UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola d'Enginyeria de Barcelona Est

TRABAJO FINAL DE GRADO

Grado en Ingeniería Biomédica

DISEÑO DE UN SISTEMA PERSONALIZADO DE NOTIFICACIONES PARA PACIENTES CON RIESGO CARDIOMETABÓLICO A PARTIR DEL ANÁLISIS DE DATOS CLÍNICOS Y ESTILO DE VIDA

Memoria y Anexos

Autor: Ana Molina Soler

Director: Beatriz F. Giraldo Giraldo

Convocatoria: Octubre 2017

RESUMEN

Las Enfermedades No Transmisibles (ENT) son la principal causa de mortalidad en nuestro país y en todo el mundo, y a fin de agrupar sus factores de riesgo surge la denominación de Síndrome Metabólico (SM). Las ENT, podrían reducirse y prevenirse eficazmente con una modificación en el estilo de vida moderno, lo que ha dado lugar, gracias al avance de las nuevas tecnologías, al surgimiento de herramientas para la monitorización del estilo de vida. Surge así PREDIRCAM; una Plataforma Inteligente para la Monitorización, Tratamiento y Prevención Personalizados de la Diabetes Mellitus, el Riesgo CardioMetabólico y la Insuficiencia Renal. Esta plataforma ha sido objeto de un estudio clínico, en el que pacientes con sobrepeso y obesos registran datos de su rutina diaria, así como, son asistidos telemáticamente mediante recomendaciones. De esta forma se dispone de una base de datos en la que se encuentran todos los registros de nutrición y ejercicio.

En este contexto, este proyecto consiste en llevar a cabo un proceso de extracción de conocimiento de la base de datos, a fin de realizar varios modelos predictivos, y determinar las variables más significativas en el transcurso del tratamiento de pérdida de peso. Finalmente, a partir de estas, se han definido un conjunto de reglas que se incorporan en el diseño de un sistema personalizado de notificaciones y se adaptan a la evolución y al cumplimiento de la prescripción de dieta y ejercicio físico.

Palabras clave: Síndrome metabólico, Predircam, minería de datos, sistema de recomendación.

RESUM

Les Malalties No Transmissibles (ENT), són la principal causa de mortalitat al nostre país i a tot el món, i a fi d'agrupar els seus factors de risc sorgeix la denominació de Síndrome Metabòlica (SM). Les ENT, podrien reduir-se i prevenir-se eficaçment amb una modificació en l'estil de vida modern, la qual cosa ha donat lloc, gràcies a l'avanç de les noves tecnologies, al sorgiment d'eines per al monitoratge de l'estil de vida. Sorgeix així PREDIRCAM; una Plataforma Intel·ligent per al Monitoratge, Tractament i Prevenció Personalitzats de la Diabetis Mellitus, el Risc CardioMetabólico i la Insuficiència Renal. Aquesta plataforma ha estat objecte d'un estudi clínic, en el qual pacients amb sobrepès i obesos registren dades de la seva rutina diària, així com són assistits telemàticament mitjançant recomanacions. D'aquesta forma es disposa d'una base de dades en la qual es troben tots els registres de nutrició i exercici.

En aquest context, aquest projecte consisteix a dur a terme un procés d'extracció de coneixement de la base de dades, a fi de realitzar diversos models predictius, i determinar les variables més significatives en el transcurs del tractament de pèrdua de pes. Finalment, a partir d'aquestes, s'han definit un conjunt de regles que s'incorporen en el disseny d'un sistema personalitzat de notificacions i s'adapten a l'evolució i al compliment de la prescripció de dieta i exercici físic.

Paraules clau: Síndrome metabòlica, Predircam, mineria de dades, sistema de recomanació.

ABSTRACT

Non transmitable Diseases (NTDs) are the main cause of mortality in our country and around the world, and in order to group their risk factors they are classified as a Metabolic Syndrome (MS). NTDs could be effectively reduced and prevented by a change in the modern lifestyle, which has given rise to the emergence of tools to monitor lifestyle, thanks to the advancement of new technologies. Thus arises PREDIRCAM; an Intelligent Platform for Personalized Monitoring, Treatment and Prevention of Diabetes Mellitus, CardioMetabolic Risk and Renal Insufficiency. This platform has been the subject of a clinical study, in which overweight and obese patients record data from their daily routine, as well as being assisted by telematics through recommendations. In this way a database became available in which all the records of nutrition and exercise are found.

In this context, this project consists in carrying out a process of extracting knowledge from the database, in order to perform several predictive models, and to determine the most significant variables in the course of the treatment of weight loss. Finally, from these, a set of rules have been defined that are incorporated in the design of a personalized system of notifications; and are adapted to the evolution and the accomplisment of the diet prescription and physical exercise.

Key words: Metabolic syndrome, Predircam, data mining, system of recommendation.

AGRADECIMIENTOS

Después de unos años intensos, este trabajo supone el fin de una importante etapa personal. Ha sido un período de aprendizaje constante, tanto en el campo científico como a nivel personal, y a pesar de que estoy segura que este camino es el principio de otros, no quisiera desaprovechar la oportunidad de mencionar a todos los que de alguna manera han sido partícipes de este proyecto:

A mi padres y hermanas que han confiado en mí de principio a fin, me han ayudado, apoyado y han estado ahí siempre que lo he necesitado.

A mis tíos y primos de Barcelona, que me han hecho sentir mejor que en casa, en una ciudad completamente nueva.

A mi chico, por estar y por esas llamadas en situaciones claves.

A Lexa Nescolarde, una excelente profesora y mejor persona, gracias a ella he consolidado conceptos complejos de manera sencilla, y ha sido un gran apoyo en momentos difíciles de mi etapa universitaria.

También quiero agradecer toda la ayuda que me ha ofrecido el Grupo de Bioingeniería y Telemedicina (GBT) de la Universidad Politécnica de Madrid, en especial a José Iniesta y José Tapia, que me han acompañado y aconsejado a lo largo de todo el proyecto, y siempre que lo he necesitado han buscado un hueco para escucharme.

Finalmente, a mi tutora Beatriz Giraldo que me ha ayudado y orientado en este trabajo.

Sin ellos este proyecto no hubiese tenido ni principio ni fin, por eso,

A todos vosotros, gracias.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN Y OBJETIVOS

1.1 Introducción

Las Enfermedades No Transmisibles (ENT) constituyen la principal causa de muerte en nuestro país y en el mundo. Las principales ENT son la diabetes, las enfermedades cardiovasculares, el cáncer, las enfermedades respiratorias crónicas y la enfermedad renal, y se caracterizan por compartir los mismos factores de riesgo: tabaquismo, mala alimentación, falta de actividad física, alcohol excesivo, el sobrepeso y la obesidad.

Con el fin de agrupar todas las patologías y factores de riesgo surge la denominación de Síndrome Metabólico, es un indicador, que constituye un importante catalizador para el desarrollo de estas enfermedades de gravedad, erigiéndose, de este modo, como un problema de salud mundial de primer orden que, según las actuales previsiones, alcanzará tintes dramáticos en las próximas décadas.

A pesar de la gran problemática e impacto que plantea el síndrome cardiometabólico, las ENT derivadas de este síndrome, podrían reducirse significativamente y prevenirse eficazmente con un cambio en la principal causa de expansión: el estilo de vida moderno. Por tanto, es un cambio que debe centrarse en la lucha contra el sedentarismo y los malos hábitos alimenticios.

Todo esto, se encuentra sumergido en una etapa en el que el mundo de las tecnologías de la información y telecomunicaciones está teniendo un gran auge e incidencia en el ámbito sanitario, lo que lleva al surgimiento de nuevas herramientas, como PREDIRCAM, una Plataforma Inteligente destinada a la Monitorización, Tratamiento y Prevención Personalizados de la Diabetes Mellitus, el Riesgo CardioMetabólico y la Insuficiencia Renal, mediante un seguimiento del registro de las variables corporales, la alimentación y la actividad física, fomentando la motivación y la adherencia hacia estilos de vida saludables a través del uso de la tecnología.

En la plataforma, se registran todas las variables corporales, la alimentación y la actividad física, así como, dispone de un sistema de recomendaciones que se basa en felicitar, alertar o recomendar en aquellos aspectos que lo requieran, a los usuarios. Estas notificaciones de nutrición, ejercicio y uso de la plataforma, se basan en el cumplimiento o incumplimiento de unos límites prefijados. Es decir, a partir de los valores de nutrición, ejercicio y uso, se evalúa si los usuarios cumplen o no con las recomendaciones y límites establecidos por los profesionales en función de nivel de cumplimiento. Estos límites, son determinados en base a la pirámide nutricional y a las recomendaciones actuales de nutrición y actividad física, avaladas por la comunidad médica, siendo estos estándares y aplicadas a todos los sujetos indistintamente.

La plataforma Predircam ha sido objeto de un estudio clínico, que aún se encuentra en vías de desarrollo y próximo a la finalización. De esta forma, se dispone de una base de datos en la que se encuentran todos los registros de nutrición, ejercicio, y uso de la plataforma, que corresponden a la evolución de los pacientes sometidos al programa.

Asimismo, el desarrollo tecnológico a niveles exponenciales en el área de cómputo, la gran potencia de los nuevos ordenadores para realizar operaciones de análisis, así como la transmisión de datos y una mejor gestión del manejo y almacenamiento de la información, conduce a un gran interés por el análisis de las bases de datos, surgiendo así: la minería de datos, Data mining. Es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones, comportamientos, tendencias o reglas que puedan generar algún modelo que permita comprender mejor el comportamiento de los datos. Y tras este proceso de análisis de datos que consta en examinar: perfil, ingestas diarias, actividad física, hábitos… del usuario y la extracción de conocimiento a partir de estos, en los últimos años, han proliferado los sistemas de recomendación personalizados, los cuales proporcionan a los usuarios: información, sugerencias, advertencias…individualizadas.

Los aspectos comentados en párrafos anteriores, la relevancia de la problemática, el desarrollo tecnológico y la implementación de herramientas y sistemas que apoyen el cambio de comportamientos y la adopción de hábitos de vida saludables, justifican la utilidad de analizar los datos recogidos en el estudio clínico del proyecto Predircam y se alinea con el objetivo de este proyecto.

1.2 Objetivos

Este Trabajo Final de Grado se enmarca dentro del proyecto Predircam y tiene por objetivo principal el diseñar un sistema de recomendaciones personalizadas que, a partir de las variables de cada paciente, de la plataforma tecnológica Predircam, proporcione recomendaciones con la finalidad de apoyar la adherencia al tratamiento y mejorar los resultados hacia un cambio de estilo de vida saludable.

El desarrollo de este trabajo ha sido realizado bajo la supervisión de la profesora Mª Elena Hernando Pérez y los investigadores José Iniesta y José Tapia de la Universidad Politécnica de Madrid, y con el apoyo de profesionales médicos y nutricionistas del Hospital Sant Pau de Barcelona.

Objetivos específicos

Para ello, los objetivos específicos de este proyecto son:

- Estudiar las funcionalidades y módulos de la plataforma Predircam, así como analizar el sistema de recomendación que dispone actualmente.
- Llevar a cabo un proceso de extracción de conocimiento (KDD) de la base de datos obtenida del estudio clínico, con el propósito de determinar las variables más significativas de nutrición y ejercicio en la pérdida de peso. Y, a partir de estas, realizar un conjunto de reglas que se incluirían en el diseño del sistema de recomendación personalizado.

Las tareas planificadas para la consecución de estos objetivos específicos y por tanto del objetivo principal, son:

- Entender cada uno de los valores de nutrición y actividad física adquiridos mediante la plataforma, a fin de comprender cada una de las variables de la base de datos obtenida del estudio clínico.
- Instruirse en las diferentes técnicas de minería de datos: técnicas de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, aprendizaje bayesiano, entre otros.
- Estudiar los diferentes sistemas de recomendación existentes.

2. ESTADO DEL ARTE

En este apartado, se desarrolla el contexto del proyecto, definiendo los conceptos médicos fundamentales y exponiendo el importante papel de la tecnología en el ámbito médico, para el control y monitorización de pacientes. Asimismo, se presenta la plataforma PREDICRCAM, el proceso de extracción de conocimiento a partir de los datos y los diferentes sistemas de recomendación.

2.1. Las enfermedades no transmisibles

Las enfermedades no transmisibles (ENT), también conocidas como enfermedades crónicas, son afecciones de larga duración y resultan de la combinación de factores genéticos, fisiológicos, ambientales y conductuales. Son la principal causa de mortalidad en nuestro país y en todo el mundo, afectando a todos los grupos de edad [1].

