \end{cases} \tag{5.4}$$

- **Transformación sigmoide** : Con el punto de corte calculado según la metodología anterior se transforma el rango $0-1$ mediante una función sigmoide tal que $f : [0,1] \rightarrow [0,1]$ y $f(x_{corteOpt}) = 0{,}5$. En la figura 5.14 se muestra un ejemplo de la transformación sigmoide usada con la variable OQ8. La idea de utilizar esta transformación es la de suavizar la transformación binaria con punto de corte óptimo

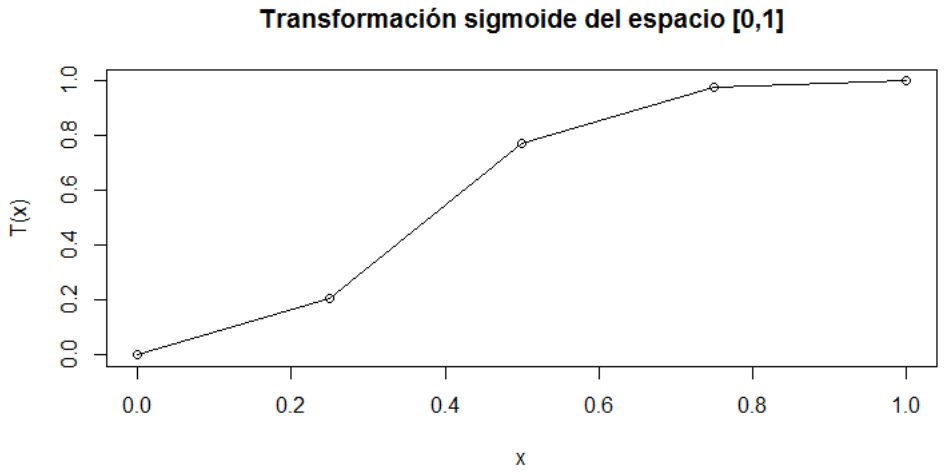

Figura 5.14: Transformación sigmoide aplicada a la variable OQ8, con punto de corte óptimo 0,375.

- **Transformación logit** : En análisis de percepción de clientes/pacientes es común utilizar transformaciones de tipo logit, y que acentúa los extremos y homogeneiza valores centrales. En la figura 5.15 se muestra la aplicación de esta transformación a la misma pregunta OQ8.

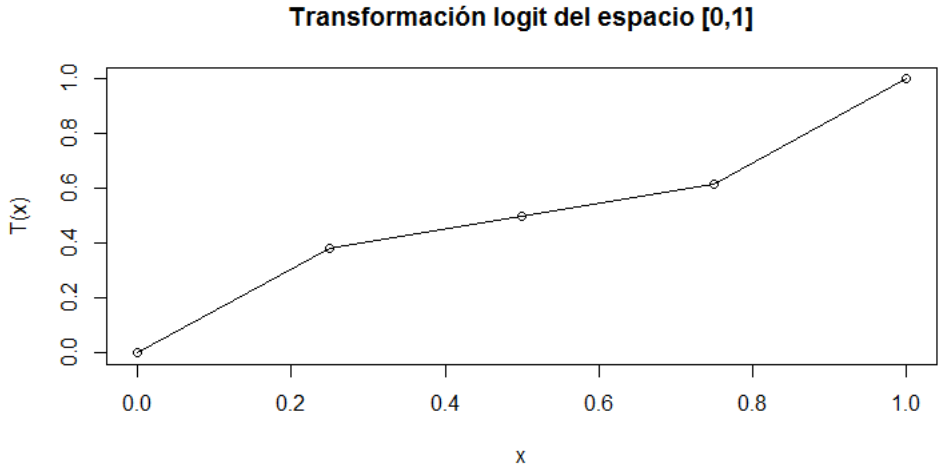

Figura 5.15: Transformación logit aplicada a la variable OQ8.

5.4. Selección de variables de instrumentos

Recordar que la cantidad de campos asociados a los instrumentos clínicos dentro de la base de datos asciende a 210 considerando los 5 instrumentos utilizados. Dado esto, es necesario una reducción de atributos antes de empezar las labores de modelado, al igual como se hizo con las variables descriptivas.

El primer paso de la reducción se realiza mediante el algoritmo de selección de atributos, esta vez, en vez de hacer el test chi-cuadrado con campos categóricos se realiza con las variables de los instrumentos normalizadas, las que se trabajan como variables numéricas continuas.

Con esta herramienta se seleccionan sólo aquellos campos definidos como "importante" en la relación con la variable de riesgo suicida, es decir, los campos con un $p - valor$ menor a 0,05. Con esta elección los campos asociados a los instrumentos se reducen de 210 a 129 atributos.

5.5. Reducción de atributos según correlación

Con el primer filtro sobre las variables ya se ha reducido el tamaño de la base analítica a utilizar en el modelado. El siguiente paso para reducir un poco más la cantidad de variables viene por el análisis de correlaciones, eliminando aquellas variables altamente correlacionadas entre sí.

Antes del análisis mismo, y para poder incluir todas las variables en el análisis (descriptivas y de instrumentos), se recodifican las variables descriptivas con tipo de dato categórico en nuevas variables dummies, esto es, para cada atribut con n categorías se generan $n-1$ nuevas variables binarias, donde un 1 representa que el paciente está asociado a dicha clase o 0 si no. Así, al recodificar los campos categóricos se cuenta con 139 campos.

Para el análisis de correlaciones se hace uso del lenguaje estadístico `R`, el cual posee una amplia gama de funciones y paquetes para el análisis de datos. En la figura 5.16 se observa la matriz de correlaciones para los 139 campos, donde se pueden observar grupos de variables correlacionadas entre sí, por ejemplo, el extremo inferior izquerdo de la matriz corresponde a las preguntas del RFL, el cual se ve presenta una mayor correlacion entre sí que con el resto de los campos. En la figura 5.17 se presenta la matriz de correlaciones considerando sólo las variables del RFL.

Para analizar de mejor manera los campos que pudieran estar correlacionados se crea un función en `R` que entrega todas las parejas de variables que tienen una correlación por sobre un umbral (considerando valor absoluto). Para efectos de distinguir aquellas variables correlacionadas se define un umbral de 0,7, esto es, campos cn correlaciones por sobre 0,7 o

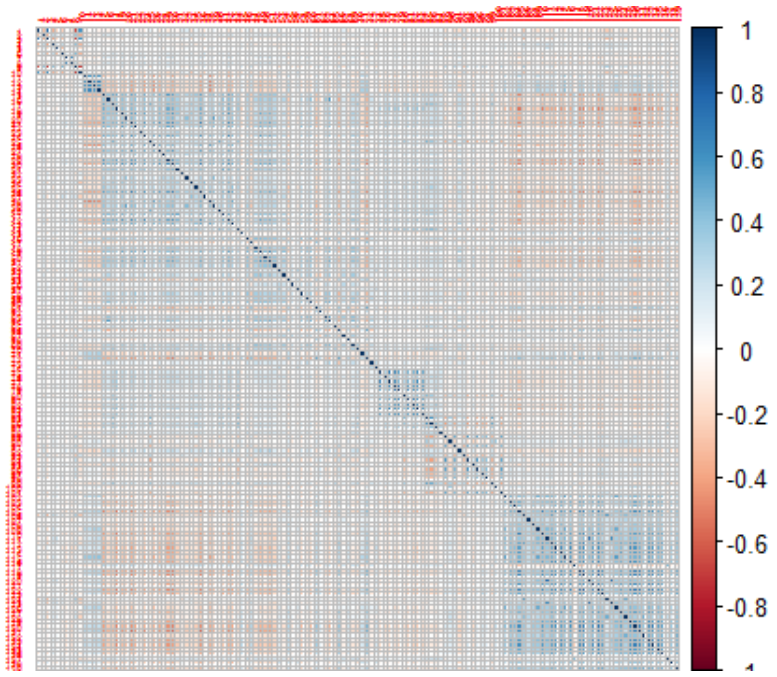

Figura 5.16: Matriz de correlaciones para los 139 atributos.

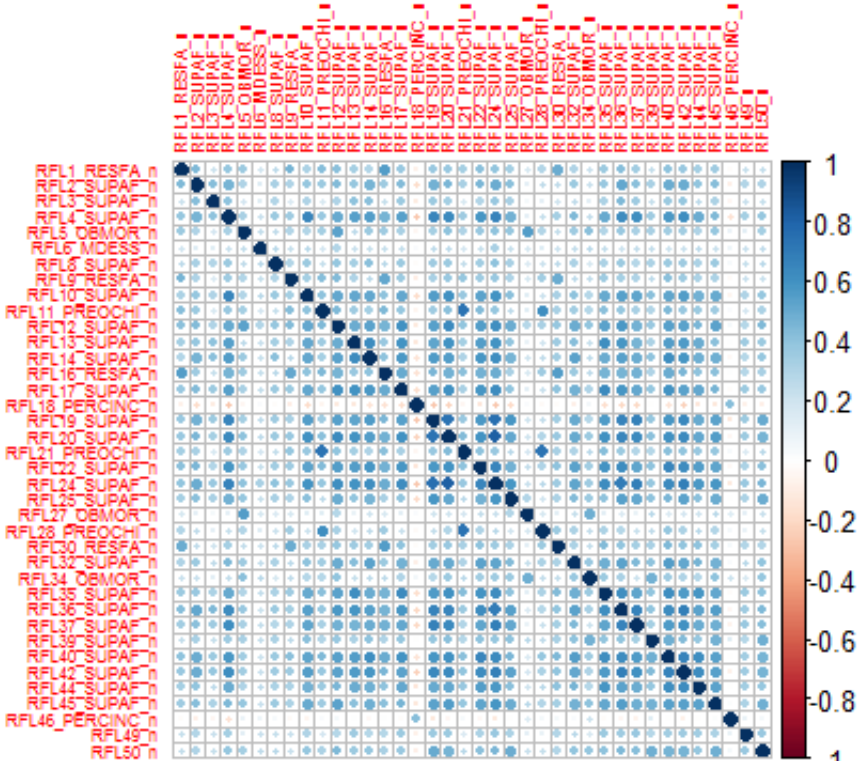

Figura 5.17: Matriz de correlaciones para las variables asociadas al RFL.

por debajo de $-0{,}7$ se consideraran altamente correlacionados. Luego, los campos detectados como correlacionados despues de un primer análisis se muestran en el código 5.1:

Luego son 11 las variables con correlaciones altas, las cuales se dividen en 4 grupos:

```
> correl <- VarCorr(atributos,0.7)
[1] "1 -. La correlacion entre TramoEtario_00.22 y OCUPACION_Estudiante es
    alta ( 0.748711186918159 )"
[1] "2 -. La correlacion entre STAXIPRE4_EDO_n y STAXIPRE5_EDO_n es alta (
    0.735163889484509 )"
[1] "3 -. La correlacion entre RFL11_PREOCHI_n y RFL21_PREOCHI_n es alta (
    0.737586509717814 )"
[1] "4 -. La correlacion entre RFL19_SUPAF_n y RFL20_SUPAF_n es alta (
   0.730097361145789 )"
[1] "5 -. La correlacion entre RFL19_SUPAF_n y RFL24_SUPAF_n es alta (
   0.75870658254802 )"
[1] "6 -. La correlacion entre RFL20_SUPAF_n y RFL24_SUPAF_n es alta (
   0.8170735538473 )"
[1] "7 -. La correlacion entre RFL21_PREOCHI_n y RFL28_PREOCHI_n es alta (
    0.73198445900364 )"
```

Resultado de R 5.1: Variables con correlaciones mayores a 0.7

1. TramoEtario_00.22 y OCUPACION_Estudiante
2. STAXIPRE4_EDO_n y STAXIPRE5_EDO_n
3. RFL11_PREOCHI_n, RFL21_PREOCHI_n y RFL28_PREOCHI_n
4. RFL19_SUPAF_n, RFL20_SUPAF_n y RFL24_SUPAF_n

Para el caso 1 la correlación es evidente, ya que los pacientes que son estudiantes se concentran en ese rango etario, luego se procede a eliminar el campo OCUPACION_Estudiante. Por su parte, para los casos 2, 3 y 4, la correlación también es razonable ya que todas (entre cada grupo) corresponden a preguntas del mismo instrumento e incluso a la misma subescala o dimensión dentro de este. Sin embargo, para mantener una estabilidad en las respuestas de los pacientes y disminuir el ruido, las preguntas correlacionadas no son eliminadas, si no que se promedian para obtener un nuevo campo que engloba tales dimensiones. Una vez creadas las nuevas variables se eliminan las originales, quedando sólo los promedios:

- STAXIPRE4_5_EDO_n
- RFL11_21_28_PREOCHI_n
- RFL19_20_24_SUPAF_n

A continuación se vuelve a calcular la correlación, para verificar si las nuevas variables introducidas presentan nuevas correlaciones. Los resultados son presentados en el código 5.2.

Y efectivamente una de las variables creadas, la RFL19_20_24_SUPAF_n genera 4 nuevas correlaciones altas, con las preguntas RFL4, RFL36, RFL37 y RFL42 todas de la subescala "SUPAF", por lo que ahora sí en este caso, y dado que la variable RFL19_20_24_SUPAF_n ya es un promedio de 3 preguntas, se opta por eliminar las 4 nuevas variables correlacionadas.

Así, tras esta reducción de campos según correlación es posible disminuir de 139 a 129 atributos. Los que se enumeran en el cuadro de resultados 5.3:

```
> correl2 <- VarCorr(atributos2 ,0.7)
[1] "1 -. La correlacion entre  RFL4_SUPAF_n  y  RFL19_20_24_SUPAF_n  es alta
    ( 0.709256514284807 )"
[1] "2 -. La correlacion entre  RFL36_SUPAF_n  y  RFL19_20_24_SUPAF_n  es alta
    ( 0.731784057852067 )"
[1] "3 -. La correlacion entre  RFL37_SUPAF_n  y  RFL19_20_24_SUPAF_n  es alta
    ( 0.723138724532199 )"
[1] "4 -. La correlacion entre  RFL42_SUPAF_n  y  RFL19_20_24_SUPAF_n  es alta
    ( 0.710330806997419 )"
```

Resultado de R 5.2: Variables correlacionadas tras la primera iteración

```
> colnames(atributos3)
  [1] "TramoEtario_41.54"       "ECIVIL_Cas_UniLib_Conv"
  [3] "TIENE_1_HIJO"            "TIENE_3_o_mas_HIJOS"
  [5] "T_E_DEPRE_MOD_o_SEV"     "TRASTORNO_ANIMO"
  [7] "EDAD_n"                  "RATIO_ESC_EDAD_1_n"
  [9] "APGAR_PRE_ADAP_n"        "APGAR_PRE_PAR_n"
 [11] "APAR_PRE_AFECTO_n"       "APGAR_PRE_RESOL_n"
 [13] "OQPRE3_SD_n"             "OQPRE5_SD_n"
 [15] "OQPRE6_SD_n"             "OQPRE8_SD_n"
 [17] "OQPRE9_SD_n"             "OQPRE12_RS_n"
 [19] "OQPRE13_SD_n"            "OQPRE15_SD_n"
 [21] "OQPRE16_RI_n"            "OQPRE18_RI_n"
 [23] "OQPRE19_RI_n"            "OQPRE20_RI_n"
 [25] "OQPRE21_RS_n"            "OQPRE22_SD_n"
 [27] "OQPRE23_SD_n"            "OQPRE24_SD_n"
 [29] "OQPRE25_SD_n"            "OQPRE26_RI_n"
 [31] "OQPRE28_RS_n"            "OQPRE29_SD_n"
 [33] "OQPRE30_RI_n"            "OQPRE31_SD_n"
 [35] "OQPRE35_SD_n"            "OQPRE37_RI_n"
 [37] "OQPRE38_RS_n"            "OQPRE39_RS_n"
 [39] "OQPRE40_SD_n"            "OQPRE42_SD_n"
 [41] "OQPRE43_RI_n"            "OQPRE44_RS_n"
 [43] "DEQPRE3_n"               "DEQPRE7_n"
 [45] "DEQPRE8_n"               "DEQPRE10_n"
 [47] "DEQPRE11_n"              "DEQPRE13_n"
 [49] "DEQPRE16_n"              "DEQPRE17_n"
 [51] "DEQPRE18_n"              "DEQPRE19_n"
 [53] "DEQPRE21_n"              "DEQPRE22_n"
 [55] "DEQPRE24_n"              "DEQPRE25_n"
 [57] "DEQPRE28_n"              "DEQPRE30_n"
 [59] "DEQPRE33_n"              "DEQPRE38_n"
 [61] "DEQPRE41_n"              "DEQPRE43_n"
 [63] "DEQPRE44_n"              "DEQPRE48_n"
 [65] "DEQPRE50_n"              "DEQPRE54_n"
 [67] "DEQPRE55_n"              "DEQPRE58_n"
 [69] "DEQPRE61_n"              "DEQPRE62_n"
 [71] "DEQPRE64_n"              "DEQPRE66_n"
 [73] "STAXIPRE1_EDO_n"         "STAXIPRE2_EDO_n"
 [75] "STAXIPRE3_EDO_n"         "STAXIPRE6_EDO_n"
 [77] "STAXIPRE7_EDO_n"         "STAXIPRE8_EDO_n"
 [79] "STAXIPRE9_EDO_n"         "STAXIPRE10_EDO_n"
 [81] "STAXIPRE12_RGO_n"        "STAXIPRE16_RGO_n"
```

```
 [83] "STAXIPRE19_RGO_n"        "STAXIPRE20_RGO_n"
 [85] "STAXIPRE21_CTR_n"        "STAXIPRE22_EXP_n"
 [87] "STAXIPRE24_CTR_n"        "STAXIPRE26_GDO_n"
 [89] "STAXIPRE27_EXP_n"        "STAXIPRE28_CTR_n"
 [91] "STAXIPRE30_GDO_n"        "STAXIPRE31_CTR_n"
 [93] "STAXIPRE38_CTR_n"        "STAXIPRE40_CTR_n"
 [95] "STAXIPRE42_EXP_n"        "STAXIPRE43_EXP_n"
 [97] "STAXIPRE44_CTR_n"        "RFL1_RESFA_n"
 [99] "RFL2_SUPAF_n"            "RFL3_SUPAF_n"
[101] "RFL5_OBMOR_n"            "RFL6_MDESS_n"
[103] "RFL8_SUPAF_n"            "RFL9_RESFA_n"
[105] "RFL10_SUPAF_n"           "RFL12_SUPAF_n"
[107] "RFL13_SUPAF_n"           "RFL14_SUPAF_n"
[109] "RFL16_RESFA_n"           "RFL17_SUPAF_n"
[111] "RFL18_PERCINC_n"         "RFL22_SUPAF_n"
[113] "RFL25_SUPAF_n"           "RFL27_OBMOR_n"
[115] "RFL30_RESFA_n"           "RFL32_SUPAF_n"
[117] "RFL34_OBMOR_n"           "RFL35_SUPAF_n"
[119] "RFL39_SUPAF_n"           "RFL40_SUPAF_n"
[121] "RFL44_SUPAF_n"           "RFL45_SUPAF_n"
[123] "RFL46_PERCINC_n"         "RFL49_n"
[125] "RFL50_n"                 "TramoEtario_00_22"
[127] "RFL11_21_28_PREOCHI_n"   "STAXIPRE4_5_EDO_n"
[129] "RFL19_20_24_SUPAF_n"
```

Resultado de R 5.3: Atributos seleccionados

Capítulo 6

Modelamiento

A lo largo del presente capítulo se muestran las técnicas y/o herramientas para la labor de modelamiento, tomando como base lo desarrollado en los capítulos previos.