Los principales tipos de ENT son las enfermedades cardiovasculares (como los ataques cardiacos y los accidentes cerebrovasculares), el cáncer, las enfermedades respiratorias crónicas (como la enfermedad pulmonar obstructiva crónica y el asma) y la diabetes [1]. Éstas se deben en gran medida a cuatro factores de riesgo comportamentales que se han afianzado de forma generalizada como parte de la transición económica, los rápidos procesos de urbanización y los modos de vida del siglo XXI: el consumo de tabaco, las dietas malsanas, la inactividad física y el uso nocivo del alcohol [2].

Síndrome metabólico

En el contexto de las enfermedades no transmisibles se inserta el síndrome metabólico (SM) [3,4].Desde principios del siglo XX, la comunidad médica ha reconocido la existencia de una agrupación de factores de riesgo de origen metabólico que aumenta, en quienes la sufren, la probabilidad de padecer Diabetes Mellitus o enfermedad cardiovascular.

Y, desde entonces, muchos han sido los intentos de dotar a este conjunto de factores de riesgo de una definición y diagnostico definitivos, pero ante la confusión existente, la Federación Internacional de la Diabetes (IDF) [5] solicitó la creación de un grupo de expertos de las distintas regiones del mundo para establecer una nueva definición mundial del síndrome metabólico. Como resultado de este trabajo conjunto, se estableció la definición que, hoy por hoy, es internacionalmente aceptada. Esta definición se muestra a continuación [6] .

"El síndrome metabólico es un conjunto complejo de factores de riesgo interrelacionados que conllevan a un aumento del riesgo de padecer una enfermedad cardiovascular o diabetes mellitus tipo 2. Los factores de riesgo incluyen: la hiperglucemia, la elevación de las concentraciones de triglicéridos, el aumento de la presión arterial (PA), la disminución de las concentraciones del colesterol unido a las lipoproteínas de alta densidad (cHDL), y la obesidad de distribución central".

Antes de profundizar en la utilidad de esta definición, conviene clarificar de manera concisa los siguientes componentes:

- **Glucemia**. Es el término utilizado para referirse a la medida de concentración de azúcar en el plasma sanguíneo mientras que la **hiperglucemia** es el término técnico utilizado para referirse a altos niveles de glucosa (azúcar)en sangre. El sistema endocrino regula la cantidad de azúcar que se almacena y utiliza la necesaria como energía, para el correcto funcionamiento de las células.
- **Hipertrigliceridemia (elevación de las concentraciones de triglicéridos).** Se refiere al exceso de triglicéridos en sangre. Los triglicéridos son el principal tipo de lípido (grasa), y sirven cómo almacenamiento de energía en las células.
- **Hipertensión** [7]. Es una patología continua o sostenida que consiste en la elevación de los niveles de presión arterial. La presión arterial es la fuerza que ejerce la sangre contra las paredes de las arterias.
- **Disminución de lipoproteínas de alta densidad (cHDL),** denominadas generalmente por las siglas en inglés HDL (high density lipoproteins), son un tipo de lipoproteínas que transportan el colesterol, los triglicéridos y otras grasas, desde las células de la pared vascular y otros órganos del cuerpo hasta el hígado, es por ello que se le conoce como el colesterol o lipoproteína buena.

  Las LDL realizan el transporte en sentido contrario, llevando estos lípidos a las regiones periféricas del cuerpo y su exceso produce la adhesión de los lípidos a las paredes arteriales. Por el contrario, el HDL ayuda a evitar estas acumulaciones.
- **Obesidad** [8]**.** Es una enfermedad crónica tratable que aparece cuando existe un exceso de tejido adiposo (grasa) en el cuerpo. La obesidad se clasifica en dos tipos: central o androide y periférica o imoide.

  La obesidad androide o central localiza la grasa en el tronco, el tejido adiposo se concentra en la mitad superior del cuerpo, sobre todo en la región abdominal. Es la que mayor riesgo presenta, por estar la grasa más cerca de órganos importantes, y se asocia directamente con un aumento del riesgo de desarrollar enfermedad cardiovascular y enfermedades como diabetes tipo 2, ateroesclerosis e hiperlipidemia. Esto es debido a que la grasa intraabdominal posee características metabólicas diferentes a las de otros tejidos adiposo.

  De manera práctica se puede determinar fácilmente el tipo de obesidad del individuo obteniendo el Índice Cintura-Caderas del mismo (en inglés, Waist-Hip Ratio o WHR) mediante la división del perímetro de cintura entre el de las caderas. Si este cociente es superior a 1 en el hombre y a 0,9 en el caso de la mujer, la obesidad se considera de tipo androide.

De entre todos estos factores, la nueva definición de la IDF, ha tenido en cuenta la gran cantidad de datos que indican que la adiposidad central (abdominal) es común a todos los componentes del síndrome metabólico. Al ser el perímetro de la cintura un parámetro bien aceptado como sustituto de la adiposidad abdominal, se considera en la actualidad un requisito necesario para establecer el diagnóstico de síndrome metabólico.

Esta consideración tiene la ventaja de que la simple determinación del perímetro de la cintura representa una primera prueba de detección del síndrome y ésta se puede realizar de manera sencilla y con un coste bajo, en cualquier parte del mundo.

El criterio para diagnosticar el Síndrome Metabólico, exige que, además del elevado valor de la adiposidad abdominal se den en el individuo 2 de los otros 4 factores citados en la definición, como se recoge en la siguiente *Tabla 2.1*:

Tabla 2.1. *Criterios para el diagnóstico del Síndrome Metabólico según criterios IDF* [5]

Factores	Condición
Elevación de los triglicéridos	≥ 150 mg/dL (1,7 mmol/l) o Recibe tratamiento farmacológico.
Disminución del cHDL	Hombres ≤ 40 mg/dL (0,9 mmol/l) Mujeres ≤ 50 mg/dL (1,1 mmol/l) o Recibe tratamiento farmacológico.
Elevación de la presión arterial	Sistólica ≥ 130 mmHg Diastólica ≥ 85 mmHg o Recibe tratamiento farmacológico de la hipertensión.
Elevación de la glucemia en ayunas	≥ 100 mg/dL o Recibe tratamiento farmacológico de la hiperglucemia.

Aunque los criterios anteriormente definidos, determinan de manera precisa el concepto teórico de Síndrome Metabólico, se ha dudado de su utilidad práctica como diagnóstico. Así, en el año 2010, la Organización Mundial de la Salud recomendó no contemplar como personas afectadas por el síndrome a los enfermos de diabetes o alguna enfermedad cardiovascular pues en ellos la identificación de los factores implicados ya no permite realizar una prevención primaria.

De esta manera, la OMS prefiere definir el Síndrome Metabólico como un estado premórbido (para remarcar esta diferencia suele denominarse Síndrome Metabólico Premórbido o SMP) más que un diagnóstico clínico [9], es decir, como un indicador que se utiliza para identificar un conjunto de factores de riesgo ante los cuáles es posible aplicar un tratamiento, con el objetivo de centrar el uso clínico del síndrome en la prevención primaria de la diabetes y la enfermedad cardiovascular.

El Darios fue el primer estudio en el que se actualizó la prevalencia del SM en España siguiendo la nueva definición acuñada por la OMS de este síndrome. Así, mediante esta nueva definición, llamada síndrome metabólico premórbido, la incidencia en España se sitúa en el 26 por ciento de los hombres y el 24 por ciento de las mujeres.

Y es que este estudio junto con otros, han demostrado que este síndrome duplica en quien lo sufre el riesgo de padecer una enfermedad cardiovascular y multiplica por cinco la probabilidad de desarrollar diabetes. Este hecho adquiere especial relevancia cuando se analiza la ya elevada prevalencia global de estas enfermedades.

Impacto global

La mayoría de las enfermedades no transmisibles, derivan de los factores de riesgo del síndrome metabólico, y son la principal causa de muerte a nivel mundial, y afectan a todos los grupos de edad, siendo 15 millones de todas las muertes que se producen, atribuidas a las ENT, en personas entre los 30 y los 69 años de edad.

En las últimas décadas, tanto las enfermedades relacionadas con el Síndrome Metabólico como las patologías constituyentes del mismo, han aumentado exponencialmente a nivel mundial. Esto es debido principalmente por la adopción generalizada de diferentes costumbres características del estilo de vida moderno. Así, la urbanización de la población se ha asociado a un cambio radical del estilo de vida caracterizado por una dieta basada en alimentos de alto contenido calórico, un descenso importante de la actividad física y el consumo de tabaco y alcohol [2,10] .Estas conductas han dado lugar a una serie de factores de riesgo que desencadenan en patologías graves, las llamadas enfermedades no transmisibles.

La influencia de estas conductas de riesgo y de otras causas metabólicas y fisiológicas de la epidemia mundial de ENT abarca lo siguiente [2] .

- **El tabaquismo**: alrededor de 6 millones de personas mueren a causa del tabaco cada año, tanto por el consumo directo como por el pasivo. Se estima que el tabaquismo causa aproximadamente el 71% de los casos de cáncer de pulmón, el 42% de las enfermedades respiratorias crónicas y alrededor del 10% de las enfermedades cardiovasculares.

- **El sedentarismo**: aproximadamente 3,2 millones de personas mueren a causa del sedentarismo cada año. Al menos un 60% de la población mundial no realiza la actividad física necesaria para obtener beneficios para la salud, esto conlleva a que corren un mayor riesgo que las otras de morir por cualquier causa.

- **El uso nocivo del alcohol**: aproximadamente 2,3 millones de personas mueren a causa del uso nocivo del alcohol cada año, lo que representa alrededor del 3,8% de todas las muertes que tienen lugar en el mundo.

- **La dieta no saludable**: el consumo de fruta y verdura en cantidades suficientes reduce el riesgo de padecer enfermedades cardiovasculares, cáncer de estómago y cáncer colorrectal. La mayoría de las poblaciones consumen niveles de sal mucho más elevados que los recomendados por la OMS para prevenir enfermedades; un consumo elevado de sal es un factor determinante que aumenta el riesgo de padecer hipertensión y enfermedades cardiovasculares.

- **La hipertensión**: se estima que la hipertensión causa 7,5 millones de muertes, lo que representa alrededor del 12,8% del total. Es un factor de riesgo muy importante de las enfermedades cardiovasculares.
- **El sobrepeso y la obesidad**: al menos 2,8 millones de personas mueren cada año por sobrepeso u obesidad. El riesgo de padecer cardiopatías, accidentes cerebrovasculares y diabetes crece paralelamente al aumento del índice de masa corporal (IMC). Un IMC elevado aumenta asimismo el riesgo de padecer ciertos tipos de cáncer.
- **El hipercolesterolemia**: se estima que la hipercolesterolemia causa 2,6 millones de muertes cada año; aumenta el riesgo de padecer cardiopatías y accidentes vasculares cerebrales.
- **Las infecciones relacionadas con el cáncer**: al menos 2 millones de casos de cáncer anuales, el 18% de la carga mundial de cáncer, pueden atribuirse a ciertas infecciones crónicas. Los principales agentes infecciosos son el virus del papiloma humano, el virus de la hepatitis B, el virus de la hepatitis C y Helicobacter pylori. Dichas infecciones pueden prevenirse en gran medida con vacunas y medidas para evitar la transmisión, o bien pueden tratarse.

La situación en España es muy similar a la analizada a nivel mundial: las enfermedades cardiovasculares son la primera causa de muerte representando el 29,66% del total de fallecimientos, lo que las sitúa por encima del cáncer (27,86%) y de las enfermedades del sistema respiratorio (11,08%),[11,12] y en torno al 10% de la población sufre diabetes. Además, el 31% de la población adulta padece el Síndrome Metabólico, porcentaje que se reduce, pero al ceñirnos a la última propuesta de la OMS de definición para el Síndrome Metabólico Premórbido, es decir excluyendo a quienes ya sufren la diabetes y/o la enfermedad cardiovascular, disminuye el porcentaje de personas afectadas siendo la incidencia en España de un 26 % en hombres y un 24 % en mujeres.

Con todos los datos expuestos, queda claro el gran impacto que tienen estas enfermedades en el mundo actual. Un impacto que ya tiene carácter de epidemia y que en un futuro será aún mayor, debido a diversos factores que hacen presagiar un empeoramiento de la situación en los próximos años. Es por todo ello, que muchos se refieren al Síndrome Metabólico y sus enfermedades asociadas, como la mayor amenaza para la salud mundial en las próximas décadas y se los identifique conjuntamente como “La Epidemia del Siglo XXI”.

Por lo tanto, el Síndrome Metabólico según la IDF, se puede concluir con la *Figura 2.1*.

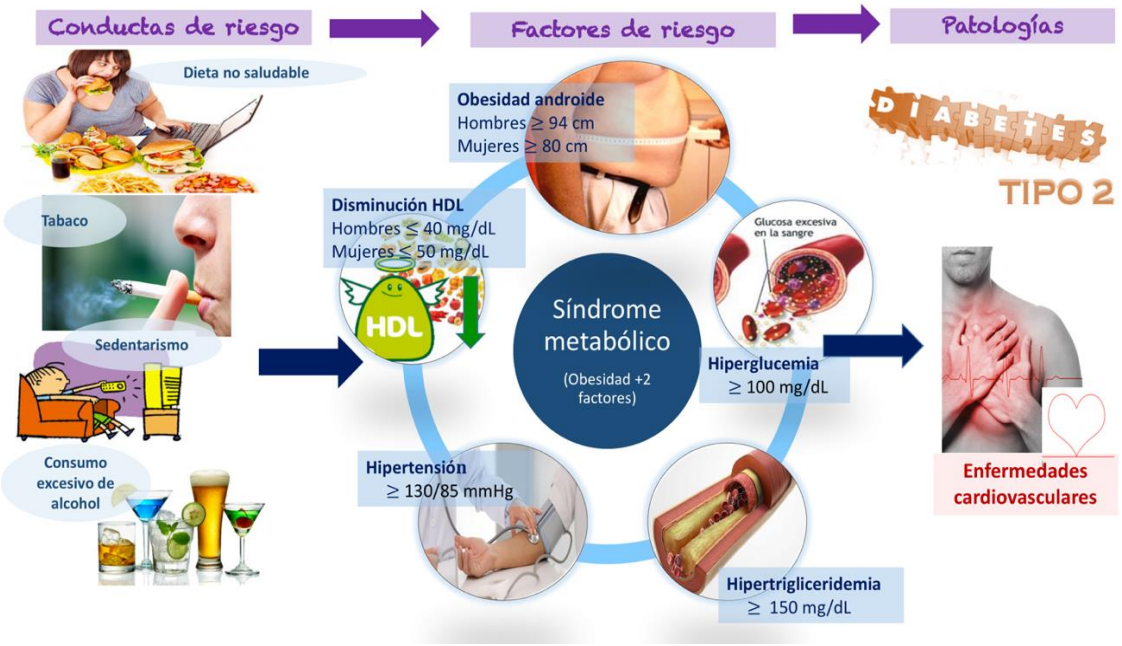


***Figura 2.1**. Síndrome metabólico según IDF.*

Elemento clave de prevención: la modificación del estilo de vida

Las proporciones de epidemia que han alcanzado las ENT junto con la rápida expansión del Síndrome Metabólico, otorgan una negativa perspectiva de futuro. Pero a pesar de la gran incidencia mundial, en la mayor parte de los casos, las ENT podrían reducirse de manera significativa y prevenirse eficazmente con una modificación en la principal causa de su expansión: el estilo de vida moderno. Un cambio que se debe centrar, en la lucha contra el sedentarismo y los malos hábitos alimenticios [2].

Sedentarismo

El sedentarismo constituye parte integral del síndrome metabólico (SM), y hace referencia a aquel individuo que invierte menos del 10% de su gasto energético diario.

Morris y Paffenbarger, manifestaron una base cuantitativa del efecto protector del ejercicio y se centraron en el estudio de la actividad física del tiempo libre y la enfermedad coronaria, y realizaron diversos estudios e informes, que demostraron que la actividad física aporta múltiples beneficios para la salud, aumentando la longevidad (*Figura 2.2*), y ejerce un efecto protector contra la mortalidad temprana por enfermedad coronaria [13].

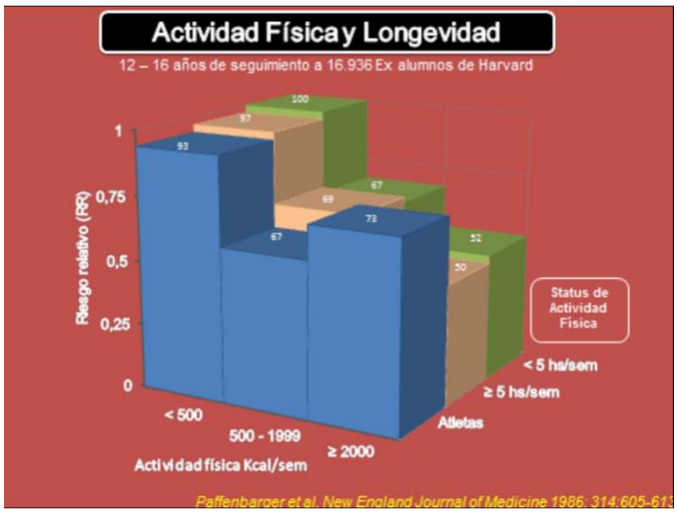


Figura 2.2. *Estudio de Paffenbarger, sobre la influencia de la actividad física en la longevidad en 1986*

Así pues, las personas con poca actividad física corren un riesgo entre un 20% y un 30% mayor que las otras de morir por cualquier causa. Y diversos estudios prospectivos en poblaciones de alto riesgo han concluido que la actividad física regular está asociada con un menor riesgo de DM2 [14] .

Por lo tanto, habiéndose probado, que el sedentarismo se asocia con ganancia de peso y aumento de grasa visceral, lo cual predispone al individuo a una adipocitopatía proinflamatoria con resistencia insulínica y aparición del fenotipo característico del Síndrome Metabólico, se puede afirmar que la actividad física es una herramienta clave y de gran eficacia para la prevención del SM.

Malos hábitos alimenticios

En la sociedad sedentaria no sólo se han reducido las oportunidades del gasto energético a través del ejercicio físico, sino que al mismo tiempo ha aumentado el consumo excesivo de calorías baratas, con el consiguiente problema creciente de la obesidad a nivel mundial [15] .Por tanto, desde el punto de vista médico, el concepto de sedentarismo debe extenderse a este doble significado, y enfocarse al desajuste calórico que hoy día afecta a gran parte de la humanidad.

Asimismo, muchas investigaciones recientes, dejan evidente que la conducta alimenticia y los hábitos de vida en lo referente a los patrones de alimentación, tienen un impacto de máxima importancia en la prevención de las ENT y SM, y por tanto deben formar parte de cualquier tipo de herramienta o medida de prevención que pretenda actuar contra este tipo de problemas.

2.2. Telemedicina

En el desarrollo de las Tecnologías de la Información y las Comunicaciones (TIC), el conjunto de técnicas y dispositivos empleados para el tratamiento y la transmisión de datos, surge la Telemedicina como una forma de luchar contra las barreras geográficas incrementando la accesibilidad a los cuidados de salud, especialmente en zonas rurales y países en desarrollo.

La Telemedicina significa medicina a distancia (diagnóstico, tratamiento, seguimiento…) mediante recursos tecnológicos que optimizan la atención, ahorrando tiempo y costes, y aumentando la accesibilidad. De forma más amplia y con matices la OMS [16] define la telemedicina como: "*Aportar servicios de salud, dónde la distancia es un factor crítico, por cualquier profesional de la salud, usando las nuevas tecnologías de comunicación para el intercambio válido de información en el diagnóstico, el tratamiento y educación continuada de los proveedores de salud, todo con el interés de mejorar la salud de los individuos y sus comunidades*". Asimismo, analizando y centrándose en los objetivos del presente proyecto, la definición propuesta por INSALUD en el "Plan de Telemedicina del Insalud" [17] focaliza el término de telemedicina en el paciente, y lo define como "*la utilización de las tecnologías de la información y de las comunicaciones como un medio de proveer servicios médicos, independientemente de la localización tanto de los que ofrecen el servicio, como de los pacientes que lo reciben, y la información necesaria para la actividad asistencial*".

A partir de estas definiciones se puede establecer que el término Telemedicina incluye los tipos de servicios que se muestran en la *Figura 2.3.*

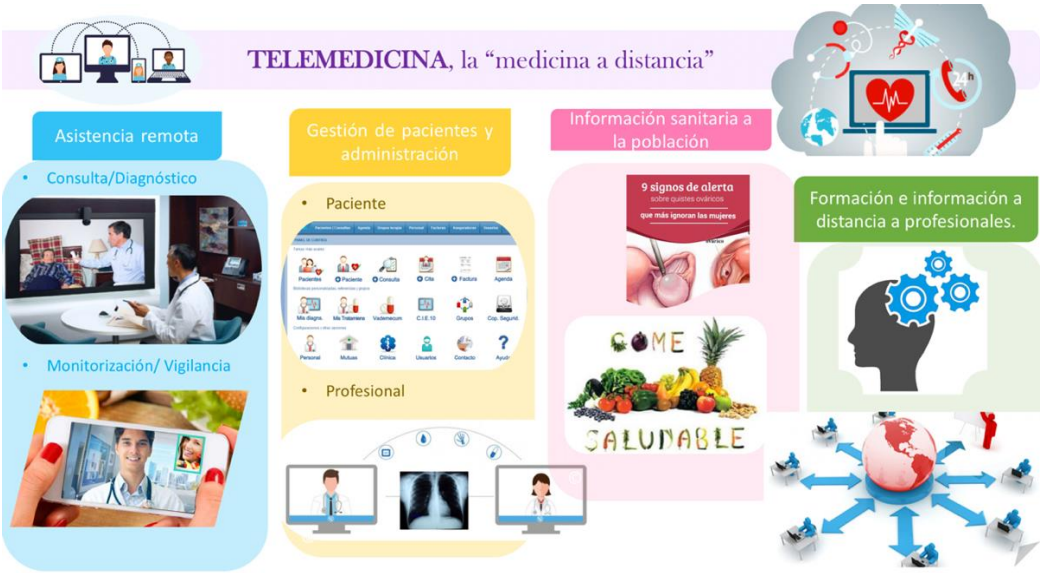


***Figura 2.3**. Servicios que ofrece la Telemedicina.*

Por tanto, la Telemedicina tiene como propósito proveer servicios de salud destinados a mejorar la salud general o mantener el bienestar de la sociedad. Estos servicios engloban: prestación asistencial a los pacientes, facilidad en los procesos administrativos y proporción de información sanitaria. De este modo, los usuarios de un sistema de Telemedicina pueden ser tanto profesionales sanitarios (personal médico, de enfermería, administrativos…) como los pacientes y ciudadanos en general.

2.3. PREDIRCAM

Tal y como se ha detallado, es una prioridad urgente y global la eliminación del sedentarismo y la modificación de la conducta alimenticia, del mismo modo que el cambio del resto de hábitos perjudiciales, todo ello requiere que la persona lleve a cabo un cambio conductual amplio. Es decir, aunque se efectúe apoyo institucional y de la sociedad, debe ser el propio individuo el que, con gran fuerza de voluntad, vele constantemente por la modificación de sus costumbres.

Pero, este tan necesario cambio de hábitos presenta una serie de debilidades: dificultad del registro de ingestas diarias y actividad física, ausencia de una figura profesional que ejerza un seguimiento y asesoramiento continuado, falta de motivación… Por tanto, se aprecia, que es necesario facilitar algún modo para el individuo que desee un cambio en su estilo de vida, que lo haga más activo y que junto a otras actuaciones complementarias, como el control de la alimentación, la renuncia al tabaco o la disminución del consumo de alcohol, se reduzca de manera significativa el riesgo de padecer el Síndrome Metabólico y sus enfermedades asociadas.

Paralelamente, cómo se ha expuesto, el exponencial crecimiento y desarrollo de las TIC ejerce un papel importante y realmente decisivo. Y aplicadas al ámbito sanitario, dónde surge la Telemedicina, aportan enormes capacidades para monitorizar y controlar de manera efectiva a diferentes tipos de pacientes.

Es en este contexto, en el que surge el **proyecto PREDIRCAM**[18] una Plataforma Inteligente de Tecnologías Biomédicas para la monitorización, prevención y tratamiento personalizados de la diabetes mellitus, el riesgo cardiometabólico y la insuficiencia renal. Es una plataforma tecnológica, para pacientes y profesionales de la salud que tiene como propósito principal mejorar la eficacia en las modificaciones del comportamiento del estilo de vida a través del uso intensivo de las nuevas tecnologías de la información y la comunicación, con objetivo de prevenir las ENT derivadas de los factores de riesgo del SM y así, revertir la catastrófica tendencia actual.

La plataforma tecnológica, asociada a PREDIRCAM trata de abordar de forma integral el manejo del sobrepeso y la obesidad. Para ello, facilita la monitorización y el registro de variables como el peso, la tensión arterial, la alimentación, el movimiento y la actividad física, fomentando la motivación y la adherencia a través de recomendaciones y un feedback, a través de mensajería web, continuado entre el usuario y el profesional.