Se muestra lo realizado para 3 modelos distintos un árbol de decisión, un modelo KNN y un SVM, adicionalmente se expondrán variantes del árbol de desición, dentro del contexto de *ensemble models* o modelos combinados. Los modelos combinados aquí evaluados corresponden a los algoritmos Random Forest y AdaBoost, los que se detallan en la sección 6.4. El ajuste de los parámetros y resultados preliminares se muestran en las secciones siguientes, mientras que una completa evaluación y comparación de los modelos generados se presentan en el capítulo 7.

Para la generación de los modelos y la posterior evaluación de estos, se utiliza el lenguaje estadístico `R`, el cual cuenta con una completa librería de técnicas y herramientas para el análisis de los datos.

6.1. Modelo CART

El primer modelo a presentar es un árbol de decisión de tipo CART, el cual además de servir para el problema mismo de clasificación, sirve para ir conociendo un poco más los datos y las relaciones que se presentan, y que no son evidentes a simple vista. Además, el árbol de decisión entrega información valiosa acerca de las variables más relevantes.

Para generar y ajustar el modelo CART se utiliza la librería `rpart` de `R`.

6.1.1. Ajuste del modelo

El parámetro a ajustar para este modelo tiene que ver con el valor c_p, asociado al parámetro de complejidad. Para ajustar el mejor valor de c_p en función del poder de generalización del árbol, se procede a generar un árbol máximal, esto es, con $c_p = 0$ y con la máxima profundidad posible. Una vez generado tal árbol se evalúa la precisión del modelo resultante para distintos valores de c_p y se escoge aquel que minimiza el error en el conjunto de validación. La estimación del error se realiza usando validación cruzada con $n = 10$.

El esquema del árbol resultante se muestra en el figura 6.1, donde se aprecia una gran cantidad de divisiones, volviendo la interpretabilidad del árbol un tanto difícil. Esto último remarca la necesidad de podar el árbol generado (al aumentar el valor de c_p).

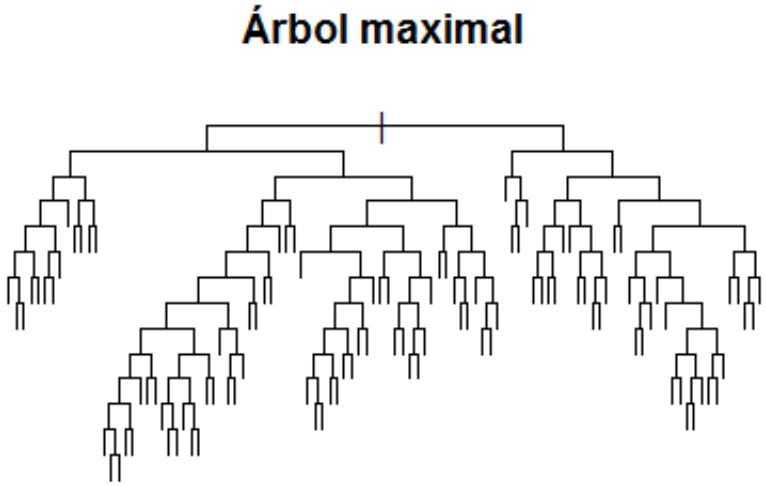

Figura 6.1: Árbol maximal con la mayor profundidad posible.

Del modelo generado se puede obtener el error de clasificación para varios valores de c_p, los que se muestran en la figura 6.2 y en la tabla 6.1. De aquí, el mejor valor en la elección de c_p es 0,0216763, con el que el modelo tiene un error relativo en el conjunto de validación igual a 0,55202 [1]

Una vez ajustado el parámetro de complejidad se vuelve a entrenar el árbol, obteniéndose un esquema como el de la figura 6.3. En él se aprecia que tiene tan sólo 3 niveles y utiliza 3 variables, OQPRE8_SD_n, TIENE_1_HIJO y T_E_DEPRE_MOD_o_SEV.

[1] El error relativo se calcula tomando como base al error del nodo raíz que es igual a 0,49358, luego el error real es 0,272466

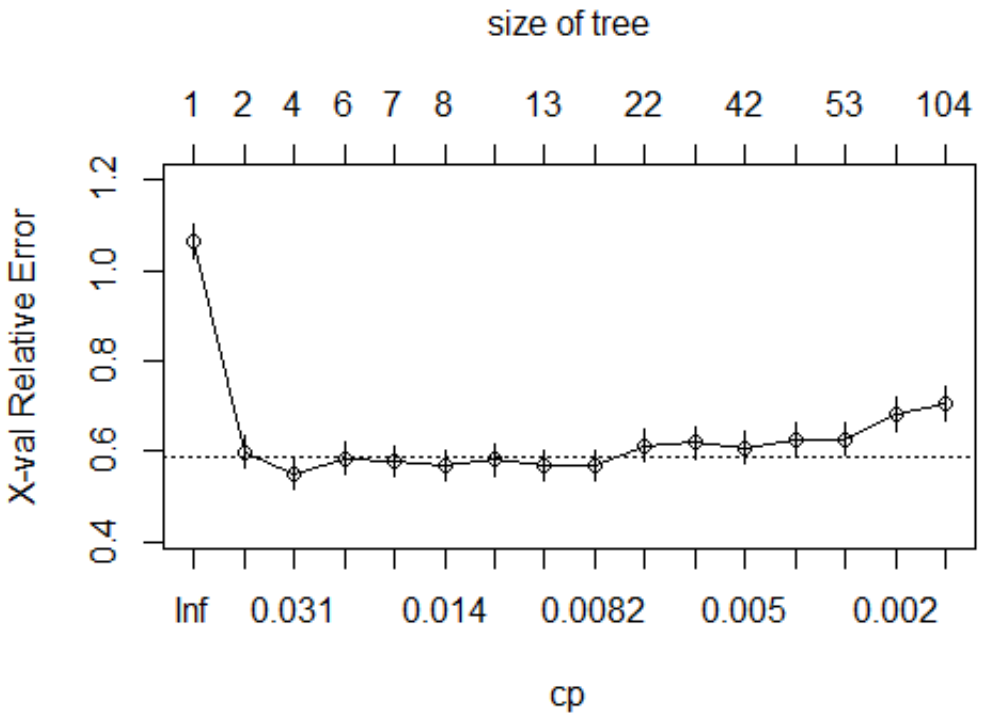

Figura 6.2: Ajuste del parámetro de complejidad c_p.

CP	Número de splits	Error de entrenamiento	Error de validación	Desviación estándar del error
0.4190751	0	1	1.06358	0.038213
0.0447977	1	0.580925	0.59827	0.034907
0.0216763	3	0.491329	0.55202	0.03407
0.0202312	5	0.447977	0.58382	0.034657
0.017341	6	0.427746	0.57803	0.034554
0.0115607	7	0.410405	0.56936	0.034396
0.0096339	9	0.387283	0.58092	0.034606
0.0086705	12	0.358382	0.56936	0.034396
0.0077071	17	0.315029	0.56936	0.034396
0.0072254	21	0.283237	0.61272	0.035147
0.0057803	23	0.268786	0.6185	0.03524
0.0043353	41	0.16185	0.60694	0.035052
0.0036127	47	0.135838	0.62717	0.035377
0.0028902	52	0.115607	0.62717	0.035377
0.0014451	81	0.031792	0.68208	0.036162
0	103	0	0.7052	0.036452

Tabla 6.1: Error relativo de entrenamiento y validación para diferentes valores de c_p. El error es relativo al error del nodo raiz, el cual es 0,49358, asociado a clasificar todas las observaciones en una misma clase. Así, al usar un $c_p = 0{,}0216763$ el error relativo es 0,55202, y el error de clasificación viene dado por $0{,}55202 * 0{,}49358 = 0{,}272466$

Adicionalmente a las variables seleccionadas en la construcción del modelo, es posible extraer información acerca de las variables más relevantes mediante un ranking generado por el mismo modelo, esto se muestra en la figura 6.4 donde se han seleccionado las 10 variables más importantes según este modelo. La variable más relevante corresponde a la pregunta OQPRE8_SD_n del instrumento *Outcome Questionaire*, mientras que la segunda m+as importante corresponde al promedio de las variables 19, 20 y 24 del instrumento RFL.

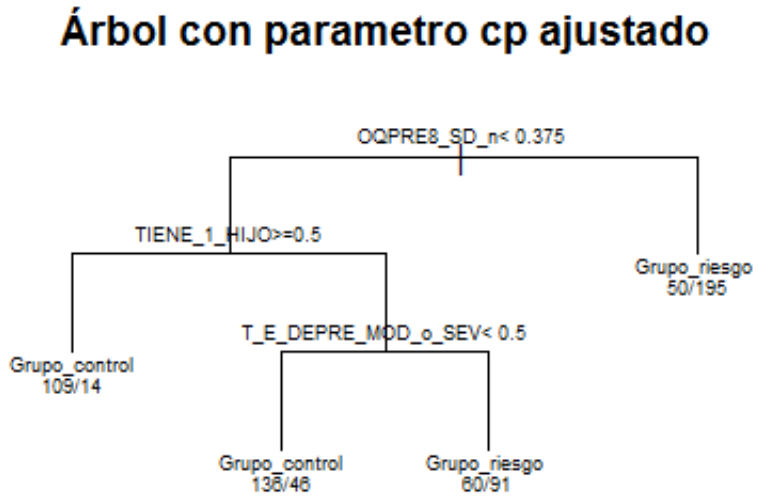

Figura 6.3: Esquema del árbol de decisión con el parámetro c_p ajustado.

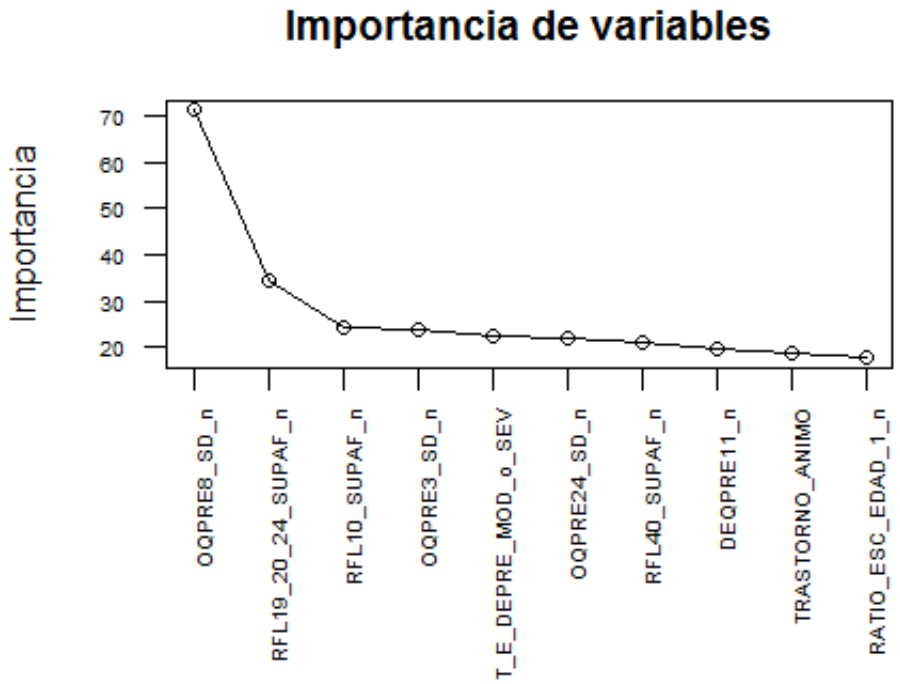

Figura 6.4: 10 variables más importantes para el modelo CART generado

Para analizar un poco más en detalle el modelo obtenido se genera una instancia de entre-

```
Confusion Matrix and Statistics

                Reference
Prediction       Grupo_control Grupo_riesgo
  Grupo_control             25            8
  Grupo_riesgo              10           27

               Accuracy : 0.7429
                 95% CI : (0.6244, 0.8399)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 2.925e-05

                  Kappa : 0.4857
 Mcnemar's Test P-Value : 0.8137

            Sensitivity : 0.7714
            Specificity : 0.7143
         Pos Pred Value : 0.7297
         Neg Pred Value : 0.7576
             Prevalence : 0.5000
         Detection Rate : 0.3857
   Detection Prevalence : 0.5286
      Balanced Accuracy : 0.7429

       'Positive' Class : Grupo_riesgo
```

Resultado de R 6.1: Matriz de confusión y estadísticas asociadas al modelo CART

namiento y validación, dejando el 90 % de los datos para entrenar y el 10 % restante para la validación. Los resultados del modelo en esta instancia se presentan en el resultado 6.1. Del aquí se observa que el modelo tiene una precisión de 0,7429, y donde la sensibilidad es levemente superior a la especificidad. Lo anterior se puede ver también en el espacio ROC, como se muestra en la figura 6.5.

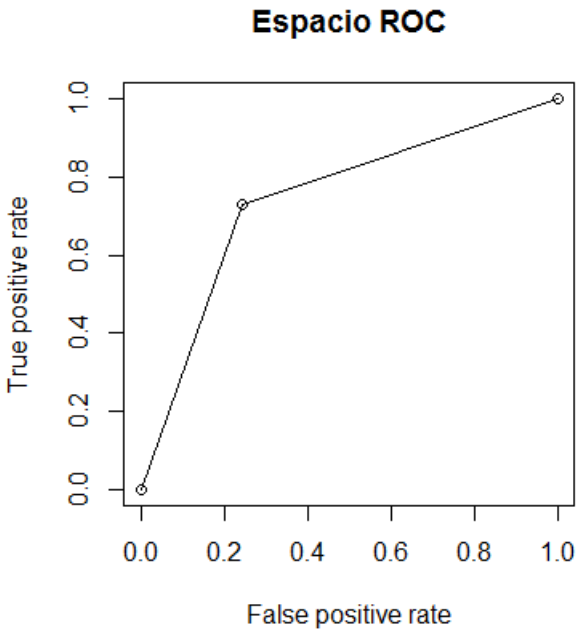

Figura 6.5: Precisión de modelo CART en el espacio ROC para una instancia de entrenamiento-validación fija.

6.2. Modelo SVM

El siguiente modelo a ajustar corresponde a un SVM. Para su ajuste se utilizan las librerías `caret` y `kernlab` de `R` que tienen implementado algoritmos para el entrenamiento y validación de estos modelos. Antes de hacer el ajuste mismo de los parámetros del modelo se procede a reducción de los atributos como se muestra a continuación.

6.2.1. Reducción de atributos

A diferencia del árbol de decisión, el SVM no tiene incorporada una selección de atributos en el algoritmo de entrenamiento, por lo que una vez definido el conjunto de atributos el modelo los utilizará todos. Dado esto, y ya que uno de los objetivos de este trabajo es reducir la cantidad variables involucradas es que se procede a una reducción de variables antes de ajustar el modelo.

En esta oportunidad y de forma similar a lo que se hizo en el capítulo 5.2 , la reducción se hace a modo de filtro, utilizando un árbol de decisión de tipo CART. Esto es, se entrena un árbol con un sólo split, y se obtiene un ranking de las variables más relevantes usando los competidores del split y el nivel de mejora como indice para rankear cada atributo. En el resultado 6.2 se presentan las 30 variables más importantes utilizando este criterio. La columna `improve` corresponde a la mejora de cada atributo en la creación del árbol, calculada

```
> competidores[1:30,c(3,4)]
                          improve      index
OQPRE8_SD_n              68.85477  0.3750000
RFL19_20_24_SUPAF_n      39.58873  0.6000000
RFL25_SUPAF_n            39.51552  0.7000000
RFL12_SUPAF_n            33.77100  0.5000000
OQPRE13_SD_n             27.93182  0.3750000
RFL5_OBMOR_n             27.86721  0.5000000
OQPRE31_SD_n             27.23881  0.3750000
RFL10_SUPAF_n            26.71089  0.7000000
T_E_DEPRE_MOD_o_SEV      26.11764  0.5000000
OQPRE24_SD_n             24.76890  0.3750000
TIENE_1_HIJO             24.11974  0.5000000
RFL45_SUPAF_n            23.85561  0.7000000
RFL17_SUPAF_n            23.56185  0.5000000
RFL22_SUPAF_n            23.19636  0.5000000
DEQPRE62_n               22.37086  0.5833335
RFL50_n                  20.75243  0.7000000
OQPRE3_SD_n              20.67827  0.1250000
RFL2_SUPAF_n             19.92760  0.5000000
RFL40_SUPAF_n            19.63845  0.5000000
RFL14_SUPAF_n            19.54278  0.5000000
RFL32_SUPAF_n            19.36551  0.7000000
DEQPRE11_n               19.06104  0.2500000
RFL35_SUPAF_n            18.15385  0.7000000
DEQPRE48_n               16.20633  0.4166665
RFL39_SUPAF_n            15.73881  0.9000000
OQPRE40_SD_n             15.72367  0.6250000
OQPRE23_SD_n             15.40155  0.6250000
DEQPRE16_n               15.23665  0.7500000
RFL1_RESFA_n             15.09061  0.9000000
STAXIPRE10_EDO_n         15.04715  0.2500000
```

Resultado de R 6.2: 30 variables más relevantes

en base a la función de ganancia del modelo CART, donde se observa que nuevamente el campo OQPRE8_SD_n es el más importante (anteriormente se había obtenido un resultado similar usando un test chi-cuadrado para rankear los atributos). Por su parte, el campo `index` indica el punto de corte que se hubiera utilizado si la variable se hubiera ocupado en el split en vez de la división primaria escogida.