Al principio, se desarrolló una plataforma tecnológica que fue terminada y validada, por tanto, se planteó y ejecutó un estudio piloto, con sujetos ajenos al equipo de desarrollo, para comprobar el funcionamiento. El estudio piloto se llevó a cabo con éxito, de forma que el paso siguiente consistía en la realización de un estudio clínico en el que tomarían parte pacientes y profesionales médicos reales. Pero, no se pudo llevar a cabo dicho estudio clínico, motivo por el que el proyecto quedó aparcado y la plataforma desactualizada.

Más tarde se reactivó un nuevo proyecto y desarrolló una nueva versión de la plataforma, PREDIRCAM 2, haciendo uso de las nuevas tecnologías y herramientas comentadas en puntos anteriores, con el objetivo de realizar un pequeño estudio piloto y el pretendido estudio clínico.

Módulos de la plataforma PREDIRCAM 2 [19]

Esta nueva plataforma, es destinada, tanto para profesionales como para los pacientes, pero cada uno de ellos, dispone de un tipo de pantalla y opciones determinadas.

Los profesionales médicos, tienen disponible un escenario para el registro de las diferentes visitas asociadas a cada usuario, permitiéndoles almacenar y visualizar los parámetros medidos obtenidos a partir de las pruebas médicas solicitadas para cada paciente. A través de estas visitas, se realizan también las diferentes prescripciones de actividad física y nutrición que los pacientes deben seguir hasta la realización de la siguiente visita.

Por otro lado, **los pacientes** disponen de tres módulos que les permiten el registro y monitorización de la nutrición y la actividad física para un seguimiento y una valoración de sus hábitos de vida:

- El módulo de nutrición (*Figura 2.4*), permite el registro de las ingestas diarias. Y en base a todos los alimentos registrados, automáticamente, se realiza una valoración nutricional que es reflejada en un conjunto de gráficos e indicadores que permiten la visualización del estado diario, tanto a nivel de los diferentes grupos de alimentos, como a nivel calórico y de macronutrientes, identificando excesos y deficiencias

***Figura 2.4**. Gráficos e indicadores de nutrición Predircam*

Y tras la investigación y puesta en común con el personal médico de Sant Pau, se establecieron unos límites para cada uno de los elementos del área de nutrición. Estos límites se especifican en la *Tabla 2.2*, y se encuentran basados, en personas con sobrepeso y obesidad, cuya prescripción de dieta oscila entre 1500 y 1900 kilocalorías diarias.

Los valores de macronutrientes, están condicionados a la prescripción, es por ello que en la tabla aparecen unos rangos, el primer valor es la cantidad que debería consumir una persona cuya prescripción es 1500 kcal/día; mientras que el segundo valor corresponde a la prescripción de 1900 kcal/día. Es decir, tomando el rango de proteínas (75-95) g/día, significa que una persona con una prescripción de 1500 kcal/día debe consumir 75 g/día (lo que corresponde al 20%) mientras que una con una prescripción de 1900 kcal/día debe consumir 95g/día de proteínas.

***Tabla 2.2**. Recomendaciones de nutrición.*

Grupo	Alimento	Mínimo	Máximo
Grupos principales	Cereales y derivados, legumbres y tubérculos	300 kcal	50 % total de kcal/día.
	Carnes, embutidos, pescados, mariscos y huevos.	280 kcal	360 kcal
	Verduras y hortalizas	140 kcal	Sin límite
Macronutrientes La cantidad depende de la prescripción de kcal/día.	Grasas (30%)	Según prescripción: (50-63.3) g/día	
	Proteínas (25%)	Según prescripción: (75-95) g/día	
	Carbohidratos (50 %)	Según prescripción: (187,5-237,5) g/día	
Grasas	Saludables	Se deberían consumir un mayor porcentaje de grasas "buenas" que de grasas "malas", por lo que el balance es positivo cuando el porcentaje de **grasas "buenas" era superior al 50%** del total consumido.	
	No saludables		
Lácteos	Leche, queso, yogurt.	260 kcal (2 unidades)	390 kcal (3 unidades)
Frutas		140 kcal (2 unidades)	210 kcal (3 unidades)
Aceites y frutos secos		90 kcal (2 unidades)	180 kcal (4 unidades)
Alimentos no saludables		No existe ningún límite inferior, puesto que el límite superior es c**ualquier cantidad que supere el 0.**	
Bebidas	Agua y bebidas saludables	1500 ml (6 vasos)	3000 ml (12 vasos)
	Bebidas no saludables	No existe ningún límite inferior, puesto se recomienda que el **consumo de este tipo de bebidas sea 0.**	
	Alcoholes	0	140 kcal (2 unidades)

- El módulo de ejercicio (*Figura 2.5*), posibilita la inserción de la actividad física realizada, ya sea manualmente o mediante la sincronización de pulsímetros Polar RS400. Y a partir de todos los datos de ejercicio, se generan una serie de gráficos e indicadores semanales, que permiten reflejar el estado de consecución del plan de ejercicio, teniendo en cuenta las calorías consumidas, la duración y sesiones realizadas.

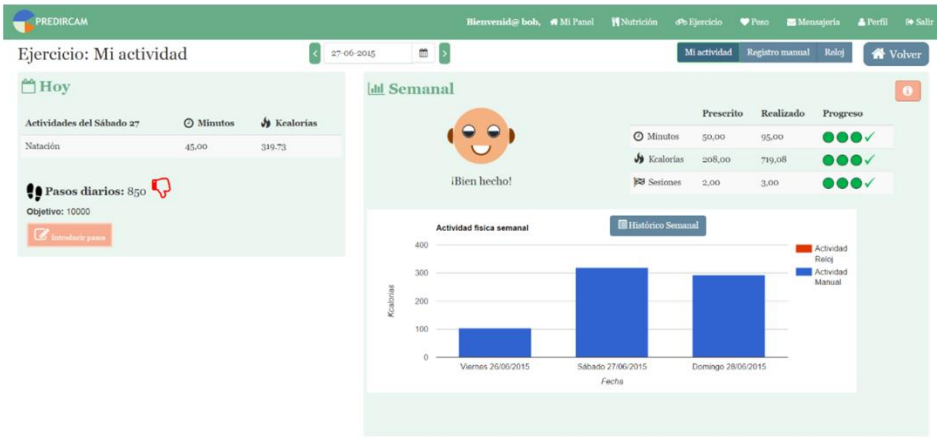


***Figura 2.5**. Gráficos e indicadores de ejercicio Predircam.*

- Por último, el módulo de Peso e ICC (*Figura 2.6*), permite la introducción de los diferentes parámetros corporales de interés, que son, el peso, las medidas de cintura/cadera y los cálculos asociados al índice de masa corporal e índice cintura-cadera. Y estos datos, se representan gráficamente con el fin de mostrar una tendencia que permita observar la evolución temporal.

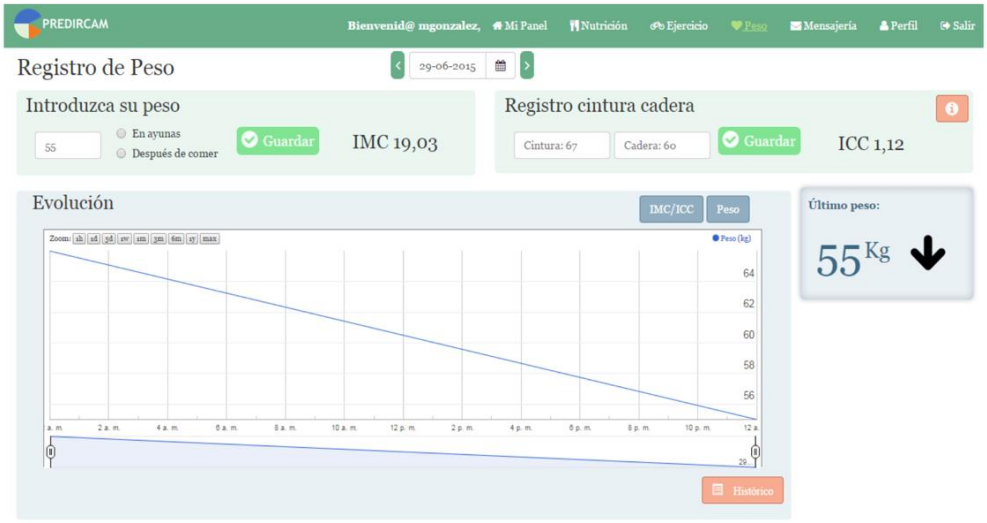


***Figura 2.6**. Gráfico de pesos Predircam.*

Sistema de recomendaciones [19]

En base a los datos registrados en la plataforma, a través de los módulos de nutrición, actividad física y peso, así como la frecuencia de uso del sistema por parte de los usuarios, se generan un conjunto de notificaciones que realimentan de forma periódica la actividad de estos.

Estas, tienen el fin de indicar a los usuarios si están realizando los registros necesarios, así como felicitarles o aconsejarles por los logros y los errores que comenten en el desempeño de la misma.

Todas ellas, se realizan en base a las recomendaciones actuales de nutrición y actividad física, y existen 3 tipos bien diferenciados:

- **Felicitaciones** (*Tabla 2.3*): son el refuerzo positivo que obtienen los pacientes, cuando consiguen los objetivos de la plataforma, tales como el cumplimiento de los parámetros de actividad física o el consumo adecuado y estructurado de alimentos.

Tabla 2.3. *Condiciones de felicitaciones Predircam.*

Tipo	Ámbito	Condiciones
Felicitaciones	**Nutrición**	✓ Desviación kcal < +/- 10% del objetivo
		✓ Desviación macronutrientes < +/- 10% del objetivo
		✓ Min< Media grupos funcionales < Max
	Ejercicio	✓ Cumple Kcal objetivo ✓ Cumple sesiones objetivo ✓ Cumple tiempo objetivo
		✓ Cumple kcal objetivo ✓ Cumple sesiones objetivo
		✓ Cumple kcal objetivo ✓ Cumple tiempo objetivo
		✓ Kcal quemadas >750 ✓ Más de 3 días con sesiones

- **Recomendaciones** (*Tabla 2.4*)**:** estas recogen todos los aspectos que no se han cumplido a lo largo del periodo bajo análisis, en cualquiera de los ámbitos del sistema. Por tanto, los usuarios, reciben este tipo de notificaciones, en las que se les dice aquello en lo que no se han mantenido dentro de los límites recomendados y se les anima a conseguirlo en las sucesivas jornadas.

Tabla 2.4*. Condiciones de recomendaciones Predircam.*

Tipo	Ámbito	Condiciones
Recomendaciones	**Nutrición**	✓ Desviación kcal > +/- 10% del objetivo
		✓ Desviación macronutrientes > +/- 10% del objetivo
		✓ Consumo de grupos funcionales < mínimo recomendado ✓ Media grupos funcionales > máximo recomendado
	Ejercicio	✓ Kcal quemadas < 750

- **Alertas** (*Tabla 2.5*): estas notificaciones tienen como objetivo, no valorar si el paciente está siguiendo correctamente el plan de actividad y nutrición acordado, sino de identificar y tratar de informar a aquellos pacientes que no están haciendo un uso continuado de la aplicación.

Tabla 2.5. *Condiciones de advertencias Predircam.*

Tipo	Condiciones
Alertas	✓ No hay registros en fin de semana.
	✓ Si número de ingestas < 12, con objetivo de recordar que se deben registrar 3 días completos (desayunos, comidas, cenas, meriendas o snacks).
	✓ Si el número de registro de ejercicio es < 3.

Las notificaciones se generan con cierta periodicidad, por lo que se utilizan diferentes temporizadores:

Tanto las **felicitaciones como las recomendaciones** se centran en los dos focos principales de la plataforma, que son la nutrición y la actividad física, y para la generación de estas, se emplea un temporizador de carácter semanal y se obtienen los datos que los pacientes han introducido en la plataforma a lo largo de la semana, a fin de analizarlos y procesarlos.

Por otro lado, las **notificaciones de alertas** se generan con periodos más cortos, y se centran en el análisis del uso de los pacientes de los módulos de nutrición, actividad física y en el control de la frecuencia de acceso de los usuarios, así pues, se crean los temporizadores necesarios.

Por último, cuando el usuario accede a la plataforma, las notificaciones se muestran por pantalla para su consulta. Y en esta representación, el usuario visualiza el título de la notificación y la fecha y hora a la que se genera (*Figura 2.7*), de forma que el propio usuario puede seleccionar aquella que quiere leer, desplegando el contenido.