Con este nuevo ranking se generan 3 nuevos conjuntos de datos, uno con 10, 20 y 30 atributos, y se procede a ajustar los parámetros del SVM para cada uno de ellos. El conjunto de datos y el modelo finalmente escogido se muestran en las secciones siguientes.

6.2.2. Ajuste de parámetros para el SVM

Para el ajuste del modelo SVM se necesita determinar dos parámetros, el *costo* y el parámetro *sigma* o *gamma*, asociado a un kernel de tipo Radial Basis Function. Si bien otros tipos de kernel fueron testeados, acá solo se presentan los resultados obtenidos por uno de tipo RBF que fué el que presentó mejores resultados. Y además es necesario determinar el conjunto de

datos a utilizar (el de 10, 20 o 30 variables), luego, es posible ver la selección de atributos como un parámetro más a ajustar.

El ajuste del parámetro *costo* se hace con valores entre $\{2^k : k = -4, ..., 4\}$, mientras que el parámetro *sigma* se evalúa entre $\{2^j : j = -6, ..., -1\}$, y se utiliza validación cruzada con $k = 5$ y $n = 10$. Los resultados del ajuste se muestran en las figuras 6.6 , 6.7 y 6.8.

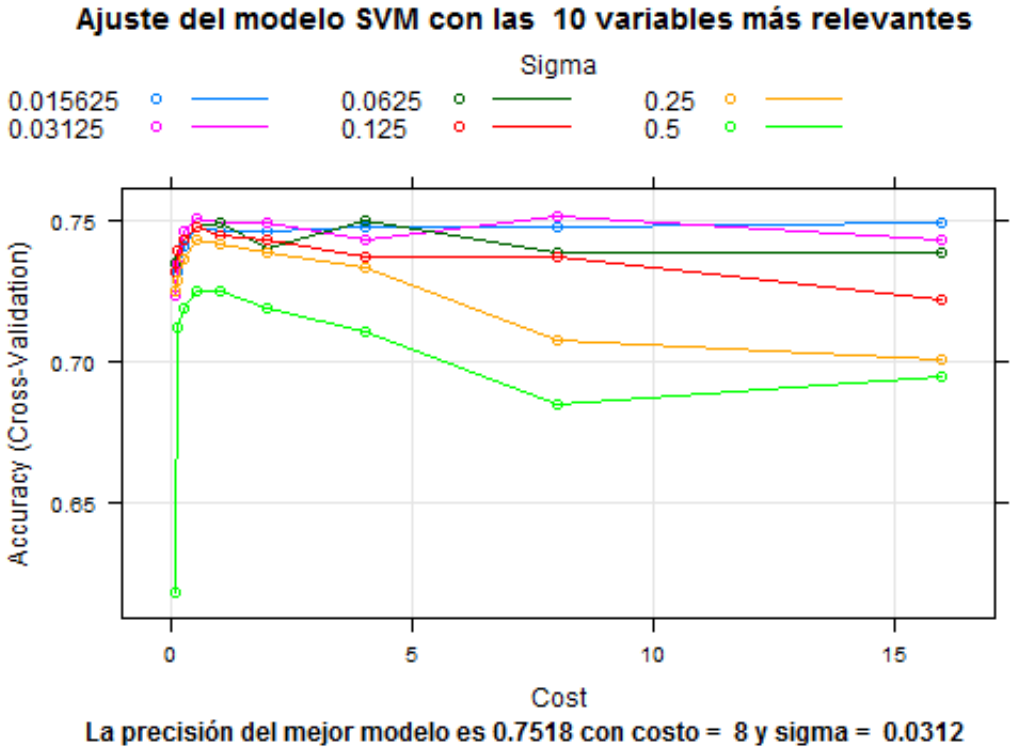

Figura 6.6: Ajuste del modelo SVM entrenado con las 10 variables más relevantes.

De los resultados obtenidos, se tiene que el mejor ajuste se tiene en el conjunto de entrenamiento con las mejores 20 variables, donde si bien la diferencia con el conjunto con 30 atributos es muy pequeña, se favorece aquel con menos atributos. Así, usando este conjunto de datos, el ajuste de los parámetros indica que la mejor elección es tomar el $costo = 1$, y el valor de $sigma = 0{,}0625$.

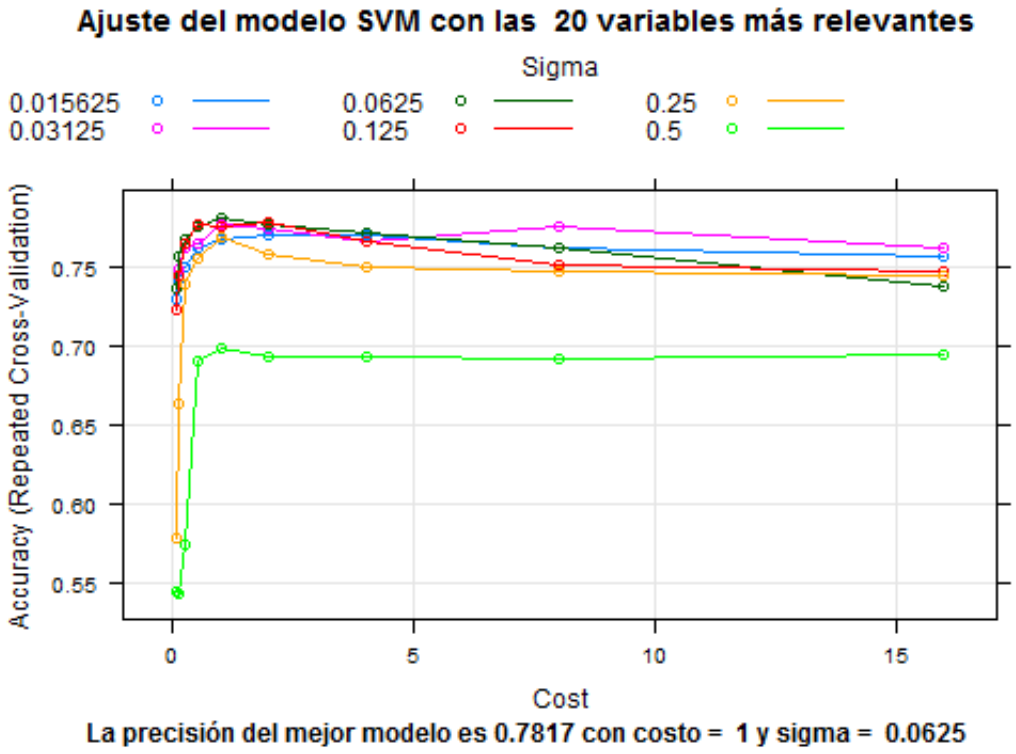

Figura 6.7: Ajuste del modelo SVM entrenado con las 20 variables más relevantes.

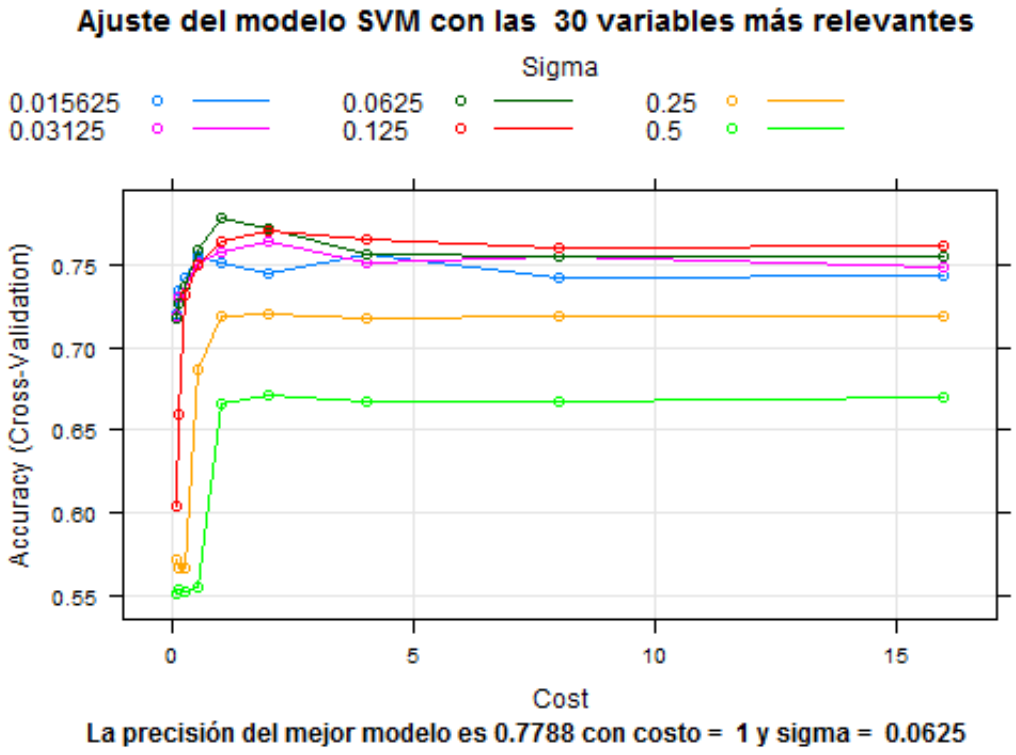

Figura 6.8: Ajuste del modelo SVM entrenado con las 30 variables más relevantes.

A continuación se genera una instancia de aprendizaje y validación para analizar la matriz

de confusión. De forma análoga a lo realizado al ajustar el modelo CART se entrena con el 90 % de las observaciones donde se realiza validación cruzada con $n = 10$. Y el 10 % restante de los datos se utiliza como conjunto de comprobación, y de donde se extraen estadísticos descriptivos de su desempeño para esta instancia dada.

El resultado del entrenamiento se observa en el resultado 6.3. Por su parte, los resultados del modelo en el conjunto de comprobación se presentan en el resultado 6.4.

```
> svm.opt
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 1

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0625

Number of Support Vectors : 420

Objective Function Value : -267.8463
Training error : 0.122029
Cross validation error : 0.221949
```

Resultado de R 6.3: Entrenamiento del modelo SVM con parámetros óptimos de *sigma* y *costo*

Notar que la precisión usando validación cruzada sobre el 90 % de los datos es levemente menor (0,778) al obtenido en la instancia de entrenamiento-comprobación, donde la precisión era 0,8143. Para evitar estas diferencias asociadas a la elección de los conjuntos de entrenamiento y prueba es que en el capítulo 7 se presentan estimadores más robustos para todos los modelos, utilizando k-validación cruzada de n-folds con $k = 100$ y $n = 10$.

```
Confusion Matrix and Statistics

              Reference
Prediction       Grupo_control Grupo_riesgo
  Grupo_control             28            6
  Grupo_riesgo               7           29

               Accuracy : 0.8143
                 95% CI : (0.7034, 0.8972)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 5.145e-08

                  Kappa : 0.6286
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8286
            Specificity : 0.8000
         Pos Pred Value : 0.8056
         Neg Pred Value : 0.8235
             Prevalence : 0.5000
         Detection Rate : 0.4143
   Detection Prevalence : 0.5143
      Balanced Accuracy : 0.8143

       'Positive' Class : Grupo_riesgo
```

Resultado de R 6.4: Matriz de confusión y estadísticas del modelo SVM generado

6.3. Modelo KNN

El modelo KNN está implementado en R mediante varias librerías, `class`, `kknn` y `rknn` entre otras, sin embargo, se puede acceder a varios de estos paquetes mediante la librería `caret`, que permite tanto calibrar los modelos como seleccionar atributos entre muchas otras cosas.

Al igual que lo que sucede con el modelo SVM, el modelo KNN no incluye de forma nativa una selección de atributos como el CART, luego es necesario reducir la cantidad de atributos antes del ajuste de los parámetros y de la determinación de su precisión.

En este sentido, la selección de atributos también puede considerarse como un ajuste de parámetros, en especial cuando el método utilizado para seleccionar y/o reducir variables es del tipo *wrapper*, esto es, cuando la selección de parámetros está embebida en el modelo mismo, o dicho de otra manera, cuando el agregar o quitar variables se se evalúa según la precisión del mismo modelo. De esta manera, la selección de variables deja de ser un filtro puramente, si no, que se va ajustando con el resto de los parámetros.

Sin embargo, aún cuando este enfoque parece el más apropiado, posee un alto coste computacional, ya que por cada instancia de los parametros a ajustar se debiera evaluar el modelo

para una gran cantidad de combinaciones distintas de las variables de entrada. Además, en la práctica, tal tarea no siempre rinde sus frutos. Un mayor análisis de esto se puede ver en [7], donde se evalúa la eficacia de dos metodologías distintas para la selección de atributos de tipo *wrapper*.

En esta memoria, y en particular para el modelo aqui presentado donde se utiliza una reducción de este estilo se opta por un enfoque separado para los ajustes, por una parte se ajusta el conjunto de variables más relevante utilizando los parámetros del modelo como fijos, y luego en la sección siguiente se procede a ajustar el parámetro k de vecinos pero con los atributos ya definidos.

6.3.1. Reducción de atributos

Como se comentó, el modelo KNN no selecciona atributos de forma nativa, sin embargo, la librería `caret` proporciona un *metodo wrapper* para la selección de atributos mediante la técnica de *Recursive Featute Elimination* (RFE). En la figura 6.9 se muestra como opera este algoritmo, mientras que mas detalle se puede encontrar en [36].

Algorithm 2: Recursive feature elimination incorporating resampling

```
for Each Resampling Iteration do
    Partition data into training and test/hold–back set via resampling
    Tune/train the model on the training set using all predictors
    Predict the held–back samples
    Calculate variable importance or rankings
    for Each subset size S_i, i = 1 ... S do
        Keep the S_i most important variables
        [Optional] Pre–process the data
        Tune/train the model on the training set using S_i predictors
        Predict the held–back samples
        [Optional] Recalculate the rankings for each predictor
    end
end
Calculate the performance profile over the S_i using the held–back samples
Determine the appropriate number of predictors
Estimate the final list of predictors to keep in the final model
Fit the final model based on the optimal S_i using the original training set
```

Figura 6.9: Algoritmo RFE

Luego, se procede a calcular la precisión del modelo KNN para diferentes tamaños del conjunto de variables. La precisión es calculada usando validación cruzada con $n = 10$, y los resultados se muestran en la figura 6.10. De aquí se obtiene que el mejor tamaño corresponde a 20 variables, con una precisión de 0,7176 como se observa en el resultado 6.5. Por su parte, las variables seleccionadas por el algoritmo se detallan en el resultado 6.6.

```
> knnProfile

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold, repeated 1 times)

Resampling performance over subset size:

 Variables  Accuracy  Kappa AccuracySD KappaSD Selected
         3    0.5976 0.1994    0.06225 0.12449
         4    0.6119 0.2260    0.08383 0.16805
         5    0.6834 0.3648    0.04978 0.09932
         6    0.6947 0.3890    0.05354 0.10734
         7    0.6649 0.3297    0.04577 0.09139
         8    0.6891 0.3780    0.03521 0.07074
         9    0.6790 0.3580    0.06658 0.13304
        10    0.6633 0.3265    0.04925 0.09876
        15    0.6878 0.3755    0.03707 0.07421
        20    0.7176 0.4353    0.04632 0.09229        *
        30    0.7176 0.4350    0.06001 0.11961
        50    0.7048 0.4099    0.04800 0.09604
       100    0.6991 0.3986    0.07048 0.14073
       129    0.7148 0.4295    0.06010 0.11991

The top 5 variables (out of 20):
   OQPRE8_SD_n, RFL19_20_24_SUPAF_n, RFL25_SUPAF_n, RFL12_SUPAF_n, OQPRE13_SD_
      n
```

Resultado de R 6.5: Resultado del modelo RFE

```
> knnProfile$optVariables
 [1] "OQPRE8_SD_n"          "RFL19_20_24_SUPAF_n" "RFL25_SUPAF_n"       "RFL12_
     SUPAF_n"
 [5] "OQPRE13_SD_n"         "OQPRE31_SD_n"        "RFL5_OBMOR_n"        "
     OQPRE24_SD_n"
 [9] "RFL45_SUPAF_n"        "RFL22_SUPAF_n"       "RFL10_SUPAF_n"       "
     DEQPRE62_n"
[13] "RFL14_SUPAF_n"        "OQPRE3_SD_n"         "RFL50_n"             "
    OQPRE23_SD_n"
[17] "T_E_DEPRE_MOD_o_SEV" "DEQPRE48_n"          "RFL32_SUPAF_n"       "RFL2_
    SUPAF_n"
```

Resultado de R 6.6: Variables seleccionadas por el modelo RFE

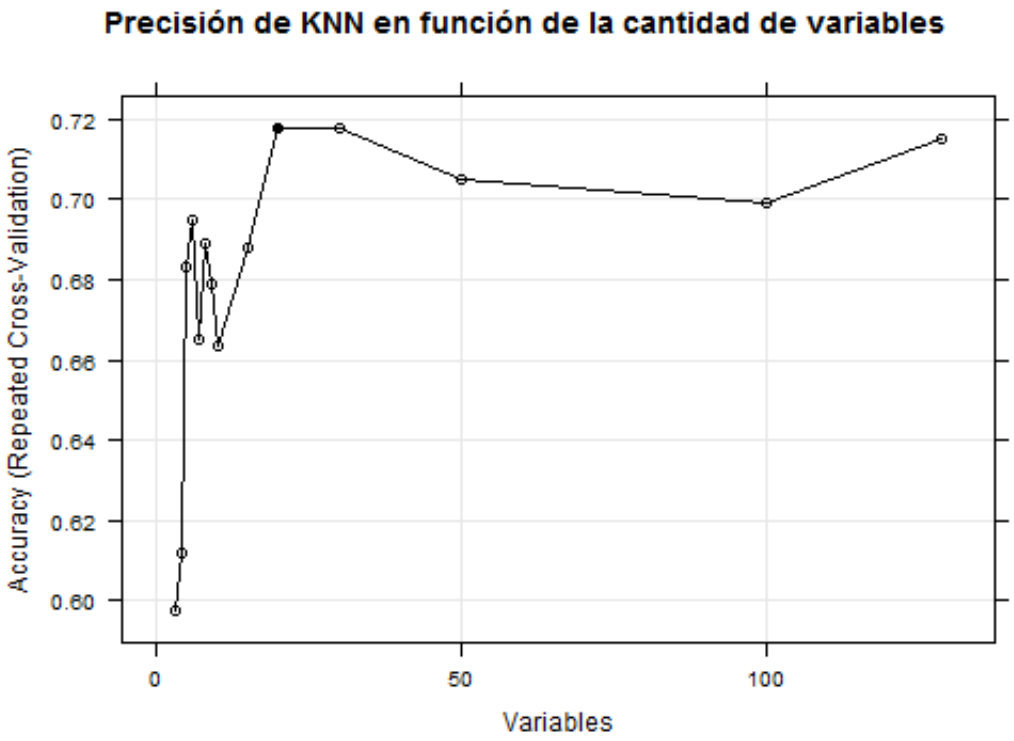

Figura 6.10: Precisión estimada del modelo KNN para diferentes tamaños del conjunto de variables

Nuevamente cabe destacar que el campo OQPRE8_SD_n y el campo RFL19_20_24_SUPAF_n han sido definidos de alta importancia, lo que da consistencia a los resultados obtenidos por los diferentes modelos a lo largo de este trabajo.