Figura 2.7. *Notificaciones vía Predircam*

Mensajería

En la plataforma, existe un módulo de mensajería, cuya finalidad es tener una vía de contacto incluida dentro del ecosistema y evitar así, hacer uso de otras vías externas.

Este módulo, por un lado, se utiliza para que los pacientes puedan comunicarse con los profesionales médicos en cualquier momento y de forma directa, por dudas sobre el uso de la plataforma, cambio de hora de las visitas o para transmitir cualquier tipo de mensaje. Por otro lado, se utiliza para llevar a cabo visitas telemáticas, en las que los pacientes no acuden al hospital y son ellos mismos quienes recogen los datos requeridos y, a través de este módulo, se los envían al médico.

2.4. Proceso de extracción de conocimiento (KDD)

Las nuevas tecnologías generan inmensas cantidades de datos a un ritmo exponencial gracias al abaratamiento y gran desarrollo del almacenamiento. Este hecho, ha llevado a entrar en la era del Big Data o datos masivos que es definida con la presencia de gran volumen, velocidad, variedad, veracidad de los datos y valor intrínseco del conocimiento extraído.

La finalidad del Big Data es convertir los datos almacenados en información útil. La información se obtiene al procesar los datos: ordenar, agrupar, analizar e interpretar. Y cuando esa información es utilizada o puesta en el contexto o marco de referencia de una persona junto con su percepción personal se transforma en conocimiento. El conocimiento es la combinación de información, contexto y experiencia y una vez resumido, validado y orientado hacia un objetivo genera inteligencia (sabiduría), la cual pretende ser una representación de la realidad [20] .

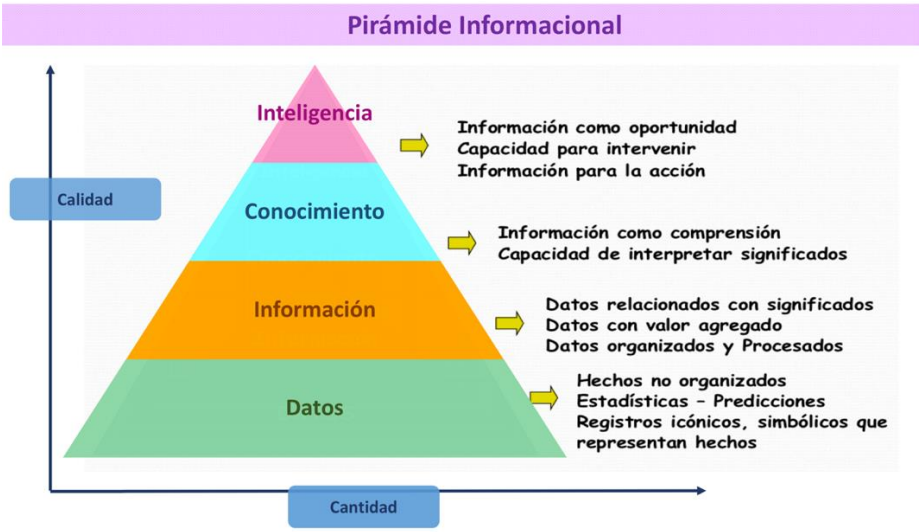


***Figura 2.8**. La Pirámide Informacional.*

Y tal y cómo se observa en la *Figura 2.8*, la Pirámide Informacional representa los 4 conceptos de forma jerárquica en función de las variables calidad vs cantidad. Por tanto, los datos son muchos y poseen poca calidad, la información tiene un volumen menor pero su calidad aumenta. Así mismo sucede con los conocimientos, más calidad y todavía menos volumen. Y, por último, la inteligencia se caracteriza por tener poco volumen, pero una alta calidad informativa, la mayor.

KDD: Proceso de extracción de conocimiento.

Con la finalidad de extraer inteligencia a partir de los datos, se ha de llevar a cabo un proceso de extracción de conocimiento (KDD) [21] el cuál, se define como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos. En esta definición se resumen cuáles deben ser las propiedades deseables del conocimiento extraído:

- **Válido**. Hace referencia a que los patrones deben seguir siendo precisos para datos nuevos (con cierto grado de certidumbre), y no sólo para aquellos que han sido usados en su obtención.

- **Novedoso.** Que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario.
- **Potencialmente útil.** La información debe conducir a acciones que reporten algún tipo de beneficio para el usuario
- **Comprensible.** La extracción de patrones no compresibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento.

Como se deduce de la anterior definición, el KDD es un proceso complejo que se organiza entorno a 5 fases [20] cómo se muestra en la *Figura 2.9.*

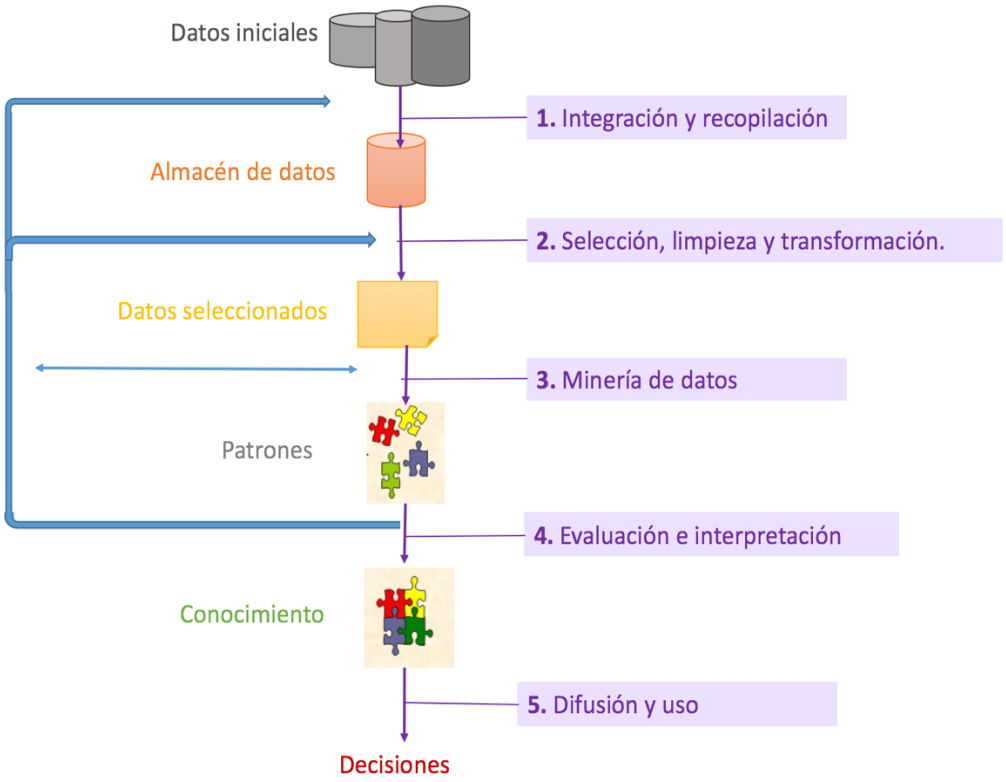


***Figura 2.9**. Fases del proceso de extracción de conocimiento.*

Tal y cómo se observa en la *Figura 2.9*, el KDD es un proceso iterativo e interactivo. Es iterativo ya que la salida de alguna de las fases puede volver a pasos anteriores y porque a menudo son necesarias varias iteraciones para extraer el conocimiento de alta calidad. Es interactivo porque el usuario, o concretamente un experto en el dominio del problema, debe ayudar en la preparación de los datos, validación del conocimiento extraído, ….

Las fases de este proceso son:

1. Integración y recopilación

Se comienza con la comprensión de un problema práctico de inteligencia de datos que surge de la actividad cotidiana, así pues, una vez identificado el problema, se acuden a los datos óptimos.

2. Preparación de datos

Una vez recopilados los datos de análisis se procede a la preparación de esto, que consiste en:

Fase de integración y recopilación de datos.

En esta fase se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas.

A continuación, con propósito de integrar los datos procedentes de diferentes fuentes, se transforman todos los datos a un formato común, frecuentemente mediante un almacén de datos que consiga unificar de manera operativa toda la información recogida detectando y resolviendo las inconsistencias.

Este almacén de datos (data warehousing), es un repositorio de información coleccionada desde varias fuentes, almacenada bajo un esquema unificado que normalmente reside en un único emplazamiento, y facilitan enormemente la "navegación" y visualización previa de los datos, para distinguir qué aspectos puede interesar que sean estudiados.

A pesar de que, un almacén de datos es muy aconsejable, no es imprescindible. En algunos casos, en especial cuando el volumen no es muy grande, se puede trabajar con los datos originales o en formatos heterogéneos.

Fase de selección limpieza y transformación

La calidad del conocimiento descubierto no sólo depende del algoritmo de minería utilizado, sino también de la calidad de los datos minados. Por tanto, debido a que los datos provienen de diferentes fuentes y pueden contener valores erróneos o faltantes, es imprescindible una etapa de selección, limpieza y transformación. En esta etapa se eliminan o corrigen los datos incorrectos (outliers) y se decide la estrategia a seguir con los datos incompletos (missing values).

Además, se proyectan los datos para considerar únicamente aquellos atributos o variables que van a ser relevantes, con el propósito de hacer más fácil la tarea propia de minería y para que los resultados de la misma sean más útiles.

Otra tarea de esta fase de preparación de datos, es la construcción de atributos, la cual consiste en construir automáticamente nuevos atributos aplicando alguna operación o función a los atributos originales con objeto de que estos nuevos atributos hagan más fácil el proceso de minería.

Por último, la fase de selección también incluye la modificación del tipo de datos para facilitar el uso de técnicas que requieren tipos de datos específicos y la discretización de los atributos continuos, es decir, transformar valores numéricos en atributos discretos o nominales.

3. Fase de minería de datos (Data mining)

La fase de minería de datos tiene como objetivo analizar los datos para extraer conocimiento.

Es la fase de modelamiento, en donde métodos inteligentes son aplicados sobre los datos, con el objetivo de extraer patrones previamente desconocidos, válidos, nuevos, potencialmente útiles y comprensibles y que están contenidos u "ocultos" en los datos.

Las herramientas específicas de minería de datos más utilizadas en la actualidad son:

- IBM SPSS Modeler
- SPSS Clementine
- SAS Enterprise Miner
- SQL Analysis Services
- Oracle Data Mining

4. Fase de evaluación e interpretación

En esta fase se evalúan los patrones obtenidos de la fase anterior y se analizan por los expertos, y en caso necesario se vuelve a fases anteriores para una nueva iteración, incluyendo resolver conflictos con el conocimiento que se disponía anteriormente.

Idealmente los patrones descubiertos deben de tener 3 cualidades: ser precisos, ser comprensibles (es decir, inteligentes) e interesantes (útiles y novedosos), pero según las aplicaciones puede interesar mejorar algún criterio y sacrificar ligeramente otro. Como, por ejemplo, en el diagnóstico médico se prefiere patrones comprensibles, aunque su precisión no sea muy buena.

5. Fase de difusión, uso y monitorización

Una vez construido y validado el modelo puede usarse con dos finalidades: para que un analista recomiende acciones basándose en el modelo y en sus resultados, o bien para aplicar el modelo a diferentes conjuntos de datos.

En ambos casos es necesario su difusión, es decir, que se distribuya y comunique a todos los posibles usuarios.

Y aunque el modelo funcione bien debemos continuamente comprobar las prestaciones del mismo, debido a que los patrones pueden cambiar. Por tanto, el modelo deberá ser monitorizado, lo que significa que de tiempo en tiempo el modelo tendrá que ser re-evaluado, re-entrenado y posiblemente reconstruido completamente.

2.5. Minería de datos (Data mining) [21]

La fase de minería de datos es la fase característica del proceso del KDD y, por esta razón, muchas veces se utiliza esta fase para nombrar todo el proceso.

El objetivo de esta fase es producir nuevo conocimiento que pueda utilizarse.

Esto se realiza construyendo un modelo basado en los datos recopilados para este efecto. El modelo, es una descripción de los patrones y relaciones entre los datos que pueden usarse para hacer predicciones, para entender mejor los datos o para explicar situaciones pasadas.

Tipos de modelos

Los modelos pueden ser de dos tipos:

- Los modelos **predictivos**, pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo o dependientes, usando otras variables o campos de la base datos, que se denominan variables predictivas o independientes.
- Los modelos **descriptivos**, identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos.