6.3.2. Ajuste de parámetros para el modelo KNN

Una vez definido los atributos a utilizar se procede a ajustar el parámetro k asociado a la cantidad de vecinos que usa el modelo para clasificar cada nueva observación. Para su ajuste se prueban varios valores de k y mediante validación cruzada (tomando número de iteraciones $= 10$ y $n = 10$) se calcula la precisión promedio.

Los resultados son presentados en la figura 6.11 , de donde se tiene que el mejor valor para k es 16, con una precisión de 0,7322, sin embargo, dado que el k óptimo es par, esto puede inducir en "empates" en la votación de los vecinos. Luego, como regla de buena práctica se escoge un k impar, en este caso el k con la mejor pecisión. Así finalmente, el k escogido es 17.

Para el análisis de los resultados se procede de manera análoga a lo hecho con los otros modelos, esto es, dar ciertas medidas estadísticas para una instancia de entrenamiento y validación, dejando un análisis más en profundidad de la precisión para más adelante en el

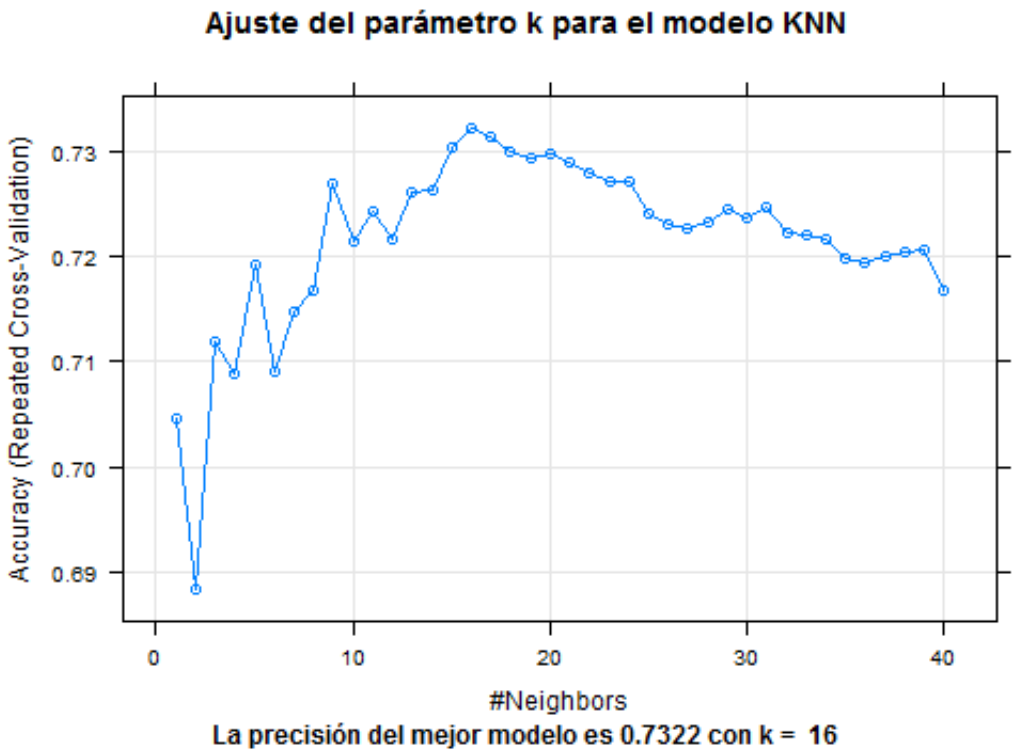

Figura 6.11: Ajuste del parámetro k

capítulo 7. Del entrenamiento y validación para una instancia dada, se obtiene lo mostrado en el resultado 6.7 . El conjunto de entrenamiento considera el 90 % de las observaciones, mientras que el de validación corresponde al 10 % restante como ha sido usual.

```
Confusion Matrix and Statistics

                Reference
Prediction       Grupo_control Grupo_riesgo
  Grupo_control             26           11
  Grupo_riesgo               9           24

               Accuracy : 0.7143
                 95% CI : (0.5938, 0.816)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.0002201

                  Kappa : 0.4286
 Mcnemar's Test P-Value : 0.8230633

            Sensitivity : 0.6857
            Specificity : 0.7429
         Pos Pred Value : 0.7273
         Neg Pred Value : 0.7027
             Prevalence : 0.5000
         Detection Rate : 0.3429
   Detection Prevalence : 0.4714
      Balanced Accuracy : 0.7143

       'Positive' Class : Grupo_riesgo
```

Resultado de R 6.7: resultados para una instancia de entrenamiento - validación para el modelo KNN

6.4. Ensemble Models

Ensemble models o *ensemble learning* consiste en ejecutar un modelo o algoritmo base muchas veces y definir una respuesta global como combinación de las respuestas de los múltiples modelos generados. En general los *ensemble models* se pueden agrupar en dos grandes categorías:

- **Bagging**: Es un método que genera múltiples versiones de un predictor, y usa todas las respuestas para obtener un predictor agregado, ya sea promediando (en caso de regresión) o por votación (en caso de clasificación). Las múltiples versiones del modelo se forman al utilizar bootstrap y generar nuevos conjuntos de entrenamiento al seleccionar observaciones de forma aleatoria y con repetición desde el conjunto de datos original.
- **Boosting**: Por su parte Boosting opera de manera similar, también se generan múltiples clasificadores, y el resultado final del modelo viene dado por una combinación de todas las versiones del modelo generada. Sin embargo, posee una diferencial sustancial, ya que a diferencia de lo que pasa al usar bagging, en este caso, cada nuevo conjunto de entrenamiento se define en base al error cometido por las versiones anteriores. Así, en cada iteración del método la distribución con la cual se escoge el nuevo conjunto de entrenamiento cambia. Adicionalmente la agregación de las respuestas de los múltiples clasificadores ya no se calcula mediante simple promedio o moda, si no que se utiliza un promedio ponderado, donde cada versión posee un peso específico distinto en el modelo combinado.

Sin embargo hay un par de algoritmos que no caen en ninguna de éstas categorías, como por ejemplo los *random forest*, que si bien tienen muchas semejanzas con los del tipo Bagging, no se clasifican dentro este grupo. En este capitulo ajustan dos *ensemble models*, un modelo de tipo Boosting, denominado AdaBoost, y un random forest, los resultados de los ajustes se muestran en las secciones siguientes.

6.4.1. Algoritmo AdaBoost

El algoritmo AdaBoost desarrollado por Schapire y Freund ([22] , [21]) consiste en generar múltiples árboles de decisión, sobre subconjuntos de las observaciones, donde cada observación es elegida con una distribución de probabilidad calculada en cada iteración del algoritmo, y el resultado final del modelo se obtiene mediante la votación ponderada de cada uno de los árboles generados. El algoritmo AdaBoost se encuentra dentro del enfoque de algoritmos de tipo Boostinng.

Para generar un modelo en `R` con este algoritmo se utilizan las librerías `caret` y `adabag`. El ajuste se realiza mediante validación cruzada con $n = 10$ folds, usando el método `train` del paquete `caret`, y donde los parámetros a ajustar son 3:

- Número de árboles
- Profundidad máxima de cada árbol

- Coeficiente α

El número de árboles controla la cantidad de árboles de decisión que se generan y que se son usados en la votación para la respuesta final del modelo. Por su parte la profundidad máxima esta asociado al máximo número de niveles que puede tener cada árbol. Y por último, el coeficiente α está asociado a la actualización de los pesos de las muestras en cada iteración del algoritmo. El parámetro α puede tomar alguno de los siguientes 3 valores:

- Breiman:
$$\alpha_{Breiman} = \frac{1}{2} ln\Big(\frac{1 - err}{err}\Big)$$
- Freund:
$$\alpha_{Freund} = ln\Big(\frac{1 - err}{err}\Big)$$
- Zhu:
$$\alpha_{Zhu} = ln\Big(\frac{1 - err}{err}\Big) + ln(nclasses - 1)$$

Luego se procede a buscar la mejor combinación de parámetros, para lo cual se hacen variar los parámetros antes descritos en los siguientes rangos / valores:

- Número de árboles: entre 3 y 100
- Profundidad máxima: entre 3 y 7
- Coeficiente α: Con los valores "Breiman", "Zhu" y "Freund"

Los resultados obtenidos para todas los combinaciones de estos 3 parámetros se muestran en la figura 6.12, donde se ha calculado la precisión mediante k-validación cruzada de n-folds, con $n = 10$ y $k = 3$. De aquí se tiene que el mejor conjunto de parámetros corresponde a tomar el número de árboles igual a 33, máxima profundidad del árbol igual a 4 y usando un coeficiente de tipo "Breiman". Con estos parámetros la precisión preliminar estimada es de 0,7783803 como se muestra en el resultado 6.8. Recordar que en el capítulo 7 se mostrarán medidas de precisión más robustas.

```
> AdaBoost.fit$bestTune
    mfinal maxdepth coeflearn
325     33        4    Breiman
> print(max(AdaBoost.fit$results[,4]))
[1] 0.7783803
```

Resultado de R 6.8: Parámetros óptimos y precisión estimada del algoritmo AdaBoost

En el cuadro de resultados 6.9 se muestran algunos indicadores para una instancia de entrenamiento y validación, usando como conjunto test el 10 % de los datos.

```
Confusion Matrix and Statistics

                Reference
Prediction       Grupo_control Grupo_riesgo
  Grupo_control             29           11
  Grupo_riesgo               7           24

               Accuracy : 0.7465
                 95% CI : (0.6292, 0.8423)
    No Information Rate : 0.507
    P-Value [Acc > NIR] : 3.199e-05

                  Kappa : 0.4921
 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 0.8056
            Specificity : 0.6857
         Pos Pred Value : 0.7250
         Neg Pred Value : 0.7742
             Prevalence : 0.5070
         Detection Rate : 0.4085
   Detection Prevalence : 0.5634
      Balanced Accuracy : 0.7456

       'Positive' Class : Grupo_control
```

Resultado de R 6.9: Indicadores de precisión para el modelo AdaBoost para una instancia fija de entrenamiento y validación.

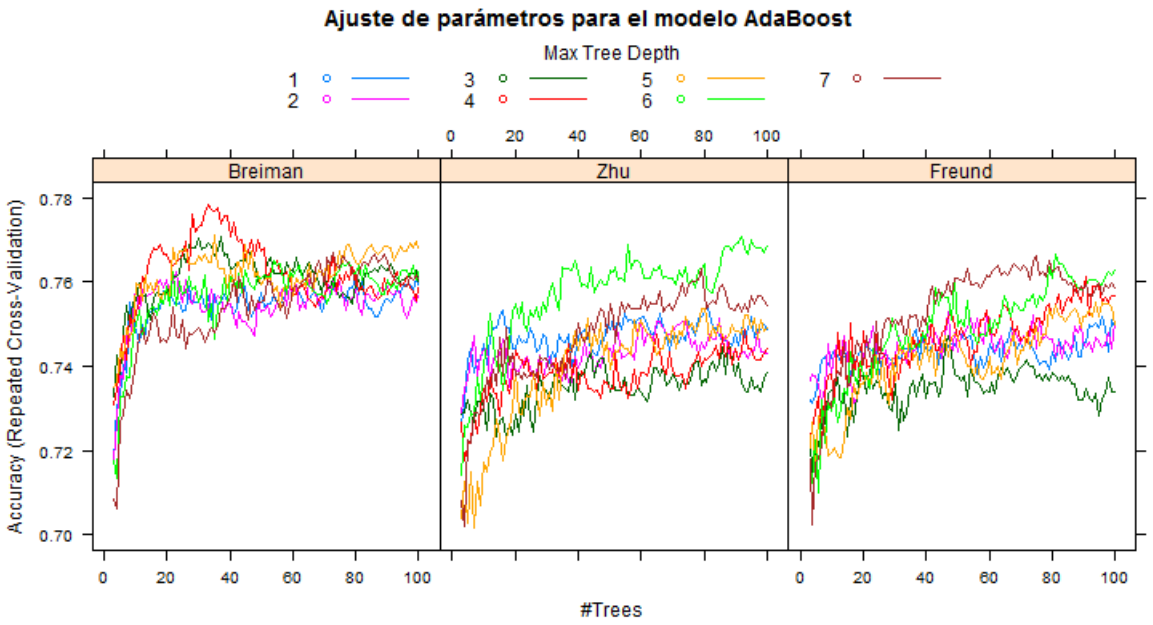

Figura 6.12: Ajuste de parámetros para el algoritmo AdaBoost.

6.4.2. Random Forest

Random Forests es una técnica de agregación desarrollada por Leo Breiman [10], que mejora la precisión en la clasificación mediante la incorporación de aleatoriedad en la construcción de cada clasificador individual. Donde a diferencia de lo que sucede con el *Bagging* en cada split el algoritmo escoge sólo un subconjunto de los predictores, esto es, la aleatoriedad está presente tanto en las muestras de entrenamiento como en el conjunto de atributos.

El algoritmo *Random Forest* está implementado en `R` mediante la librería `randomForest`, también como sucede con la mayoría de los modelos se puede acceder a él mediante la función `train` de la librería `caret`.

Para el ajuste de parámetros en el algoritmo *random forest* se deja la cantidad de árboles a generar fija e igual a 500, mientras que la optimización se realiza sobre el parámetro asociado a la cantidad de variables que se seleccionan aleatoreamente en cada iteración, que llamaremos *mtry*. Dado que la selección de variables es aleatoria en cada iteración, y son 500 iteraciones, es que finalmente pueden utilizarse (casi seguramente) todas las variables en el modelo final, dado esto, se generan dos instancias de random forest, una con todos los campos, que son 129, y otra donde sólo se utilizan las 30 variables más relevantes según el filtro generado por CART, el detalle de las variables se muestra en el resultado 6.2.

En la figura 6.13 y en la figura 6.14 se muestra el ajuste del párametro asociado a la cantidad de variables seleccionadas en cada iteración para los dos conjuntos de datos antes descritos. Las precisiones obtenidas fueron calculadas usando validación cruzada con $n = 10$ folds.