Tareas de la minería de datos

Dentro de la minería de datos se ha de diferenciar tipos de tareas, cada una de las cuales puede considerarse como un tipo de problema a ser resuelto por un algoritmo de minería de datos. Esto significa que cada tarea tiene sus propios requisitos, y que el tipo de información obtenida con una tarea puede distinguirse mucho de la obtenida con otra.

Las distintas tareas pueden ser:

Predictivas

Clasificación

El objetivo del algoritmo es maximizar la razón de precisión de la clasificación de las nuevas instancias, la cual se calcula como cociente entre las predicciones correctas y el número total de predicciones (correctas e incorrectas).

Regresión

Consiste en aprender una función real que asigna a cada registro un valor real. El objetivo es minimizar el error (generalmente el error cuadrático medio) entre el valor predicho y el valor real.

Descriptivas

Agrupamiento (clustering)

Consiste en obtener grupos “naturales” a partir de los datos. Los datos son agrupados basándose en el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos.

Es decir, se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo, son muy diferentes a los objetos de otro grupo.

Al agrupamiento también se le suele segmentación, ya que parte o segmenta los datos en grupos que pueden ser o no disjuntos.

Correlaciones

Se usan para examinar el grado de similitud de los valores de dos variables numéricas.

Una fórmula estándar para medir la correlación es el coeficiente de correlación r, el cual es un valor comprendido entre -1 y 1. Si r es 1 (respectivamente -1), las variables están perfectamente correlacionadas (perfectamente correlacionadas negativamente), mientras si es 0 no hay correlación.

Es decir, que cuando r es positivo, las variables tienen un comportamiento similar (ambas crecen o decrecen al mismo tiempo), y cuando r es negativo, tienen un comportamiento opuesto (si una variable crece la otra decrece).

Reglas de asociación

Es una tarea muy similar a las correlaciones, y su objetivo se basa en identificar relaciones no explícitas entre atributos categóricos.

La formulación más común es del estilo “si el atributo X toma el valor d entonces el atributo Y toma el valor b”.

Las reglas de asociación no implican una relación causa-efecto, es decir, puede no existir una causa para que los datos estén asociados.

Reglas de asociación secuenciales

Es un caso especial de reglas de asociación, y se utiliza para determinar patrones secuenciales (ordenados) en los datos. Estos patrones se basan en secuencias temporales de acciones y se distinguen de las reglas de asociación en que las relaciones entre los datos se basan en el tiempo.

Técnicas de minería de datos

La minería de datos es un campo interdisciplinar, dónde existen diferentes modelos detrás de las técnicas utilizadas para esta fase: técnicas de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, aprendizaje basado en instancias, algoritmos genéticos, aprendizaje bayesiano, programación logística inductiva y varios tipos de métodos basados en núcleos, entre otros.

Cada uno de estos modelos incluye diferentes algoritmos y variaciones de los mismos, así como otro tipo de restricciones que hacen que la efectividad del algoritmo dependa del dominio de aplicación, no extiendo lo que podríamos llamar el método universal aplicable a todo tipo de aplicación.

Construcción del modelo

Es en la construcción del modelo dónde se observa mejor el carácter iterativo del proceso de KDD, ya que será necesario explorar modelos alternativos hasta encontrar aquel que resulte más útil para resolver nuestro problema.

Así, una vez obtenido un modelo y a partir de los resultados adquiridos para el mismo, se podría querer construir otro modelo usando la misma técnica, pero con otros parámetros, o quizás usar otras técnicas o herramientas. Por tanto, en la búsqueda del "perfecto modelo" se ha de retroceder a fases anteriores y realizar cambios o incluso modificar el objetivo del problema.

2.6. Sistemas de recomendación

Los sistemas recomendadores (SR) son algoritmos, modelos o técnicas utilizados para proporcionar a los usuarios finales sugerencias sobre contenidos, productos o servicios (tales como libros, películas, música, viajes, etc.). Los sistemas de recomendación estudian las características de cada usuario y mediante un procesamiento de los datos, a fin de encontrar un subconjunto de elementos que pueden resultar de interés para el usuario. Estas recomendaciones son personalizadas, diferentes usuarios o grupos reciben diferentes sugerencias.

Técnicas de SR [22]

Los sistemas de recomendación se pueden clasificar, dependiendo del tipo de información que utilizan para realizar sus recomendaciones, y los principales son:

SR basado en filtrado colaborativo (*Figura 2.10*)

El sistema recopila las calificaciones de los miembros en una comunidad y usa esa información para recomendar artículos a otros usuarios, "correlación entre personas". La idea subyacente es que diferentes personas tienen diferentes gustos y clasifican las cosas de manera diferente según esos gustos. Si dos usuarios valoran un conjunto de elementos de forma similar podemos deducir que tienen ideas similares, considerando que son vecinos. Este sistema, es la técnica más implementada en los sistemas de recomendaciones.

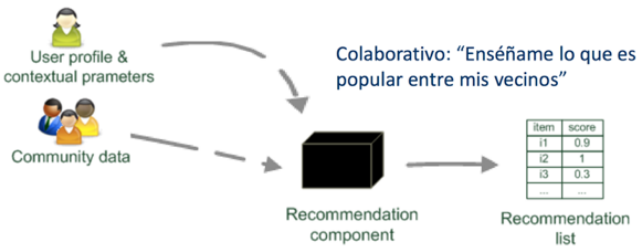


Figura 2.10. *SR basado en filtrado colaborativo*

SR con filtrado basado en contenido (*Figura 2.11*)

Las recomendaciones se generan a partir de dos fuentes: las características del producto /servicio y las calificaciones generadas por los usuarios. Por tanto, realizan las recomendaciones equiparando las preferencias del usuario (expresadas por éste de forma implícita o explícita) con las características utilizadas en la representación de los ítems, ignorando la información relativa de otros usuarios. En otras palabras, se le recomendará al usuario un ítem (servicio, producto o indicación) similar al que el mismo usuario haya elegido anteriormente.

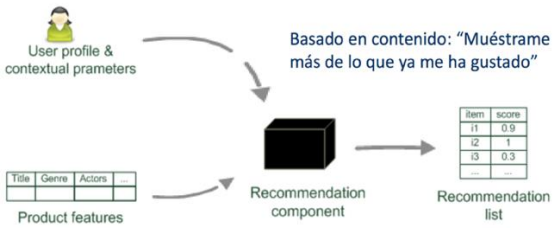


***Figura 2.11**. SR con filtrado basado en contenido.*

SR con filtrado basado en conocimiento (*Figura 2.12*)

Estos sistemas sugieren productos basados en inferencias de conocimiento de dominio sobre las preferencias de los usuarios. Una función de similitud calcula en que cantidad, las necesidades del usuario coinciden con las recomendaciones. Los sistemas basados en conocimiento, tienden a funcionar mejor al principio de la vida de los sistemas, pero necesitan ser provistos de técnicas de aprendizaje para asegurar su eficiencia.

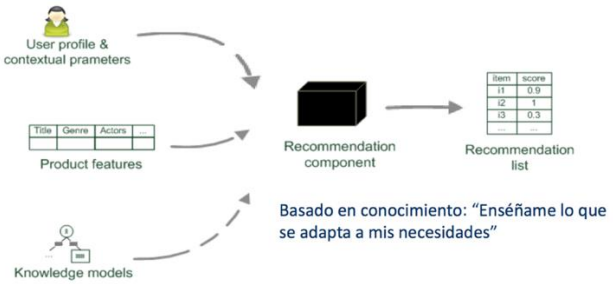


***Figura 2.12**. SR con filtrado basado en conocimiento.*

SR con métodos de filtrado híbrido (*Figura 2.13*)

Los sistemas híbridos se construyen combinando múltiples técnicas de recomendación para realizar recomendaciones e incluso se combinan con alguna otra técnica de inteligencia artificial como puede ser la lógica borrosa o la computación evolutiva.

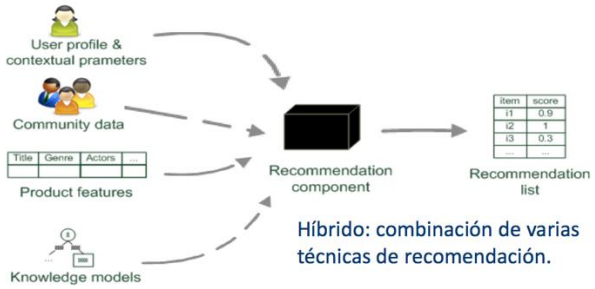


***Figura 2.13**. SR con métodos de filtrado híbrido.*

3. MATERIALES Y METODOLOGÍA

Una vez estructurado e introducido el presente proyecto, así como justificada su realización y los objetivos que se pretenden conseguir, a continuación, se detallan las herramientas que se han utilizado para su adquisición, así como la metodología empleada en su elaboración.

3.1. Herramientas

MySQL

MySQL, es la base de datos de código abierto de mayor aceptación mundial que ofrece una oferta económica de aplicaciones de bases de datos fiables, de alto rendimiento y fácilmente ampliables basadas en la web. Técnicamente, se dice que MySQL, es un sistema de administración de bases de datos relacionales de código abierto, el cual se basa en el lenguaje de consulta estructurado SQL. Se ha utilizado esta herramienta, debido a que Predircam se encuentra en esta base de datos.

Algunas de sus características más importantes son:

- Es de código abierto.
- Es de tipo relacional, lo que implica que maneja de manera eficiente y segura los datos de las distintas tablas que se están relacionadas entre sí. Utiliza código SQL.
- Es capaz de gestionar los datos de manera eficiente, sencilla y cómoda. ○ Permite recurrir a bases de datos multiusuario a través de la web y en diferentes lenguajes de programación que se adaptan a diferentes necesidades y requerimientos.
- Es conocida por desarrollar alta velocidad en la búsqueda de datos e información.
- Gestor de base de datos más utilizado por informáticos, desarrolladores y diseñadores web.

IBM SPSS Modeler ANEXOS

IBM SPSS Modeler es una plataforma de análisis predictiva diseñada para aportar inteligencia predictiva a decisiones llevadas a cabo por personas, grupos, sistemas y la empresa. Proporciona un rango de algoritmos y técnicas avanzados, incluidos el análisis de texto, el análisis de entidad, la gestión y optimización de decisiones, para ayudar a seleccionar las acciones que dan como resultado un mejor resultado. Disponible en varias ediciones, incluida una versión basada en la nube, SPSS Modeler puede escalar desde despliegues de escritorio a la integración dentro de los sistemas operativos. SPSS Modeler está diseñado para:

- Mejorar las decisiones y los resultados.
- Ayudar a extraer el valor de los datos.
- Integrarse de forma más sencilla en los sistemas existentes.

3.2. Metodología

En la ejecución de este proyecto, se han llevado a cabo dos etapas de desarrollo muy diferenciadas. En primer lugar, se ha realizado una primera etapa de investigación, prospección y formación para poder encontrar aquellas herramientas y elementos que permitieran dar respuesta de valor adecuada y satisfactoria a los diferentes objetivos planteados. Posteriormente, se ha llevado a cabo una segunda etapa en la que se han puesto en práctica aquellos conocimientos adquiridos y las herramientas seleccionadas, con el propósito de a partir de datos clínicos, obtener los modelos predictivos y las reglas de decisión a fin de desarrollar el diseño de un sistema de recomendación personalizado.

A pesar de tratarse de dos etapas consecutivas, debido a factores tales como el desconocimiento de las variables y valores clínicos, la falta de experiencia en minería de datos, así como el descubrimiento de nuevas necesidades y dificultades durante el trascurso del proyecto, han sido necesaria sucesivas iteraciones de estas dos fases con el fin de dar respuesta a cada una de estas complicaciones y eventualidades.

A continuación, se detallan de forma concisa y clara los puntos más significativos de estas dos etapas:

Investigación y formación

Con el objetivo de resolver, a través de las nuevas tecnologías, el problema planteado, y que esa respuesta fuera innovadora y aportara conocimiento nuevo, era necesaria una primera etapa de investigación, tanto a nivel del problema médico que se pretende minimizar, como a nivel tecnológico y de analítica predictiva.

Es relevante comentar que, al comienzo del proyecto, tan sólo se contaba con conocimiento básicos del significado de variables clínicas, base de datos y analítica predictiva, adquiridos en las asignaturas de biología, estadística y sistemas de información y comunicaciones en la sanidad, por lo que fue necesaria una profundización más exhaustiva para cada uno de estos ámbitos.

El primer paso fue conocer los módulos, funcionalidades y base de datos adquirida de la plataforma de Predircam. Esto, consistió en evaluar la problemática del síndrome cardiometabólico y los elementos claves para su prevención, a fin de entender, cada una de las variables clínicas que se registran en ella. Para la adquisición de este conocimiento, se consultaron diferentes estudios, artículos y publicaciones que son los que han permitido la realización del estado del arte analizado con anterioridad y cuyas referencias pueden encontrarse también en dicho epígrafe de este proyecto. Además, se contó con la ayuda de los expertos de la base de datos de Predircam.

Así pues, una vez adquirida la comprensión completa de la base de datos, fue necesario formarse en la analítica predictiva a fin de realizar un proceso de extracción de conocimiento, y para ello se estudiaron algunas de las herramientas que ofrece IBM.

Primeramente, se utilizó Watson Analytics, y a pesar de que es una herramienta que es utilizada para descubrir patrones y significado en los datos, limitaba mucho el estudio. Para poder llevar a cabo el análisis con ella, se han de cumplir unos requisitos muy específicos, referidos a la estructura de los datos, así como, no permitía la generación de un modelo predictivo.

Por tanto, se decidió utilizar IBM SPSS Modeler, que es una plataforma de analítica predictiva que proporciona diferentes algoritmos avanzados y técnicas de análisis. Por último, para la interpretación de los resultados, se estudiaron las diferentes técnicas de evaluación y se contó con la ayuda del equipo experto de la base de datos analizada.

Descripción del estudio clínico

Para evaluar la eficacia, con el uso de la nueva plataforma PREDIRCAM 2, de la intervención de la obesidad y la prevención del riesgo cardiovascular, en noviembre de 2015, se comenzó un estudio clínico, que aún está en marcha y cercano a su finalización. Este estudio comprende un periodo de 12 meses en el que, profesionales de tres hospitales (Hospital de la Santa Creu i Sant Pau – Barcelona, Hospital Universitari i Politècnic La Fe – Valencia y Hospital Universitario Virgen de la Victoria – Málaga) realizan a través de doce visitas programadas, el seguimiento y monitorización de pacientes tanto tecnológicos (utilizaban la plataforma), como no tecnológicos, obtenidos de un reclutamiento y selección previa

El objetivo de este programa, es la pérdida de peso y la mejora de la salud con ayuda de asistencia telemática continuada mediante recomendaciones estándares y globales basadas en las pautas de nutrición y ejercicio actuales, avaladas por la comunidad médica. Dando lugar a una base de datos en la que se almacenan todos los diferentes parámetros adquiridos de los pacientes, a lo largo de su progreso.

Criterios de inclusión y exclusión

Los requisitos de incorporación en el estudio son (*Figura 3.1 y Figura 3.2*):

Criterios de INCLUSIÓN

- **Capacidad aceptable de manejo tecnológico y dispositivo con acceso a Internet.**
- **Edad comprendida entre 25 y 65 años.**
- **Obesidad definida por un IMC ≥ 30 kg/m2:**
  - Grado I o II sin síndrome de apnea del sueño ni cardiopatía isquemia.
  - Ausencia de dislipemia o hipertensión arterial que precisen de tratamiento farmacológico.

***Figura 3.1**. Criterios de inclusión*

Criterios de EXCLUSIÓN

- **Enfermedad grave y/o incapacitante.**
- **Sujetos con criterios definitorios de diabetes según ADA.**
- **Sujetos con patología moderada-grave en el momento de inclusión.**
- **Tratamiento que interfiera con el metabolismo glucídico.**
- **Presencia de enfermedad cardiovascular.**
- **Dislipemia o HTA que precisen de tratamiento farmacológico.**
- **Consumo activo de drogas o enolismo moderado-severo.**
- **Gestación**

***Figura 3.2.** Criterios de exclusión.*

Procedimiento

El estudio consta de doce visitas que se espacian en el tiempo, las primeras visitas (1,2 y 3) se basan en la selección de los pacientes y determinación de forma aleatoria de quiénes serán tecnológicos y entonces, llevarán a cabo una intervención tecnológica mediante el uso de la plataforma Predircam y quiénes no.

La visita 4 corresponde a la semana 1, ya que es en esta en la que se realiza la primera prescripción por parte del profesional médico, el cual establece unas pautas nutricionales y de actividad física que el paciente debe cumplir durante el periodo comprendido entre esta y la siguiente. Por tanto, es a partir de la visita 4 desde donde se comienzan a contar los 12 meses, es decir, las 48 semanas de tratamiento.

Los pacientes tecnológicos pueden hacer uso de la plataforma durante todo este periodo, registrando sus ingestas y su actividad física, así como sus parámetros corporales, y todo ello con la asistencia de un sistema de notificaciones. Así cómo, algunas de las visitas (6,8,9 y 11) de estos pacientes, los tecnológicos, son telemáticas, tal y como se observa en la *Figura 3.3*, es decir, son ellos mismos los que recogen los valores requeridos y se los envían al médico mediante el módulo de mensajería de la plataforma.

De esta manera, en cada visita, el profesional médico puede valorar a través de los reportes de nutrición y actividad física realizada, que le proporciona la plataforma sobre el paciente, y de las medidas directas que se obtienen durante la visita (valores antropométricos y/o analítica), si se han seguido las pautas acordadas y si éstas deben ser modificadas para una futura mejora o, por el contrario, deben mantenerse y continuar con el progreso.

Por último, los profesionales médicos se encargan, tanto de la gestión de los usuarios y su acceso a la plataforma, como de la gestión y creación de los contenidos de la misma.

Figura 3.3. *Procedimiento y variables de estudio en cada una de las visitas*

Base de Datos

El subset obtenido de la base de datos, está compuesto por aquellos pacientes que han hecho uso de la plataforma de Predircam, y se ha empleado para el desarrollo de este proyecto, dispone de 5 entidades, detalladas a continuación:

- **Entidad 1:** Visitas (visita 1 a la 10). En esta tabla, se encuentran los registros asociados a cada uno de los campos medidos correspondientes a cada visita.
- **Entidad 2:** Ejercicio. En esta tabla, se encuentran los numerosos registros de ejercicio correspondientes a cada una de las visitas de los diferentes usuarios.
- **Entidad 3:** Nutrición. En esta tabla, se encuentran los numerosos registros de nutrición, correspondientes a cada una de las visitas de los diferentes usuarios.
- **Entidad 4:** Peso_cintura_cadera. En esta tabla, se encuentran los diferentes valores de peso, cintura y cadera correspondientes a cada uno de los usuarios.
- **Entidad 5:** Notificaciones. En esta tabla están registradas cada una de las notificaciones que se le ha enviado a cada uno de los usuarios, a través de la plataforma, a lo largo de todo el programa.

Cada usuario realiza N visitas, por lo que para cada entidad de visita la relación existente con las otras entidades, es de 1: N, tal y como se muestra en la *Figura 3.4*.

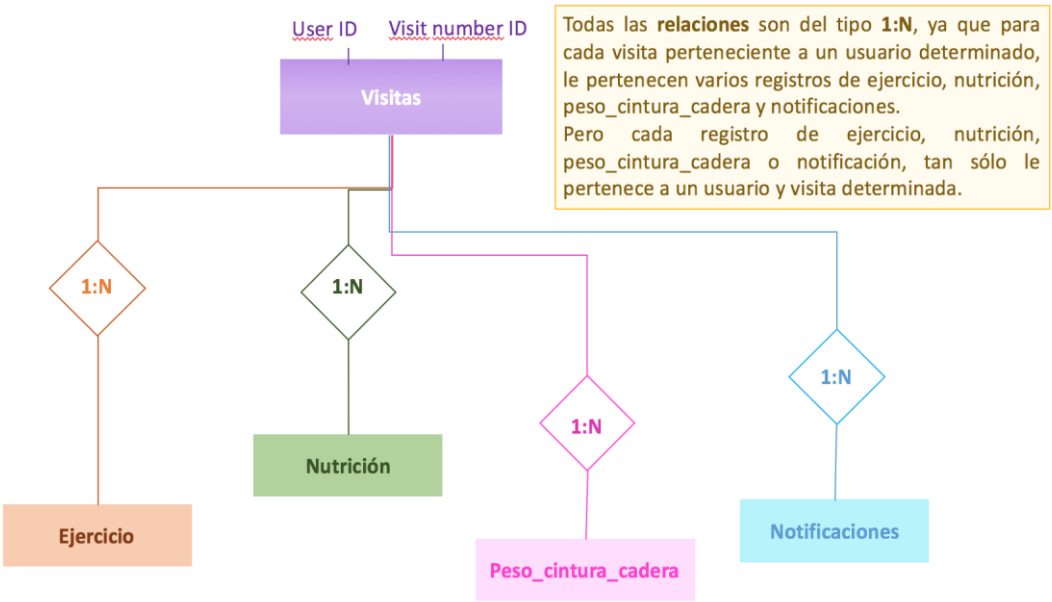


Figura 3.4. *Relaciones entre las entidades de la base de datos de PREDIRCAM.*

Análisis y modelado

Una vez obtenida la base de datos del estudio clínico, se llevó a cabo un completo proceso de extracción de conocimiento a partir de estos (KDD). Para ello, se ha utilizado la herramienta de IBM SPSS Modeler, que dispone de una metodología completa de KDD a fin de ordenar las tareas de la minería de datos: el modelo CRISP-DM (Cross-Industry Standard Process for Data Mining) mostrado en la *Figura 3.5.*

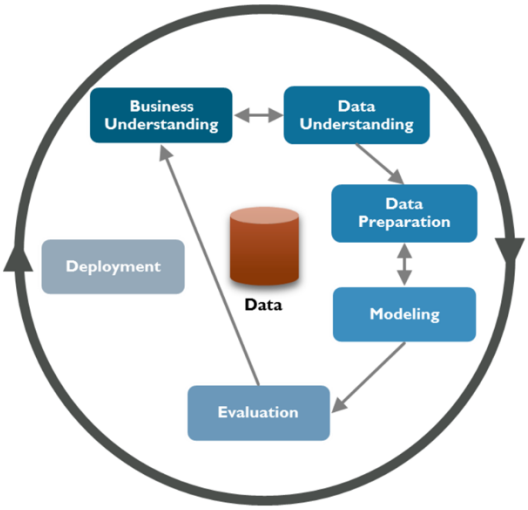


Figura 3.5. *Proceso KDD con la metodología CRISP-DM.*

Por tanto, se lleva a cabo el siguiente proceso cíclico CRISP-DM:

1. Determinación del propósito del análisis

El propósito de analizar la base de datos, consiste en determinar que variables de nutrición y ejercicio de los 3 primeros meses (registros de las vistas 4,5,6,7), son las más significativas en la pérdida de peso a los 6 meses (visita 10).

Para ello, se determina que aquellos pacientes que han conseguido disminuir su peso un 5% al cabo de los 6 meses, son los que han logrado el objetivo del programa. Este valor está basado en análisis previos y, se ha determinado, junto con los médicos, que el 5% es un valor significativo para poder estudiar las diferencias de las variables entre aquellos pacientes que si han logrado el éxito y los que no.

A partir de aquí, en cada una de las tablas se genera el campo *Purpose_Objetive_yes_no*, con los valores:

- 0, no ha logrado bajar un 5% su peso a los 6 meses.
- 1, si ha logrado bajar un 5% su peso a los 6 meses

2. Comprensión de los datos

Para esta investigación, de entre los 183 pacientes que han participado en el estudio, el subconjunto seleccionado está formado por aquellos sujetos que son tecnológicos y han completado un seguimiento de 6 meses.

Por lo tanto, los sujetos analizados son 43, siendo el 88.37 % mujeres y 11.62 % hombres. Así como, el 53.48 % de los pacientes no ha logrado el éxito mientras que el 46.52% sí lo ha conseguido.