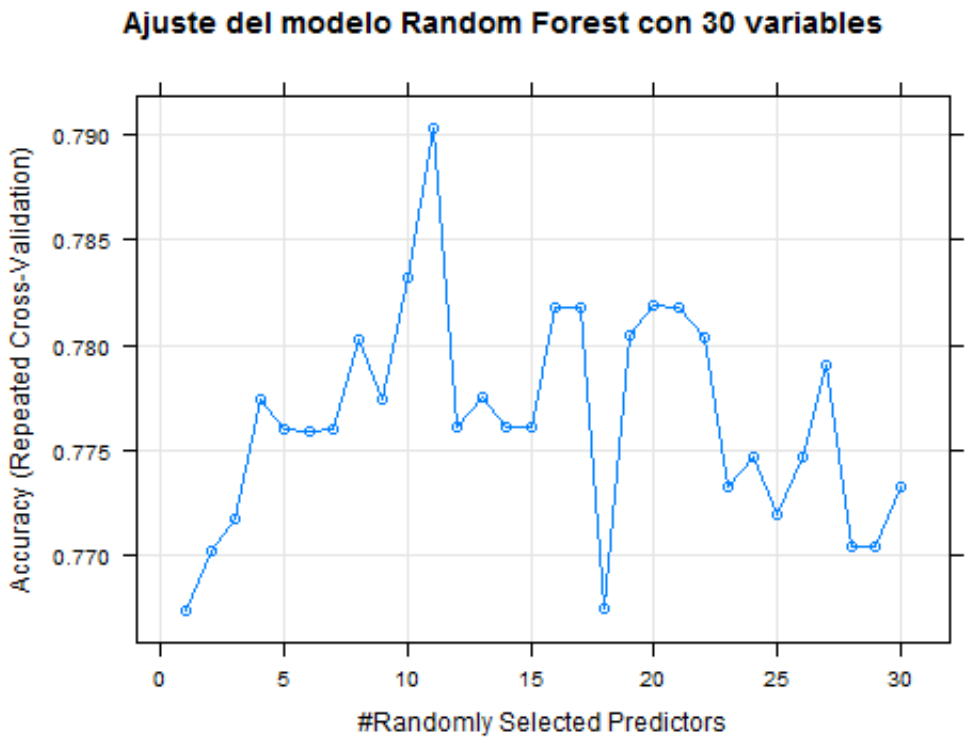

Figura 6.13: Ajuste del modelo random forest usando el conjunto de las 30 variables más relevantes

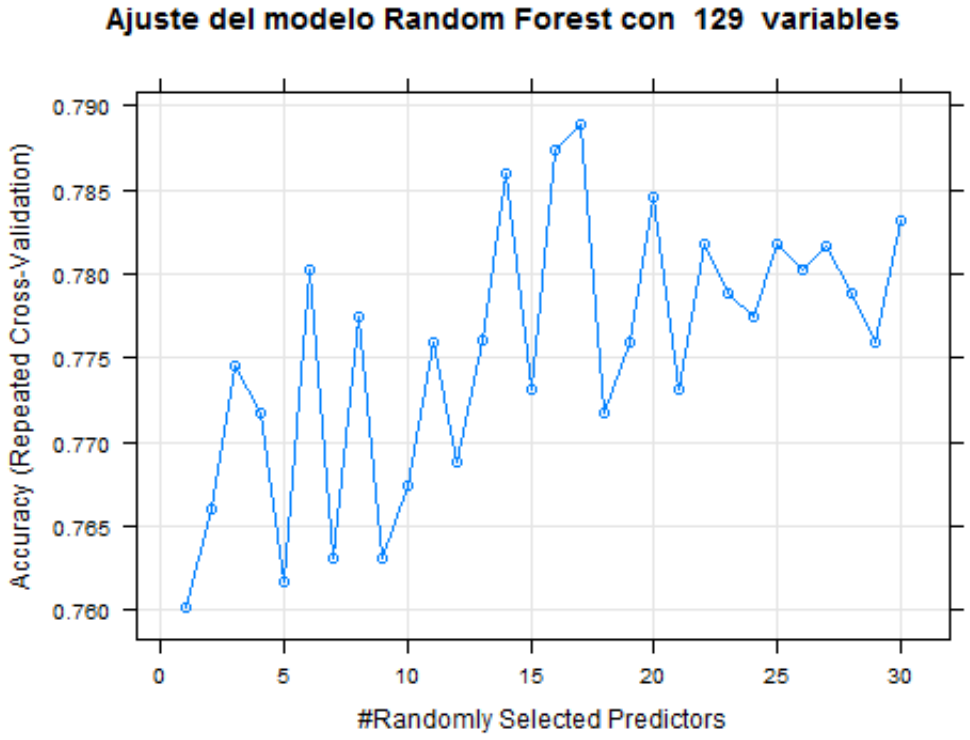

Figura 6.14: Ajuste del modelo random forest usando todas las variables (129)

Luego de la tabla 6.2 se tiene que el valor óptimo para el parámetro analizado corresponde a 11, usando el conjunto con 30 variables.

Dataset	**mtry**	**Accuracy**	**Kappa**	**AccuracySD**	**KappaSD**
129 variables	17	0.7888575	0.5775967	0.0389158	0.07797764
30 variables	11	0.7902948	0.5801993	0.04100712	0.08236513

Tabla 6.2: Resultados del mejor ajuste para ambos conjuntos de datos (con 129 y 30 variables).

Finalmente en el resultado 6.10 se muestra el resultado de una instancia dada (al azar) de entrenamiento y validación, dejando un 10 % de los datos como conjunto de validación. Una mejor estimación de todos los modelos, incluido este se presentará en el capítulo siguiente.

```
Confusion Matrix and Statistics

                Reference
Prediction       Grupo_control Grupo_riesgo
  Grupo_control             30            10
  Grupo_riesgo               6            25

               Accuracy : 0.7746
                 95% CI : (0.66, 0.8654)
    No Information Rate : 0.507
    P-Value [Acc > NIR] : 3.262e-06

                  Kappa : 0.5485
 Mcnemar's Test P-Value : 0.4533

            Sensitivity : 0.8333
            Specificity : 0.7143
         Pos Pred Value : 0.7500
         Neg Pred Value : 0.8065
             Prevalence : 0.5070
         Detection Rate : 0.4225
   Detection Prevalence : 0.5634
      Balanced Accuracy : 0.7738

       'Positive' Class : Grupo_control
```

Resultado de R 6.10: Resultados de una instancia de entrenamiento y validación con el modelo de Random Forest

Capítulo 7

Análisis de resultados

7.1. Comparación de resultados obtenidos

En este capítulo se comparan las medidas de precisión para los 5 modelos generados mediante los siguientes indicadores

- Precisión
- Sensibilidad
- Especificidad
- Curva ROC

Dichos indicadores son calculados utilizando validación cruzada de n-folds, con $n = 10$ y repitiéndose 100 veces, usando el promedio de estas medidas para cada modelo. Al iterar la validación cruzada 100 veces se espera disminuir el ruido y estimar de mejor manera el error sistemático del modelo en cada caso.

Después de las 100 iteraciones los resultados obtenidos se muestran en el cuadro de resultados 7.1. De aquí se tiene que el modelos con la mejor precisión es el SVM, con una precisión media de 0,7796839, el segundo mejor modelo corresponde al Random Forest, el cual tiene una precisión media del 0,77781147, y en tercer lugar está el modelo AdaBoost con una precisión de 0,7559938.

```
> results
                cart       svm       knn  AdaBoost RandomForest
Accuracy    0.72413239 0.7796839 0.7309044 0.7559938   0.77781147
Sensibility 0.70911470 0.7700632 0.7378200 0.7548683   0.78262317
Specificity 0.74105884 0.7908470 0.7252821 0.7581077   0.77454554
AccuracySD  0.05114694 0.0471182 0.0493103 0.0495292   0.04718801
```

Resultado de R 7.1: Indicadores de precisión para los 5 modelos generados

Por su parte las sensibilidades y especificidades de cada modelo se encuentran en torno a la precisión de cada uno, sin grandes diferencias, el que tiene una mayor diferencia entre sensibilidad y especificidad es el modelo CART, mientras que el que tiene una menor diferencia es el modelo AdaBoost, el cual de hecho tiene sus 3 indicadores casi iguales. Por otro lado, notar que el modelo SVM tiene una mayor especificidad que sensibilidad, mientras que el modelo de Random Forest es al revés, es decir, su sensibilidad es mayor que su especificidad. Sin embargo, las diferencias siguen siendo marginales.

En las figuras 7.1 , 7.2 , 7.3 , 7.4 y 7.5 se muestra información adicional de los histogramas con las precisiones obtenidas, mientras que en la figura 7.6 se resume la precisión de los 5 modelos mediante un gráfico boxplot.

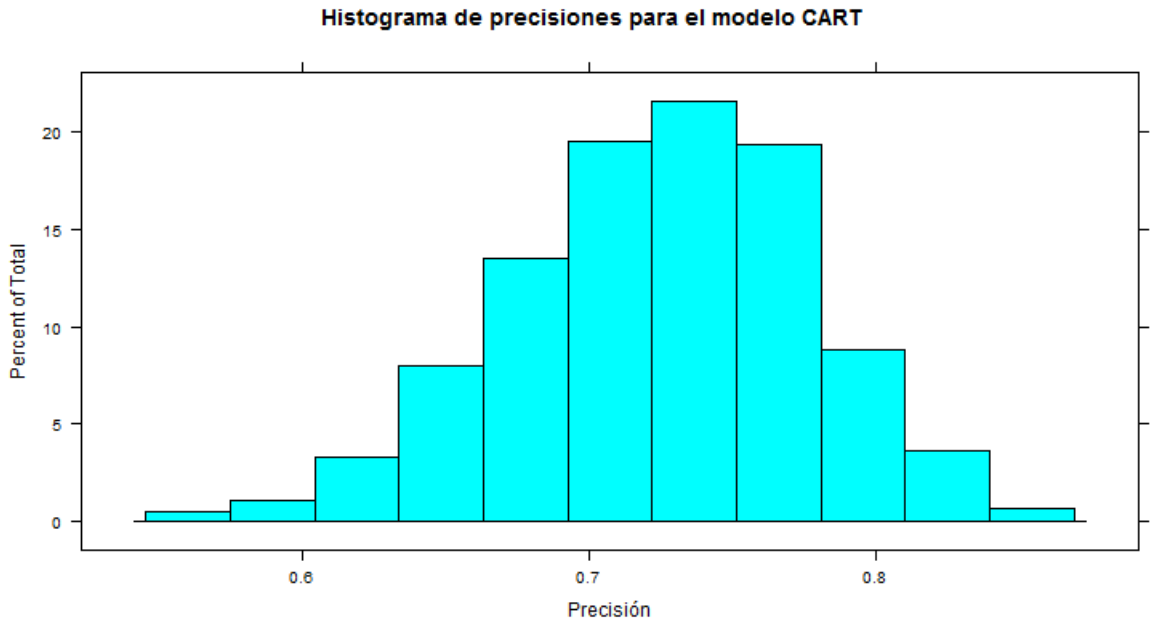

Figura 7.1: Histograma de las precisiones obtenidas por el modelo CART

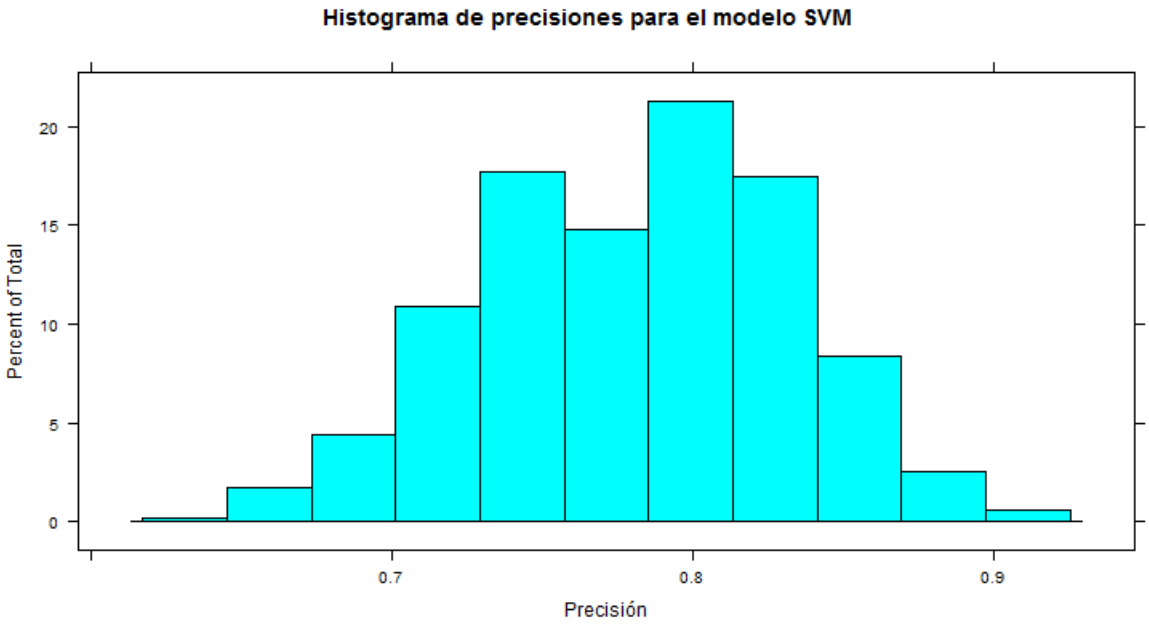

Figura 7.2: Histograma de las precisiones obtenidas por el modelo SVM

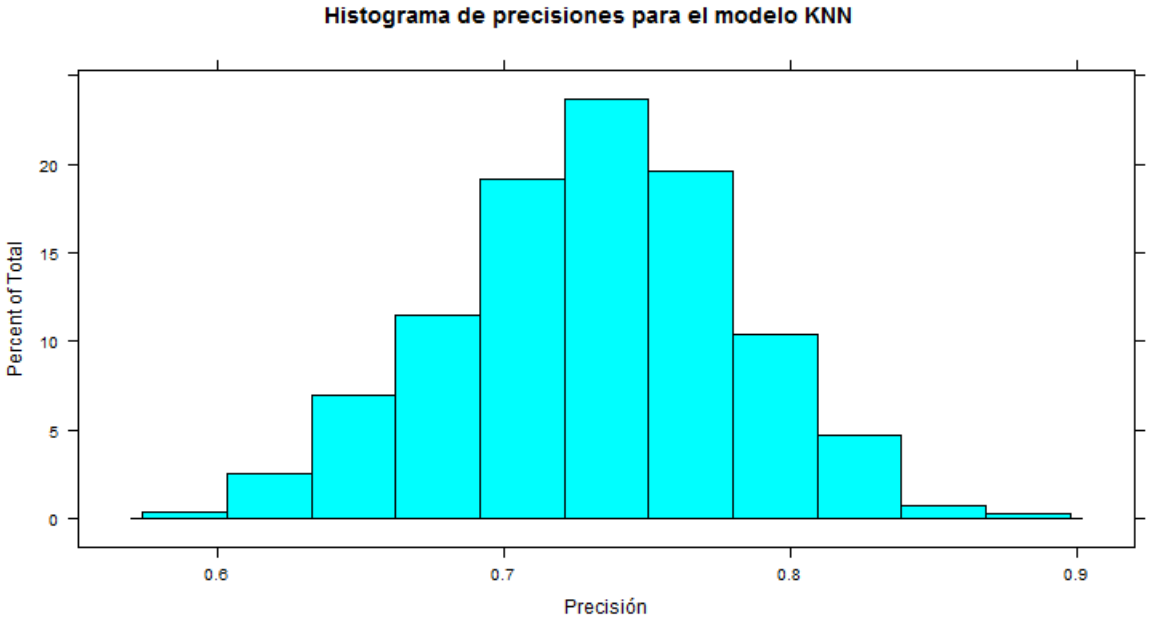

Figura 7.3: Histograma de las precisiones obtenidas por el modelo KNN

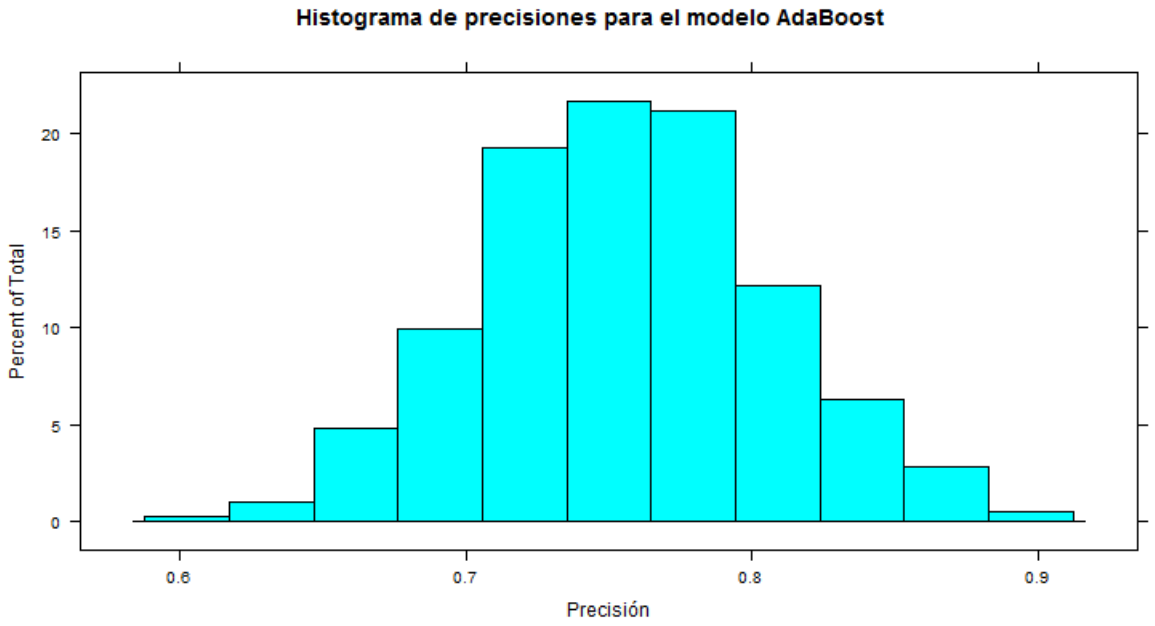

Figura 7.4: Histograma de las precisiones obtenidas por el modelo AdaBoost

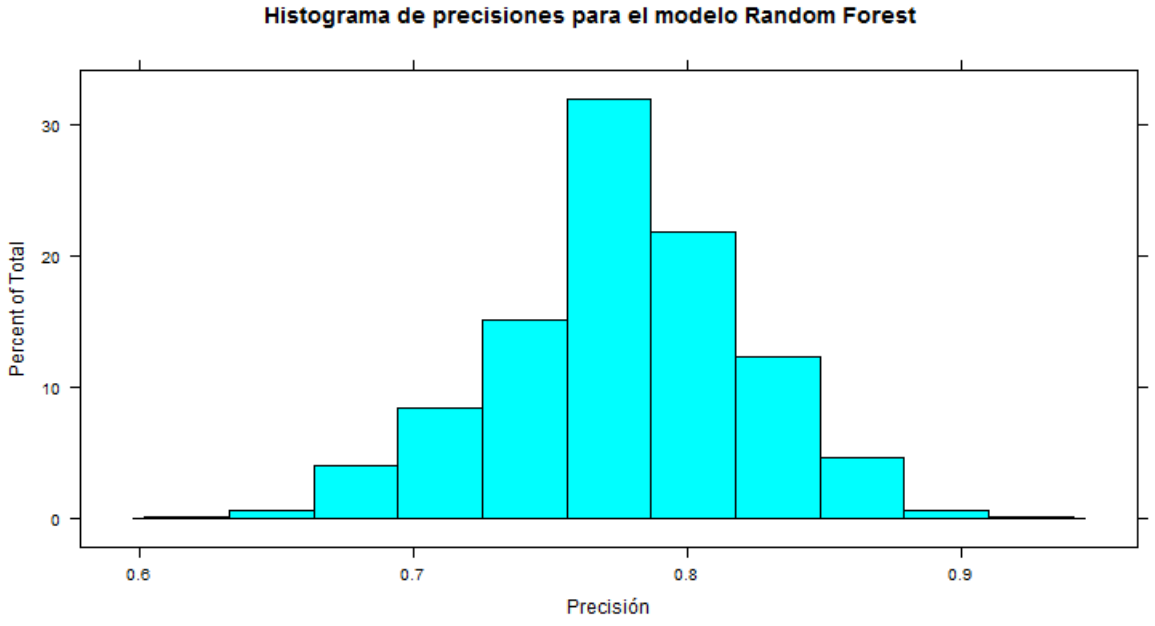

Figura 7.5: Histograma de las precisiones obtenidas por el modelo Random Forest

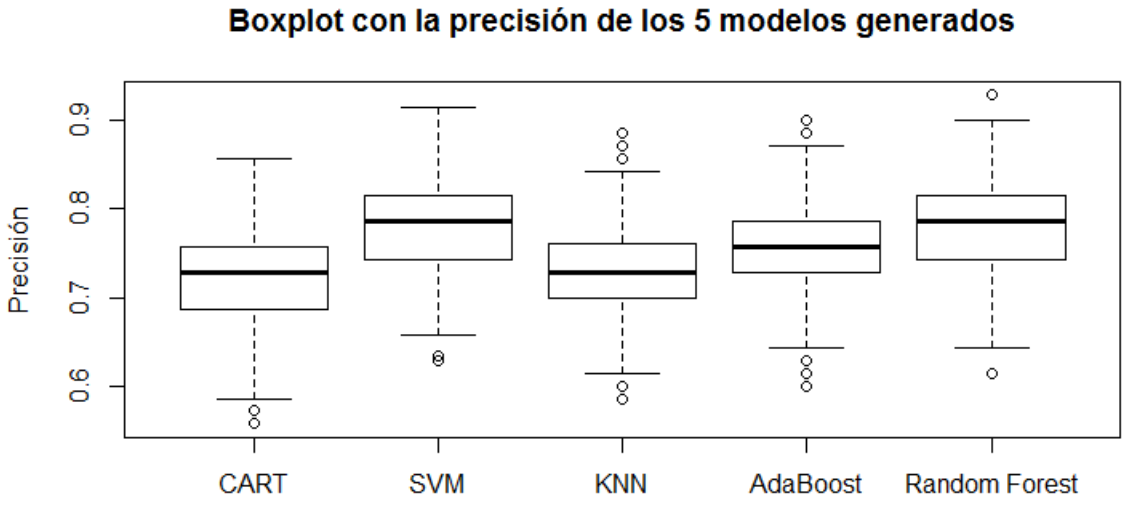

Figura 7.6: Boxplot con las precisiones obtenidas por los 5 modelos desarrollados

7.2. Curva ROC

Finalmente como parte de los resultados se gráfica cada uno de los modelos generados en el espacio ROC, usando la sensibilidad y especificidad promedio como parte de las coordenadas de cada modelo. Lo anterior se muestra en la figura 7.7.

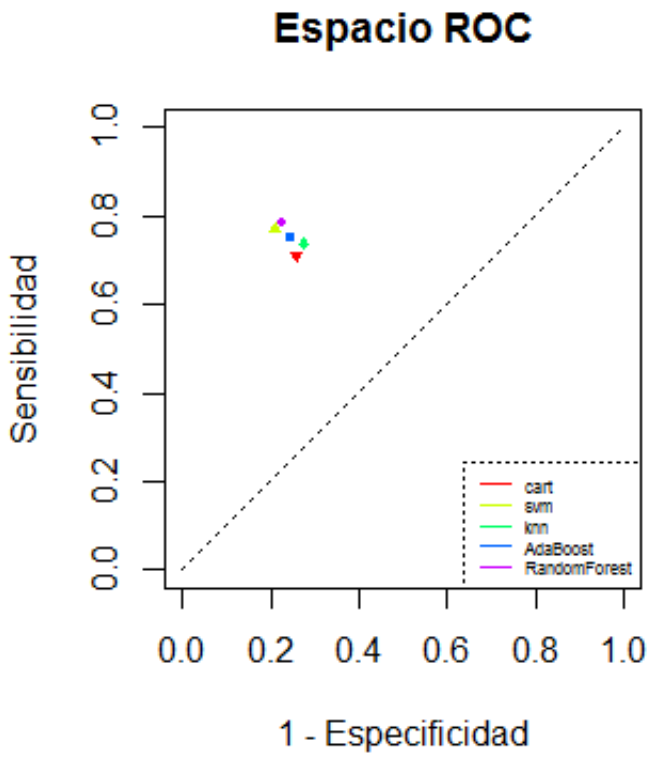

Figura 7.7: Espacio ROC con los 5 modelos generados, CART, SVM, KNN, AdaBoost y Random Forest.

Dado que todos los modelos están concentrados en un cuadrante específico del espacio ROC se hace un zoom para poder identificar de mejor manera cada punto. Esto se presenta en la figura 7.8

De la figura 7.8 se remarca el hecho que tanto como el SVM como el modelo de Random Forest ofrecen mejores resultados que las otras 3 alternativas. Siendo el SVM levemente superior al Random Forest en cuanto a precisión. Sin embargo, en el espacio ROC la diferencia es mucho menos notoria.

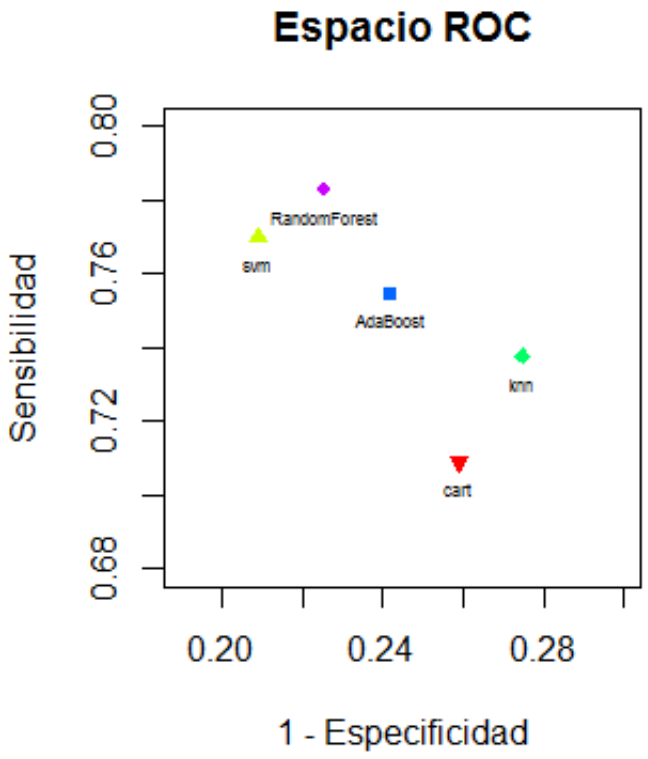

Figura 7.8: Zoom al espacio ROC con los 5 modelos generados.

Conclusión

Durante la presente memoria se estudió el problema de riesgo suicida desde una perspectiva de data mining/machine learning, incluyendo desde la preparación de los datos hasta el desarrollo de modelos matemáticos para la clasificación del riesgo suicida. Usando estas técnicas se generó un modelo de clasificación del riesgo suicida cumpliendo el objetivo de esta memoria.

La primera parte del trabajo se cumplió de manera satisfactoria, donde se depuraron y limpiaron los datos, preparándolos para su uso y cumpliendo así el primer objetivo específico de este trabajo. Se ocuparon técnicas estadísticas para la imputación de datos que, junto al criterio experto en el depurado y limpieza, permitieron crear una base de datos apta para su uso en el modelamiento. De esta manera, sólo 6 registros fueron eliminados, pues la imputación para valores perdidos fue cuidadosa y caso a caso, asegurando una buena imputación de los datos faltantes, y reduciendo al máximo la cantidad de registros eliminados.

En el análisis, transformación y selección de atributos se identificaron aquellas variables más relevantes para el problema estudiado, tales como la pregunta OQ8 (Pienso en quitarme la vida), RFL12 (La vida es lo único que tenemos y es mejor que tener nada), RFL19 (Me quiero lo suficiente como para vivir), RFL25 (Soy demasiado estable como para matarme) y OQ13 (Soy una persona feliz) entre otras. Mientras que la cantidad de hijos y el diagnóstico de tipo "episodio o trastorno depresivo moderado o severo" fueron de las variables descriptivas más relevantes. Si bien estas variables en cierta forma se validan por los conocimientos previos de los especialistas de la salud, en este trabajo se le da sustento estadístico a su importancia, generándose rankings y definiendo interrelaciones entre estas variables mediante los modelos desarrollados. Con respecto a la transformación de las variables se probaron varios tipos de transformaciones y asociación de variables tales como transformaciones binarias, sigmoideas y logit entre otras, sin embargo, los resultados obtenidos no presentaron mayores mejoras, por lo que el uso de variables (de instrumentos) fue mediante una normalización min-max únicamente. Así, el uso de técnicas de minería de datos permitió obtener 22 variables de un total de 343, lo que resulta del aporte de la depuración y parsimonia en la investigación realizada.

De los modelos generados, el que presenta un mejor desempeño corresponde al Support Vector Machine, el cual posee una precisión media de 77,96 %, con una sensibilidad del 77 % y una especificidad del 79 %. La precisión del modelo fue calculada mediante k-validación

cruzada de n-folds, con $k = 100$ y $n = 10$, mientras que el ajuste de los parámetros se realizó tomando $k = 5$ y $n = 10$. La precisión del modelo en el conjunto de entrenamiento es aproximadamente de 87,8 % como se mostró en el cuadro de resultados 6.3, luego si bien es 10 puntos porcentuales mayor a la precisión obtenida en el conjunto de validación no es tan alto como para suponer un sobreajuste, y de esta manera se acepta el modelo propuesto.

Dentro del trabajo futuro se propone seguir investigando nuevas y mejores técnicas para la reducción de atributos, en este trabajo se expusieron 3 técnicas distintas, sin embargo, dada la cantidad de variables iniciales se cree que una mejor selección de las variables podría incidir en una mayor precisión del modelo.

Por otra parte, dentro de los próximos pasos, las 22 variables aquí seleccionadas y el modelo generado serán traducidas en un instrumento de evaluación clínico, que será de fácil aplicación y adecuado para la detección del riesgo suicida, siendo utilizado en los servicios de urgencia, ambulatorios y unidades de hospitalización, donde se reciban consultantes con conducta suicida.

Bibliografía

[1] Proceso de extracción de conocimiento kdd. <http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/>. [Consultado: 2015-05-10].

[2] Alan C Acock and David H Demo. *Family diversity and well-being*. Sage Thousand Oaks, CA, 1994.

[3] Louis Appleby, Jenny Shaw, Tim Amos, Ros McDonnell, Catherine Harris, Kerry McCann, Katy Kiernan, Sue Davies, Harriet Bickley, Rebecca Parsons, et al. Suicide within 12 months of contact with mental health services: national clinical survey. *Bmj*, 318(7193):1235–1239, 1999.

[4] American Psychiatric Association et al. Practice guideline for the assessment and treatment of patients with suicidal behaviors. *American Journal of Psychiatry*, 160(11):1–60, 2003.

[5] Tomas Baader-Matthei, Paul Richter, and Christoph Mundt. Suicidios de pacientes psiquiátricos hospitalizados y sus factores de riesgo: Un estudio caso control. *Revista chilena de neuro-psiquiatría*, 42(4):293–316, 2004.

[6] A Beautrais. Suicidio: estado actual de la ciencia. *Presented on IV Jornadas de Psiquiatría, Suicidio: Prevención, evaluación y Tratamiento*, 2009.

[7] Christoph Bernau and Anne-Laure Boulesteix. Variable selection and parameter tuning in high-dimensional prediction. 2010.

[8] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[9] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[10] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[11] MJ Burgueño, JL García-Bastos, and JM González-Buitrago. Las curvas roc en la evaluación de las pruebas diagnósticas. *Med Clin (Barc)*, 104(17):661–70, 1995.

[12] G Burlingame and MJ Lambert. Administration and scoring manual: Oq-45.2, 1996.

[13] Jonathan Cheung-Wai Chan and Desiré Paelinckx. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112(6):2999–3011, 2008.

[14] Jaime Correa, Ramón Florenzano, Pilar Rojas, Juan Francisco Labra, Verónica del Río, and Juan Andrés Pastén. El uso del cuestionario oq-45.2 como indicador de psicopatología y de mejoría en pacientes psiquiátricos hospitalizados. *Revista chilena de neuro-psiquiatría*, 44(4):258–262, 2006.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[16] Organización Panamericana de la Salud. Mortalidad por suicidio en las Américas. 2010.

[17] Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple classifier systems*, pages 1–15. Springer, 2000.

[18] Thomas G Dietterich. Ensemble learning. *The handbook of brain theory and neural networks*, 2:110–125, 2002.

[19] Gabriele B Durrant. Imputation methods for handling item-nonresponse in the social sciences: a methodological review. *ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute. NCRM Methods Review Papers NCRM/002*, 2005.

[20] Ramón Florenzano, Juan Francisco Labra, Roberto Fasani, Katherine San Juan, Josefina Reynal, and Y Quevedo. Los pacientes suicidas y para-suicidas pueden ser adecuadamente diagnosticados y tratados en una red pública de atención en salud mental. *Revista Gaceta Psiquiatría Universitaria*, 3:331–9, 2007.

[21] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995.

[22] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156, 1996.

[23] Marco Galván. *Imputación de datos: teoría y práctica*, volume 54. United Nations Publications, 2007.

[24] Cristina García and Irene Gómez Moreno. Algoritmos de aprendizaje: knn & kmeans. Technical report.

[25] Julio Bobes García, Giner Ubago Giner, and Jerónimo Saiz Ruiz. *Suicidio y psiquiatría: recomendaciones preventivas y de manejo del comportamiento suicida*. Triacastela, 2011.

[26] Sara K Goldsmith, Terry C Pellmar, Arthur M Kleinman, and William E Bunney. *Reducing suicide: A national imperative.* National Academies Press, 2002.

[27] Steve R Gunn. Support vector machines for classification and regression. Technical Report May, 1998.

[28] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

[29] E Clare Harris and Brian Barraclough. Suicide as an outcome for mental disorders. a meta-analysis. *The British Journal of Psychiatry*, 170(3):205–228, 1997.

[30] Lisa M Horowitz, Jeffrey A Bridge, Stephen J Teach, Elizabeth Ballard, Jennifer Klima, Donald L Rosenstein, Elizabeth A Wharff, Katherine Ginnis, Elizabeth Cannon, and Paramjit Joshi. Ask suicide-screening questions (asq): a brief instrument for the pediatric emergency department. *Archives of pediatrics & adolescent medicine*, 166(12):1170–1176, 2012.

[31] ET Isomets et al. Suicide attempts preceding completed suicide. *The British Journal of Psychiatry*, 173(6):531–535, 1998.

[32] Graham Kalton and Daniel Kasprzyk. Imputing for missing survey responses. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, volume 22. American Statistical Association Cincinnati, 1982.

[33] Harold Kaplan, Benjamín J Sadock, and Jack A Grebb. *Sinopsis de psiquiatría*. Medica, 1999.

[34] Ron Kohavi and Foster Provost. Glossary of terms. *Machine Learning*, 30(2-3):271–274, 1998.

[35] Peter Sollich Anders Krogh. Learning with ensembles: How over-fitting can be useful. 1996.

[36] Max Kuhn. Algoritmo rfe del paquete `caret` en `R`[en línea]. <`http://topepo.github.io/caret/rfe.html`>. [Consultado: 2015-07-01].

[37] RJA Little and DB Rubin. Statistical analysis with missing data. 1987.

[38] William G Madow, Harold Nisselson, Ingram Olkin, and Donald Rubin. Incomplete data in sample surveys (vols. 1-31), 1985.

[39] R Maris, A Berman, and MM Silverman. The theoretical component in suicidology. *Comprehensive textbook of suicidology*, pages 26–61, 2000.

[40] Gobierno de Chile Ministerio de Salud. Panorama de Salud 2013: Informe OECD sobre Chile y comparación con países miembros. 2013.

[41] Gobierno de Chile Ministerio de Salud. Situación actual del suicidio adolescente en Chile, con perspectiva de género. 2013.

[42] OECD. Suicides. In *OECD Factbook 2014: Economic, Environmental and Social Statis-*

tics, pages 240–241. 2014.

[43] MA Oquendo, E Baca-Garcia, A Artes-Rodriguez, F Perez-Cruz, HC Galfalvy, H Blasco-Fontecilla, D Madigan, and N Duan. Machine learning and data mining: strategies for hypothesis generation. *Molecular psychiatry*, 17(10):956–959, 2012.

[44] Robert E Roberts, Catherine Ramsay Roberts, and Yun Xing. One-year incidence of suicide attempts and associated risk and protective factors among adolescents. *Archives of suicide research : official journal of the International Academy for Suicide Research*, 14(1):66–78, 2010.

[45] Donald B Rubin. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, 72(359):538–543, 1977.

[46] Christopher Ryan, Olav Nielssen, Michael Paton, and Matthew Large. Clinical decisions in psychiatry should not be based on risk assessment. *Australasian Psychiatry*, 18(5):398–403, October 2010.

[47] Lilian Salvo, Roberto Melipillán, and Andrea Castro. Confiabilidad, validez y punto de corte para escala de screening de suicidalidad en adolescentes. *Revista chilena de neuro-psiquiatría*, 47(1):16–23, 2009.

[48] Robert E Schapire. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer, 2003.

[49] Sonia Singh. Comparative study id3 , cart and c4 . 5 decision tree algorithm : A survey. *International Journal of Advanced Information Science and Technology (IJAIST)*, 27(27):97–103, 2014.

[50] Truyen Tran, Dinh Phung, Wei Luo, Richard Harvey, Michael Berk, and Svetha Venkatesh. An integrated framework for suicide risk prediction. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, page 1410, New York, New York, USA, 2013. ACM Press.

[51] World Health Organization. Preventing suicide: A Global Imperative. 2014.

Anexos

Apéndice A

Anexo A: Instrumentos clínicos.

ESCALA DE RIESGO-RESCATE (Weisman & Worden, 1972))

Nombre	Fecha
Circunstancias	
Puntuación Riesgo	Puntuación Rescate
Relación Riesgo/Rescate	

FACTORES DE RIESGO	**FACTORES DE RESCATE**
1. Método utilizado	**1. Lugar**
1. Ingestión, cortes, apuñalamiento	3. Familiar
2. Ahogamiento, asfixia, estrangulamiento	2. No familiar, no lejano
3. Precipitación, disparo	1. Lejano
2. Alteración de la conciencia	**2 Persona que inicia el rescate**
1. No evidencia	3. Persona "clave"
2. Confusión, semi coma	2. Profesional
3. Coma, coma profundo	1. Transeúnte
3. Lesiones/Toxicidad	**3. Probabilidad de ser descubierto por un "salvador"**
1. Leve	3. Alta, casi segura
2. Moderada	2. Descubrimiento incierto
3. Severa	1. Descubrimiento accidental
4. Reversibilidad	**4. Facilitación del rescate**
1. Buena	3. Pide ayuda
2. Favorable. Expectativa de recuperación con el tiempo	2. Deja "pistas"
3. Mala. Expectativa de secuelas, si se recupera	1. No pide ayuda
5. Tratamiento requerido	**5. Demora hasta el descubrimiento**
1. Primeros auxilios en Urgencias	3. Inmediatamente 1 hora
3. Cuidados Intensivos, tratamientos especiales	1. Más de tres horas

* El **auto rescate** determina automáticamente una puntuación de rescate de 5

* Si hay excesivo tiempo de demora en obtener tratamiento después del descubrimiento, reducir la puntuación de rescate 1 punto

* Si utiliza varios métodos puntuar el más letal.

FACTORES DE RIESGO

5. Alto riesgo (13-15 puntos riesgo)

4. Alto moderado (11-12 puntos riesgo)

3. Moderado (9-10 puntos riesgo)

2. Bajo moderado (7-8 puntos riesgo)

1. Bajo riesgo (5-6 puntos riesgo)

FACTORES DE RESCATE

1. El menos rescatable (5-7 puntos rescate)

2. Bajo moderado (8-9 puntos rescate)

3. Moderado (10-11 puntos rescate)

4. Moderado alto (12-13 puntos rescate)

5. Más rescatable (14-15 puntos rescate)

COMPUTACION DE LAS PUNTUACIONES RIESGO/RESCATE

Puntuación Riesgo Puntuación Rescate Puntuación Riesgo-Rescate

Puntuación riesgo	Puntuación rescate	Puntuación Riesgo Rescate
1	5	17
1	4	20
1	3	25
1	2	33
1	1	50
2	5	29
2	4	33
2	3	40
2	2	50
2	1	60
3	5	38
3	4	43
3	3	50
3	2	60
3	1	75
4	5	44
4	4	50
4	3	57
4	2	66
4	1	80
5	5	50
5	4	56
5	3	63
5	2	71
5	1	83

ESCALA DE INTENCION SUICIDA DE PIERCE

CIRCUNSTANCIAS RELACIONADAS AL INTENTO SUICIDA.

1.- Aislamiento.

0 Alguien presente
1 Alguien cerca o en contacto.
2 Nadie cerca o en contacto.

2.- Momento.

0 Escogido de tal manera que la intervención es probable.
1 Escogido de tal manera que la intervención no es probable.
2 Escogido de tal manera que la intervención es altamente improbable.

3.- Precauciones contra el descubrimiento y o intervención.

0 No tomo precauciones
1 Precauciones pasivas, evitación de otros, pero sin hacer nada para prevenir su intervención (estar sólo en su pieza, puertas sin llave).
2 Activas precauciones como el poner llave a la puerta.

4.- Actuación para obtener ayuda durante o después del intento.

0 Notificó a auxiliador potencial respecto al intento.
1 Contactó pero específicamente no notificó al auxiliador potencial respecto al intento.
2 No contacto ni notificó a auxiliador potencial.

5.- Actos finales en anticipación de la muerte.

0 Ninguno
1 Preparación parcial o ideación.
2 Planes definidos realizados (e.g. cambios en un testamento, tomar un seguro)

6.- Nota suicida.

0 Ausencia de nota.
1 Nota escrita pero destruida.
2 Presencia de Nota.

AUTOINFORME.

1.- Calificación de letalidad por parte del paciente.

0 Pensó que lo efectuado no le provocaría la muerte.
1 Inseguro acerca de que lo efectuado le provocaría la muerte.
2 Creyó que lo efectuado le provocaría la muerte.

2.- Intento establecido.

0 No quiso morir
1 Incierto no se preocupó de si viviría o moriría.
2 Quiso morir.

3.- Premeditación.

0 impulsivo, consideró acto por menos de una hora
1 Consideró acto por más de una hora
2 Consideró acto por menos de un día
3 Consideró acto por más de un día

4.- Reacción frente al acto

0 Paciente contento de haberse recuperado
1 Paciente inseguro de estar contento o descontento
2 Paciente descontento de haberse recuperado

RIESGO

1,.-Resultado predecible en términos de la letalidad del acto del paciente y de las circunstancias.

0 Sobreviva segura
1 Muerte improbable
2 Muerte probable o segura

2.- Habría ocurrido la muerte sin tratamiento médico.

0 No
1 Incierto
2 Sí

RESULTADOS:

0 – 3 Baja intención suicida.
4 – 10 Mediana intención suicida
10 o más Alta intención suicida.

Agradecemos que conteste las siguientes preguntas

Cuestionario de resultados OQ 45.2

Edad:_____ **Sexo:** M ☐ F ☐

Nombre:_____

Nº Ficha:_____ **Sesión Nº**_____ **Fono:**_____ **Fecha:**_____

Instrucciones: Para ayudarnos a entender como se ha estado sintiendo, básese en los **ÚLTIMOS SIETE DIAS**, incluyendo el día de hoy. Lea cuidadosamente las frases y seleccione la categoría que mejor describa como se siente esta semana. En el cuestionario el término "**TRABAJO**" se refiere al empleo, la escuela, el trabajo voluntario, ser dueña de casa, cuidar los niños,etc. Por favor no escriba en las áreas oscuras. Marque con una "**X**" en el cuadro que corresponda.

	Nunca	Casi nunca	A veces	Con frecuencia	Casi siempre	SD	IR	SR
1. Me llevo bien con otros	4	3	2	1	0		*	
2. Me canso rápidamente.	0	1	2	3	4			
3. Nada me interesa	0	1	2	3	4			
4. Me siento presionado (estresado) en el trabajo/escuela/dueña de casa	0	1	2	3	4			
5. Me siento culpable.	0	1	2	3	4			
6. Me siento irritado, molesto.	0	1	2	3	4			
7. Me siento contento con mi matrimonio/pareja.	4	3	2	1	0		*	
8. Pienso en quitarme la vida.	0	1	2	3	4			
9. Me siento débil.	0	1	2	3	4			
10. Me siento atemorizado.	0	1	2	3	4			
11.Necesito tomar bebidas alcohólicas en la mañana, después de haber tomado el día anterior. (Si esto no le ocurre marque nunca).	0	1	2	3	4			
12. Encuentro satisfacción en mi trabajo/ escuela/dueña de casa.	4	3	2	1	0			*
13. Soy una persona feliz.	4	3	2	1	0	*		
14. Trabajo/estudio/dueña de casa, excesivamente (mas de la cuenta).	0	1	2	3	4			
15. Me siento inútil.	0	1	2	3	4			
16. Me abruman (angustian) los problemas de mi familia.	0	1	2	3	4			
17. Mi vida sexual me llena.	4	3	2	1	0		*	
18. Me siento solo.	0	1	2	3	4			
19. Discuto frecuentemente.	0	1	2	3	4			
20. Me siento querido y que me necesitan.	4	3	2	1	0		*	
21. Disfruto mi tiempo libre.	4	3	2	1	0			*
22. Tengo dificultades para concentrarme.	0	1	2	3	4			

	Nunca	Casi nunca	A veces	Con frecuencia	Casi siempre	SD	IR	SR
23. Me siento sin esperanza en el futuro.	0	1	2	3	4			
24. Estoy contento conmigo mismo.	4	3	2	1	0	*		
25. Me perturban o molestan pensamientos de los que no me puedo deshacer.	0	1	2	3	4			
26. Me molesta que me critiquen porque tomo o me drogo. (No se refiere a medicamentos recetados). (Si esto no le ocurre marque nunca)	0	1	2	3	4			
27. Tengo malestares estomacales.	0	1	2	3	4			
28. Trabajo/estudio/dueña de casa, tan bien como lo hacía antes.	4	3	2	1	0			*
29. Mi corazón palpita demasiado.	0	1	2	3	4			
30. Tengo dificultades para llevarme bien con mis amigos y conocidos.	0	1	2	3	4			
31. Estoy satisfecho con mi vida.	4	3	2	1	0	*		
32. Tengo problemas en el trabajo/escuela debido a las drogas o el alcohol. (Si esto no le ocurre marque nunca).	0	1	2	3	4			
33. Siento que algo malo va a ocurrir.	0	1	2	3	4			
34. Tengo los músculos adoloridos.	0	1	2	3	4			
35. Me atemorizan los espacios abiertos, el manejar, el estar dentro de un bus, el metro, ascensores, etcétera.	0	1	2	3	4			
36. Me siento nervioso.	0	1	2	3	4			
37. Me satisfacen mis relaciones con mis seres queridos.	4	3	2	1	0		*	
38. Siento que me va bien en el trabajo/escuela/dueña de casa.	4	3	2	1	0			*
39. Tengo muchas discusiones en el trabajo/escuela/dueña de casa.	0	1	2	3	4			
40. Siento que algo anda mal con mi mente.	0	1	2	3	4			
41. Tengo dificultades para dormir, o no me puedo quedar dormido.	0	1	2	3	4			
42. Me siento triste.	0	1	2	3	4			
43. Mis relaciones con otros me satisfacen.	4	3	2	1	0		*	
44. Me enojo tanto en el trabajo/escuela/casa, que puedo hacer algo de lo que después me puedo arrepentir.	0	1	2	3	4			
45. Me dan dolores de cabeza.	0	1	2	3	4			
						+	+	
						Total=		

Developed by Michael Lambert, Ph.D. and Gary M. Burlingame, Ph.D.
Para mayor información contactar: Guillermo de la Parra C.
E-MAIL: gdelap@net.vtr
Alejandra von Bergen R.
E-MAIL: avonbergen@terra.cl

Validado en Castellano por Guillermo de la Parra C. y Alejandra von Bergen R.

Nombre:

Cuestionario de funcionamiento familiar

(Smilkstein, 1978)

Por favor indique con un X el espacio que refleje mejor la frecuencia con que está de acuerdo con las siguientes afirmaciones respecto a su familia

	Nunca 0	Algunas veces 1	Siempre 2
Me satisface la ayuda que recibo de mi familia cuando tengo algún problema y/o necesidad			
Me satisface la forma como mi familia habla de las cosas y comparte los problemas conmigo			
Me satisface como mi familia acepta y apoya mis deseos de emprender nuevas actividades			
Me satisface como mi familia expresa afecto y responde a mis emociones como rabia, tristeza o amor			
Me satisface cómo compartimos en familia el tiempo de estar juntos, los espacios en la casa o el dinero			

CUESTIONARIO DE EXPERIENCIAS DEPRESIVAS
(DEQ)

Nombre:

Sidney J. Blatt, Ph.D.; Carrie E. Schaffer, Ph.D.; Susan A. Bers, Ph.D.; Donald M. Quinlan, Ph.D., 1989.

Traducción y Adaptación al Idioma Español

Por Ps. Susana Morales Silva. Doctora en Psicoterapia P. Universidad Católica de Chile-U de Chile sumorales@med.puc.cl

Tomado de la primera adaptación al español de la escala para adolescentes, de Humberto L. Persano MD, Ph.D., 2003., Mental Health Department, School of Medicine, University of Buenos Aires.

Se le solicita que lea cada una de las siguientes frases y decide que tan bien te describen. Luego haz un círculo alrededor del número más apropiado para cada ítem mencionado, basándote en la escala que se presenta a continuación

En total desacuerdo 1	**2**	**3**	**4**	**5**	**6**	**En total acuerdo 7**

1. Establezco mis objetivos a un nivel muy alto.

1	2	3	4	5	6	7

2. Sin el apoyo de los que están cerca de mí me encontraría desamparado (a).

1	2	3	4	5	6	7

3. En general, me siento más conforme con mis planes y metas, que intentando alcanzar objetivos más elevados.

1	2	3	4	5	6	7

4. A veces me siento muy grande y en otras ocasiones muy pequeño (a).

1	2	3	4	5	6	7

5. Cuando estoy involucrado con alguien, nunca me siento celoso (a).

1	2	3	4	5	6	7

6. Realmente necesito cosas que sólo otras personas me pueden dar.

1	2	3	4	5	6	7

7. A menudo me parece que no alcanzo los ideales que me he propuesto.

1	2	3	4	5	6	7

8. Considero que siempre utilizo al máximo mis habilidades.

1	2	3	4	5	6	7

En total desacuerdo 1	**2**	**3**	**4**	**5**	**6**	**En total acuerdo 7**

9. Me molesta que cambien las relaciones entre las personas.

1	2	3	4	5	6	7

10. Si fracaso en el intento de lograr mis expectativas, me siento desvalorizado (a).

1	2	3	4	5	6	7

11. Muchas veces me siento desamparado.

1	2	3	4	5	6	7

12. Rara vez me preocupa que me critiquen por cosas que he dicho o he realizado.

1	2	3	4	5	6	7

13. Existe una gran diferencia entre como soy ahora y cómo desearía ser.

1	2	3	4	5	6	7

14. Disfruto al competir con otros.

1	2	3	4	5	6	7

15. Siento que debo enfrentar muchas responsabilidades.

1	2	3	4	5	6	7

16. A veces me siento vacío (a) por dentro.

1	2	3	4	5	6	7

17. Usualmente no me siento satisfecho (a) con lo que tengo.

1	2	3	4	5	6	7

18. Me importa alcanzar lo que los demás esperan de mí.

1	2	3	4	5	6	7

19. Me vuelvo atemorizado (a) cuando me siento solo (a).

1	2	3	4	5	6	7

En total desacuerdo 1	**2**	**3**	**4**	**5**	**6**	**En total acuerdo 7**

20. Si perdiera a un amigo (a) muy cercano (a), sería como si perdiese una parte muy importante de mí mismo (a).

1	2	3	4	5	6	7

21. La gente me aceptará sin importar cuántos errores haya cometido.

1	2	3	4	5	6	7

22. Me resulta difícil cortar con una amistad que me esté haciendo infeliz.

1	2	3	4	5	6	7

23. Frecuentemente pienso acerca del peligro de perder alguien cercano a mí.

1	2	3	4	5	6	7

24. La gente tiene altas expectativas acerca de mí.

1	2	3	4	5	6	7

25. Cuando estoy con otros, tiendo a devaluarme o "subvenderme".

1	2	3	4	5	6	7

26. Me preocupa mucho la manera en que la gente reacciona ante mí.

1	2	3	4	5	6	7

27. Siempre hay una cuota de incertidumbre y conflicto, aunque dos personas estén muy unidas.

1	2	3	4	5	6	7

28. Soy muy sensitivo (a) a los signos de rechazo de los otros hacia mí.

1	2	3	4	5	6	7

29. Es importante para mi familia que yo tenga éxito.

1	2	3	4	5	6	7

30. A menudo siento que he decepcionado a los demás.

1	2	3	4	5	6	7

En total desacuerdo 1	**2**	**3**	**4**	**5**	**6**	**En total acuerdo 7**

31. Si alguien me hace enojar, le hago saber cómo me siento.

1	2	3	4	5	6	7

32. Muy frecuentemente me desvivo por agradar o ayudar a las personas cercanas a mí.

1	2	3	4	5	6	7

33. Tengo muchos recursos personales (fortaleza interior, habilidades).

1	2	3	4	5	6	7

34. Encuentro muy difícil decir "No" a las peticiones de mis amigos (as).

1	2	3	4	5	6	7

35. Nunca me siento del todo seguro (a) en una relación cercana.

1	2	3	4	5	6	7

36. La manera en que me siento frecuentemente varía: A veces me siento extremadamente bien conmigo mismo y en otras ocasiones me siento totalmente fracasado (a).

1	2	3	4	5	6	7

37. A menudo me siento amenazado por el cambio de las cosas y situaciones.

1	2	3	4	5	6	7

38. Aunque la persona más cercana a mí se fuera, yo podría continuar solo (a).

1	2	3	4	5	6	7

39. La gente siempre tiene que esforzarse para ganarse el amor de los otros. Es decir: el amor debe ser ganado.

1	2	3	4	5	6	7

40. Soy muy sensitivo (a) a los efectos que tienen mis palabras y mis acciones , en los sentimientos de las demás personas.

1	2	3	4	5	6	7

41. A menudo me culpo por las cosas que he hecho o dicho a alguien.

1	2	3	4	5	6	7

En total desacuerdo	**2**	**3**	**4**	**5**	**6**	**En total acuerdo 7**

42. Soy una persona muy independiente.

1	2	3	4	5	6	7

43. A menudo me siento culpable.

1	2	3	4	5	6	7

44. Pienso que soy una persona muy compleja que tiene "muchas facetas".

1	2	3	4	5	6	7

45. Me preocupa mucho la posibilidad de herir o lastimar a alguien que sea cercano a mí.

1	2	3	4	5	6	7

46. La ira me asusta.

1	2	3	4	5	6	7

47. Lo que importa es el “rendimiento” o “cumplimiento” (calificaciones, logros), mucho más que “quien uno es”

1	2	3	4	5	6	7

48. Me siento bien conmigo mismo (a), ya sea que triunfe o que fracase.

1	2	3	4	5	6	7

49. Puedo dejar de lado mis propios problemas y sentimientos con facilidad y dedicar mi atención completa a los sentimientos y problemas de otros.

1	2	3	4	5	6	7

50. Si alguien que me importa se enojara conmigo, me asustaría la posibilidad de que esa persona pudieran abandonarme.

1	2	3	4	5	6	7

51. Me siento cómodo (a) cuando me dan responsabilidades importantes.

1	2	3	4	5	6	7

52. Luego de una pelea con un amigo (a), me preocupo de arreglarlo lo antes posible.

1	2	3	4	5	6	7

En total desacuerdo 1	**2**	**3**	**4**	**5**	**6**	**En total acuerdo 7**

53. Me resulta difícil aceptar mis propias debilidades.

1	2	3	4	5	6	7

54. Es más importante que yo disfrute de mi trabajo (tareas escolares), a que otros me alaben por ello.

1	2	3	4	5	6	7

55. Luego de una discusión me siento muy solo (a).

1	2	3	4	5	6	7

56. En mis relaciones con los demás, me preocupo mucho por lo que ellos me puedan darme.

1	2	3	4	5	6	7

57. Raramente pienso en mi familia.

1	2	3	4	5	6	7

58. Muy frecuentemente, los sentimientos hacia una persona muy cercana, varían: Hay momentos en los que me siento completamente enojado y otros en los que quiero muchísimo a esa persona.

1	2	3	4	5	6	7

59. Lo que digo y hago tiene un fuerte impacto en aquellos que me rodean.

1	2	3	4	5	6	7

60. A veces siento que soy "especial".

1	2	3	4	5	6	7

61. Crecí en una familia extremadamente unida.

1	2	3	4	5	6	7

62. Estoy muy satisfecho (a) conmigo mismo (a) y con las cosas que he logrado.

1	2	3	4	5	6	7

63. Necesito muchas cosas de los que están cerca de mí.

1	2	3	4	5	6	7

En total desacuerdo 1	**2**	**3**	**4**	**5**	**6**	**En total acuerdo 7**

64. Tiendo a ser muy crítico (a) conmigo mismo.

1	2	3	4	5	6	7

65. Estar solo (a) me agrada mucho.

1	2	3	4	5	6	7

66. Muy frecuentemente me comparo con las metas y parámetros que me propuse.

1	2	3	4	5	6	7

Nombre:

Spielberger, C.D. (1996). State–Trait Anger Expression Inventory (STAXI 2). Odessa, FL: Psychological Assessment Resources.

Traducción y adaptación al idioma español por Ps. Susana Morales Silva
Doctora en Psicoterapia, P. Universidad Católica de Chile - Universidad de Chile
sumorales@med.puc.cl

Parte 1 Instrucciones

A continuación, hay una serie de afirmaciones que las personas utilizan para describirse a sí mismas. Se le solicita que lea cada una y luego marque con un círculo el número que lo (a) representa a usted ahora. Recuerde que no existen respuestas correctas o incorrectas. No destine demasiado tiempo en una afirmación y por favor responda con aquella afirmación que más se ajusta a su sentimiento actual

EN TOTAL DESACUERDO 1	MEDIANAMENTE DE ACUERDO 2	MUY DE ACUERDO 3

COMO ME SIENTO AHORA

1. Estoy furioso (a)

1	2	3

2. Me siento cohibido (a)

1	2	3

3. Me siento enojado (a)

1	2	3

4. Me siento como si estuviera gritándole a alguien

1	2	3

5. Me siento como si estuviera rompiendo cosas

1	2	3

6. Yo soy enojón (a)

1	2	3

7. Me siento como que quisiera golpear una mesa

1	2	3

8. Me siento como que quisiera golpear a alguien

1	2	3

9. Me siento enrabiado (a)

1	2	3

10. Me siento como que quisiera maldecir

1	2	3

Parte 2 Instrucciones

A continuación, hay una serie de afirmaciones que las personas utilizan para describirse a sí mismas. Se le solicita que lea cada una y luego marque con un círculo el número que indica cuál lo (a) representa como usted generalmente se siente. Recuerde que no existen respuestas correctas o incorrectas. No destine demasiado tiempo en una afirmación y por favor responda con aquella afirmación que más se ajusta a su sentimiento en general.

EN TOTAL DESACUERDO 1	MEDIANAMENTE DE ACUERDO 2	MUY DE ACUERDO 3

CÓMO ME SIENTO GENERALMENTE

11. Soy temperamental		
1	2	3
12. Tengo un temperamento explosivo		
1	2	3
13. Soy una persona enojona		
1	2	3
14. Me enoja cuando me retrasan los errores de otros		
1	2	3
15. Me siento molesto (a) cuando no soy reconocido por un trabajo bien hecho		
1	2	3
16. Pierdo los estribos generalmente		
1	2	3
17. Cuando me enojo digo groserías		
1	2	3
18. Me pone furioso (a) ser criticado (a) frente a otros		
1	2	3
19. Cuando me frustro, me siento como si quisiera golpear a alguien		
1	2	3
20. Me pongo furioso (a) cuando hago un buen trabajo y obtengo una pobre evaluación		
1	2	3

Parte 3 Instrucciones

Todas las personas se enojan o se ponen furiosas de vez en cuando. Sin embargo, la gente difiere en la forma en cómo reaccionan cuando están enojados. Se señalan a continuación, una serie de afirmaciones que describe las reacciones que usualmente tienen las personas cuando se enojan ó cuando se ponen furiosos. Se le solicita que lea cada afirmación y luego marque con un círculo, el número que mejor indica la forma en que a menudo reacciona cuando se enoja. Recuerde que no hay respuestas correctas o incorrectas. No destine demasiado tiempo en ninguna de las afirmaciones.

EN TOTAL DESACUERDO 1	**MEDIANAMENTE DE ACUERDO 2**	**MUY DE ACUERDO 3**

CUANDO ME ENOJO O ME PONGO FURIOSO (A)

21. Controlo mi temperamento

1	2	3

22. Expreso mi enojo

1	2	3

23. Me guardo las cosas en mi interior

1	2	3

24. Soy paciente con los demás

1	2	3

25. Pongo mala cara o de mal humor

1	2	3

26. Me alejo de la gente

1	2	3

27. Hago comentarios sarcásticos

1	2	3

28. Me mantengo controlado (a)

1	2	3

29. Hago cosas como dar portazos

1	2	3

30. Hiervo por dentro, pero no lo demuestro

1	2	3

31. Controlo mi conducta

1	2	3

32. Soy peleador (a) con los demás

1	2	3

EN TOTAL DESACUERDO 1	MEDIANAMENTE DE ACUERDO 2	MUY DE ACUERDO 3

33 Tiendo a guardar rencores que no le cuento a los demás

1	2	3

34. Elimino (borro) todo lo que me pone furioso (a)

1	2	3

35. No puedo detenerme, pierdo la paciencia

1	2	3

36. Soy íntimamente bastante crítico (a) con los demás

1	2	3

37. Soy más rabioso (a) de lo que estoy generalmente dispuesto (a) a admitir

1	2	3

38. Me calmo más rápidamente que la mayoría de la gente

1	2	3

39. Digo groserías

1	2	3

40. Intento ser tolerante y comprensivo (a)

1	2	3

41. Me irrito mucho más de lo que las personas imaginan que lo hago

1	2	3

42. Pierdo la paciencia

1	2	3

43. Si alguien me molesta, estoy preparado (a) para decirle como me siento

1	2	3

44. Controlo mis sentimientos de rabia

1	2	3

Spielberg, C.D. (1996).

PARENTAL BONDING INSTRUMENT PBI

INSTRUCCIONES

Este cuestionario consta de 25 afirmaciones, cada una de las cuales se refiere a cómo recuerda usted a su MADRE (*) en su infancia (hasta sus 16 años).

Cada afirmación es seguida por una escala de puntaje:

Muy en desacuerdo
Moderadamente en desacuerdo
Moderadamente en acuerdo
Muy en acuerdo

Evalúe el grado en que Usted está en acuerdo o en desacuerdo con cada afirmación y marque con una cruz la celdilla indicada.

Por favor conteste en relación a los recuerdos que tiene de su **MADRE (*)**

	Muy en desacuerdo	Moderadamente en desacuerdo	Moderadamente en acuerdo	Muy en acuerdo	C	S
1. Me hablaba con voz amistosa y cálida	0	1	2	3		
2. No me ayudaba tanto como yo lo necesitaba	0	1	2	3		
3. Evitaba que yo saliera solo (a)	3	2	1	0		
4. Parecía emocionalmente fría hacia mi	3	2	1	0		
5. Parecía entender mis problemas y preocupaciones	0	1	2	3		
6. Era afectuosa conmigo	0	1	2	3		
7. Le gustaba que tomara mis propias decisiones	3	2	1	0		
8. No quería que creciera	0	1	2	3		
9. Trataba de controlar todo lo que yo hacía	0	1	2	3		
10. Invadía mi privacidad	0	1	2	3		
11. Se entretenía conversando cosas conmigo	0	1	2	3		
12. Me sonreía frecuentemente	0	1	2	3		
13. Me regaloneaba	0	1	2	3		
14. No parecía entender lo que yo quería o necesitaba	3	2	1	0		
15. Me permitía decidir las cosas por mi mismo (a)	3	2	1	0		
16. Me hacía sentir que no era deseado (a)	3	2	1	0		

17. Tenía la capacidad de reconfortarme cuando me sentía molesto (a) o perturbado (a)	0	1	2	3		
18. No conversaba mucho conmigo	3	2	1	0		
19. Trataba de hacerme dependiente de ella	0	1	2	3		
20. Sentía que no podía cuidar de mi mismo (a), a menos que ella estuviera cerca	0	1	2	3		
21. Me daba toda la libertad que yo quería	3	2	1	0		
22. Me dejaba salir todo lo que yo quería	3	2	1	0		
23. Era sobreprotectora conmigo	0	1	2	3		
24. No me elogiaba	3	2	1	0		
25. Me permitía vestirme como se me antojara	3	2	1	0		
Total						

(*) Nota: El inventario dirigido al PADRE es exactamente igual a este, pero las instrucciones se refieren al padre.

PARENTAL BONDING INSTRUMENT PBI

INSTRUCCIONES

Este cuestionario consta de 25 afirmaciones, cada una de las cuales se refiere a cómo recuerda usted a su PADRE en su infancia (hasta sus 16 años).

Cada afirmación es seguida por una escala de puntaje:

Muy en desacuerdo
Moderadamente en desacuerdo
Moderadamente en acuerdo
Muy en acuerdo

Evalúe el grado en que Usted está en acuerdo o en desacuerdo con cada afirmación y marque con una cruz la celdilla indicada.

Por favor conteste en relación a los recuerdos que tiene de su **PADRE**

	Muy en desacuerdo	Moderadamente en desacuerdo	Moderadamente en acuerdo	Muy en acuerdo	C	S
1. Me hablaba con voz amistosa y cálida	0	1	2	3		
2. No me ayudaba tanto como yo lo necesitaba	0	1	2	3		
3. Evitaba que yo saliera solo (a)	3	2	1	0		
4. Parecía emocionalmente frío hacia mi	3	2	1	0		
5. Parecía entender mis problemas y preocupaciones	0	1	2	3		
6. Era afectuoso conmigo	0	1	2	3		
7. Le gustaba que tomara mis propias decisiones	3	2	1	0		
8. No quería que creciera	0	1	2	3		
9. Trataba de controlar todo lo que yo hacía	0	1	2	3		
10. Invadía mi privacidad	0	1	2	3		
11. Se entretenía conversando cosas conmigo	0	1	2	3		
12. Me sonreía frecuentemente	0	1	2	3		
13. Me regaloneaba	0	1	2	3		
14. No parecía entender lo que yo quería o necesitaba	3	2	1	0		
15. Me permitía decidir las cosas por mi mismo (a)	3	2	1	0		
16. Me hacía sentir que no era deseado (a)	3	2	1	0		
17. Tenía la capacidad de reconfortarme cuando me sentía molesto (a) o perturbado (a)	0	1	2	3		

18. No conversaba mucho conmigo	3	2	1	0		
19. Trataba de hacerme dependiente de él	0	1	2	3		
20. Sentía que no podía cuidar de mi mismo (a), a menos que él estuviera cerca	0	1	2	3		
21. Me daba toda la libertad que yo quería	3	2	1	0		
22. Me dejaba salir todo lo que yo quería	3	2	1	0		
23. Era sobreprotector conmigo	0	1	2	3		
24. No me elogiaba	3	2	1	0		
25. Me permitía vestirme como se me antojara	3	2	1	0		
Total						

REASONS FOR LIVING

University of Washington
Clínicas de Investigación Conductual y Terapia
Linehan et. al., 1983

NOMBRE:
_____

FECHA: _____

Instrucciones:

Muchas personas han pensado en el suicidio alguna vez y otras nunca lo han considerado. Independiente de que Ud. Lo haya pensado o no, nos interesan las razones que tendría para **no** suicidarse si alguna vez se le ocurriera hacerlo o si alguien se lo sugiriera.

En las siguientes páginas hay razones que dan las personas para **no** suicidarse. Queremos saber cuán importantes son cada una de estas posibles razones para Ud. Para no suicidarse en este momento de su vida. Por favor ponga el número de la importancia en el espacio a la derecha de cada pregunta.

Cada pregunta puede calificarse desde 1 (No es importante) a 6 (Extremadamente importante). Si una razón no se aplica a Ud. O si no piensa que la afirmación sea verdadera, entonces probablemente no es importante para Ud. En este caso, debería poner un 1 (No es importante) como respuesta. Por favor utilice el rango completo de opciones para no calificar solamente desde el medio (2, 3, 4, 5) o solamente desde los extremos (1, 6).

En cada espacio ponga un número para indicar la importancia que tiene para Ud. Cada razón por la cual **no** suicidarse.

1. No es importante
2. Muy poco importante
3. Poco importante
4. Importante
5. Muy importante
6. Extremadamente importante

Incluso si nunca ha considerado el suicidio o cree firmemente que nunca pensaría en el suicidio en forma seria, es importante que califique cada razón. En este caso, la calificación debe ser basada en **por qué suicidarse no es o nunca sería una alternativa para Ud.**

1. No es importante 3. Poco importante 5. Muy importante 2. Muy poco importante 4. Importante 6. Extremadamente importante	
1. Soy responsable y estoy comprometido (a) con mi familia	
2. Pienso que puedo aprender a ajustarme a los problemas o a sobrellevarlos	
3. Pienso que yo controlo mi vida y mi destino	
4. Tengo deseos de vivir	
5. Pienso que sólo Dios tiene el derecho a terminar con una vida	
6. Le tengo miedo a la muerte	
7. Mi familia podría pensar que no los quise	
8. No pienso que las cosas se pongan tan malas o desesperanzadoras como para preferir estar muerto (a)	
9. Mi familia depende de mí y me necesita	
10. No quiero morir	
11. Quiero ver crecer a mis hijos (as)	
12. La vida es lo único que tenemos y es mejor que tener nada	
13. Tengo planes futuros que deseo llevar a cabo	
14. No importa cuán mal me sienta, sé que no durará para siempre	
15. Le tengo miedo a lo desconocido	
16. Quiero y disfruto mucho de mi familia, no la querría dejar	
17. Quiero vivir toda experiencia que la vida me ofrezca y hay muchas experiencias que no he tenido aun	
18. Tengo miedo que el método que elija para suicidarme, falle	
19. Me quiero lo suficiente como para vivir	
20. La vida es demasiado bonita y valiosa como para terminarla	
21. No sería justo dejar a mis niños (as) para que otros (as) los (as) cuidaran	
22. Pienso que puedo encontrar otra solución a mis problemas	
23. Tengo miedo de ir al infierno	
24. Siento amor por la vida	
25. Soy demasiado estable como para matarme	
1.No es importante 3.Poco importante 5.Muy importante 2.Muy poco importante 4.Importante 6.Extremadamente importante	

26. Soy cobarde y no tengo las agallas para hacerlo	
27. Mis creencias religiosas lo prohíben	
28. El efecto que tendría sobre mis hijos sería muy dañino	
29. Siento curiosidad acerca de lo que ocurrirá en el futuro	
30. Le dolería mucho a mi familia y no quiero que sufran	
31. Me preocupa lo que los demás pensarán de mi	
32. Pienso que todo generalmente resulta para mejor al final	
33. No podría decidir dónde, cuándo y cómo hacerlo	
34. Lo considero “malo” moralmente	
35. Todavía tengo muchas cosas que hacer	
36. Tengo la valentía de enfrentar la vida	
37. Estoy feliz y contento (a) con mi vida	
38. Me da miedo el acto mismo de suicidarme (ej. el dolor, la sangre, la violencia)	
39. Pienso que el suicidio no lograría ni solucionaría nada	
40. Tengo la esperanza de que las cosas mejorarán y que el futuro será más feliz	
41. Otras personas pensaría que soy débil y egoísta	
42. Tengo la motivación interior para sobrevivir	
43. No querría que las personas pensaran que no tengo control sobre mi vida	
44. Pienso que puedo encontrar un propósito en mi vida, una razón para vivir	
45. No veo razón para apurar la muerte	
46. Soy tan inepto (a) que mi método probablemente no funcionará	
47. No querría que mi familia se sintiera culpable después	
48. No querría que mi familia pensara que soy egoísta ni cobarde	
49. Tengo una responsabilidad y un compromiso con mis amigos	
50. La idea del suicidio es totalmente incomprensible para mí.	

Muchas gracias por su colaboración