Y con la finalidad de evaluar las variables de nutrición y ejercicio, se seleccionan las siguientes tablas:

- La **tabla de nutrición** *(Tabla 3.1*) abarca dos tipos de registros: las ingestas durante la semana y las realizadas durante el fin de semana, así como varios campos:

Tabla 3.1. *Tabla de nutrición.*

Tabla	Registros	Campos
Nutrición	Registros de nutrición que abarcan desde la visita 4 a la visita 10, siendo cada registro un día de ingesta completo, de un paciente determinado. Existiendo numerosos registros correspondientes a un mismo usuario.	✓ **Personales**: edad, sexo, hábitos alimenticios, antecedentes. ✓ **Purpose_Objetive_yes_no:** valores de 1 y 0. ✓ **Kcal prescriptas y Kcal consumidas** ✓ **Grupos principales**: cereales, legumbres y tubérculos. Carnes, embutidos, pescados, mariscos y huevo. Verduras y hortalizas. ✓ **Macronutrientes**: carbohidratos, proteínas y total de grasas (buenas y malas). ✓ **Bebidas**: saludables, no saludables y alcoholes ✓ **Lácteos** ✓ **Frutas** ✓ **Aceites y frutos secos** ✓ **Alimentos no saludables**

- **La tabla de ejercicio** *(Tabla 3.2),* que abarca dos tipos de registros: los insertados por los usuarios manualmente a través de la plataforma y los insertados mediante el uso de un reloj, el Polar RS400, y de una banda de detección del pulso cardíaco que se conecta con este mediante un Bluetooh, así como varios campos:

Tabla 3.2. *Tabla de ejercicio.*

Tabla	Registros	Campos
Ejercicio	Registros de ejercicio que abarcan desde la visita 4 a la visita 10, siendo cada registro una actividad física realizada, de un paciente determinado. Habiendo numerosos registros correspondientes a un mismo usuario.	✓ **Personales**: edad, sexo, hábitos alimenticios, antecedentes. ✓ **Purpose_Objetive_yes_no:** valores de 0 y 1. ✓ **Tipo de fuente de adquisición:** a partir de Reloj Polar (valor = 1) o manualmente realizada por el usuario (valor = 0). ✓ **Kcal prescriptas y quemadas**. ✓ **Duración prescripta de actividad y realizada en minutos.** ✓ **Frecuencia cardíaca:** máxima, media y mínima.

Por último, para una comprensión integra de las tablas expuestas, cabe comentar que en ellas los datos de nutrición y ejercicio, van asociados a una determinada visita, por tanto, los valores pertenecen al periodo que abarca desde la visita a la que corresponde hasta la siguiente. Es decir, por ejemplo, todos los registros realizados en el periodo comprendido entre la visita 6 y 7, corresponden a la visita 6.

3. Preparación de los datos

Una vez determinadas las tablas utilizadas para el análisis, se ha llevado a cabo una preparación de los datos que consiste en la recopilación, limpieza, selección y transformación de estos mediante los nodos de la herramienta de IBM SPSS Modeler, destinados a estas tareas.

Primeramente, debido a que los registros correspondientes a la visita 4 son aquellos que pertenecen al periodo de la 4 a la 5, es decir, la primera semana de tratamiento, en la cual los usuarios prueban la plataforma, estos registros no se incluyen en dicho análisis.

Posteriormente se preparan, de forma específica, cada una de las tablas:

Nutrición

A partir de la tabla de nutrición, en la que se encuentran todos los registros correspondientes a todas las visitas, los pasos principales que se han llevado a cabo son los siguientes:

1. Se seleccionan de la tabla, los registros pertenecientes a cada visita de interés, y se obtiene así, una tabla independiente de cada visita. En este caso se obtienen 3 tablas de nutrición correspondientes a las visitas 5,6 y 7.
2. En cada una de las tablas obtenidas, se inserta un campo nuevo, que indica si el registro pertenece a un día entre semana, es decir, de lunes a viernes; o un día en fin de semana. Se crea así el campo denominado *Registro fin de semana*, el cual tiene 2 valores: T (true) si se ha realizado en fin de semana, y F (false), no se ha realizado en fin de semana.
3. Se agrupan los registros mediante un método de agrupación por campos claves, los cuáles son el campo *UserID*, que se refiere al identificador del paciente y el campo de *Registro fin de semana* que indica si el registro pertenece al fin de semana o no.

   Por tanto, todos los valores con el mismo ID y mismo valor de registro de fin de semana se agrupan, calculando el valor medio, para los demás campos, entre los valores de los registros que se agrupan. Así cómo, se genera un campo denominado *Cantidad de registros*, que indica la cantidad de registros que se han agrupado, para la obtención de esos valores medios.

En consecuencia, para cada una de las visitas, de nutrición, se obtiene una tabla independiente como la *Tabla 3.3*, en la que se observa que para cada visita y usuario quedan 2 filas, una con la media de todos los valores de entre semana y otra con los valores medios de todos los registros realizados en fin de semana.

***Tabla 3.3**. Tabla de nutrición tras proceso de selección.*

UserID	visit_number	Registro fin de semana	Cantidad de registros	proteins_Mean
16	5	F	4	68
19	5	F	6	60
19	5	T	3	58
27	5	T	4	56
27	5	F	5	69
35	5	T	6	83
35	5	F	12	84
38	5	T	6	81
38	5	F	13	82
40	5	T	2	81
40	5	F	5	116

A continuación, a partir de esta tabla, a fin de determinar los hábitos característicos de fin de semana se realizan 3 análisis independientes: uno para la visita 5, otro para la 6 y otro para la 7.

Y en cada uno de estos análisis, se ejecutan 3 tipos de estudios:

1. Estudio entre semana, consiste en analizar los registros realizados entre semana, es decir, de lunes a viernes. A este efecto, se genera una tabla seleccionando del campo llamado registro fin de semana, aquellos registros con valor F.
2. Estudio de fin de semana, consiste en analizar los registros en fin de semana, es decir, sábado y domingo. Para esto, se genera una tabla seleccionando del campo llamado registro fin de semana, aquellos registros con valor T.

3. Estudio semanal de todos los registros juntos, es decir, todos los registros tomados de lunes a domingo. Para esto, se genera una agrupación por campo clave, en este caso el ID de usuario, es decir, se agrupan todos los registros con el mismo ID, y se calcula la media entre los valores de los demás campos.

Ejercicio

A partir de la tabla de ejercicio físico, en la que se encuentran todos los registros correspondientes a todas las visitas, los pasos principales que se han llevado a cabo son los siguientes:

1. Se seleccionan de la tabla los registros pertenecientes a cada visita de interés, y se obtiene así, una tabla de cada visita. En este caso se obtienen 3 tablas de ejercicio correspondientes a las visitas 5,6 y 7.

2. En cada una de las tablas, se agrupan los registros mediante un método de agrupación por campos claves, los cuáles son el campo *UserID*, que se refiere al identificador del paciente y el campo *Source Type* que indica si el registro es realizado manualmente por el usuario (valor 0) o proviene del Reloj Polar RS400 (valor 1).

   Por tanto, todos los valores con el mismo ID y mismo valor de tipo de fuente se agrupan, y con los demás campos, se calcula el valor medio, entre los valores de estos registros. Pero, sin embargo, para los campos de *kcal quemadas* y *duración en minutos*, se realiza la suma de todos los valores de los registros agrupados.
   Así cómo, se genera un campo denominado *Cantidad de registros*, que indica la cantidad de registros que se han agrupado, para la obtención de esos valores medios.

En consecuencia, para cada una de las visitas, de ejercicio, se obtiene una tabla independiente como la *Tabla 3.4*, en la que se observa que para cada visita y usuario quedan 2 filas, una correspondiente a los registros de actividad manual y otra con los valores de los registros obtenidos con el reloj.

***Tabla 3.4**. Tabla de ejercicio con datos seleccionados*

UserID	Visit NumberID	Source Type	kcal burned_Sum	Duration in Minutes_Sum	Cantidad de registros
16	5	0	431.200	100	3
19	5	0	110.250	30	1
19	5	1	180.000	39	1
27	5	0	421.400	80	1
27	5	1	1770.000	351	5
35	5	1	7465.000	521	9
38	5	0	436.380	109	2
38	5	1	5478.000	1242	17
40	5	0	6552.530	1310	20
40	5	1	2964.000	414	6

UserID	Visit NumberID	Source Type	kcal burned_Sum	Duration in Minutes_Sum	Cantidad de registros
16	6	0	2791.290	639	11
16	6	1	1425.000	143	3
19	6	1	2065.000	593	10
19	6	0	1688.670	440	11
27	6	1	5701.000	1000	15
27	6	0	1051.790	216	3
35	6	1	1726.000	126	3
38	6	1	10050.000	2334	29
40	6	1	4671.000	695	9
40	6	0	17913.180	3160	48

Seguidamente, a partir de las tablas obtenidas, se observa, que los valores de los campos de *kcal, duración y cantidad de registros*, representan la suma total de todos los valores de los registros realizados según el tipo de fuente de obtención. Pero, debido a que las *prescripciones de kcal, duración y registros* se elaboran semanalmente, interesa obtener estos valores en el mismo formato.

Para ello, se dividen los valores de estos campos, entre el periodo en semanas, es decir, la cantidad de semanas a las que corresponden. Por tanto:

1. La tabla correspondiente a la visita 5 se dividen los valores entre 2, ya que corresponden al periodo de la visita 5 a la 6 que abarca 2 semanas. Así pues, se divide cada uno de estos campos entre 2 y se obtiene el valor medio semanal de kcal quemadas, duración en minutos de actividad física y cantidad de registros realizados.
2. La tabla correspondiente a la visita 6 se dividen los valores entre 4, ya que corresponden al periodo de la visita 6 a la 7 que abarca 4 semanas. Así pues, se divide cada uno de estos campos entre 4 y se obtiene el valor medio semanal de kcal quemadas, duración en minutos de actividad física y cantidad de registros realizados.
3. La tabla correspondiente a la visita 7 se dividen los valores entre 4, ya que corresponden al periodo de la visita 7 a la 8 que abarca 4 semanas. Así pues, se divide cada uno de estos campos entre 4 y se obtiene el valor medio semanal de kcal quemadas, duración en minutos de actividad física y cantidad de registros realizados.

A continuación, a fin de poder realizar un correcto análisis se agrupan los registros con el mismo *UserID,* para ello:

1. Se realiza un campo de kcal medias quemadas semanalmente ingresadas manualmente y otro para las registradas mediante el reloj, y a partir de la suma de estos dos campos, se obtiene un nuevo campo que corresponde a la media de kcal quemadas semanalmente y se denomina *kcal_burned_semanal_media_Sum*.
2. Así como, se realiza un campo de duración media invertida en ejercicio semanal en minutos de los registros ingresados manualmente y otro para los ingresados mediante el reloj, y a partir de la suma de estos dos campos, se obtiene un nuevo campo que corresponde a la duración media de actividad física semanal y se *denomina duración_semanal_media_Sum*
3. También se hace un campo de cantidad de registros semanales ingresados manualmente y otro para los registrados mediante el reloj, y a partir de la suma de estos dos campos, se obtiene un nuevo campo que corresponde, a la media de los registros semanales realizados y se denomina *resgristos_semanales_media_Sum.*

Finalmente, para cada una de las visitas, se obtiene una tabla diferente, pero todas con el formato de la *Tabla 3.5*.

Tabla 3.5. *Tabla de ejercicio tras proceso de selección y agrupación.*

UserID	Visit NumberID	kcal_burned_semanal_media_Sum	kcal_burned_semanal_manual_Sum	kcal_burned_semanal_reloj_Sum
16	5	215.600	215.600	0.000
19	5	145.125	55.125	90.000
27	5	1095.700	210.700	885.000
35	5	3732.500	0.000	3732.500
38	5	2957.190	218.190	2739.000
40	5	4758.265	3276.265	1482.000

duración_minutos_semanal_media_Sum	duración_semanal_media_manual_Sum	duración_semanal_media_reloj_Sum
50.000	50.000	0.000
34.500	15.000	19.500
215.500	40.000	175.500
260.500	0.000	260.500
675.500	54.500	621.000
862.000	655.000	207.000

registros_semanales_media_Sum	registros_semanal_media_manual_Sum	registros_semanal_media_reloj_Sum
1.500	1.500	0.000
1.000	0.500	0.500
3.000	0.500	2.500
4.500	0.000	4.500
9.500	1.000	8.500
13.000	10.000	3.000

4. Modelado

Una vez preparados los datos y obtenidas las tablas de nutrición y ejercicio limpias y seleccionadas, con los registros de interés, se procede a la fase de minería de datos para cada una de las tablas que se examinan. Esta fase consiste en la utilización de sofisticados métodos de análisis a fin de extraer nuevo conocimiento de los datos, el cual se pueda utilizar. Para esto, se realiza un modelo, que es un conjunto de reglas, fórmulas o ecuaciones que se emplea para predecir un resultado basándose en un conjunto de campos o variables de entrada.

Y, para el desarrollo del modelo, se utilizan los diferentes nodos de los que dispone la herramienta IBM SPSS Modeler en la paleta de modelado.

Proceso de modelado

El proceso de modelado que se lleva a cabo para cada una de las tablas, mediante esta herramienta, consta de 3 etapas:

3. **Generación de la ruta** *(Figura 3.6)***.** Para crear la ruta que genere el modelo, los 3 elementos fundamentales son:
    - Un nodo origen que lee los datos de las tablas de nutrición y ejercicio.
    - Un nodo Tipo que especifica propiedades de campo, como el nivel de medición (el tipo de datos que contiene el campo) y el rol de cada campo como objetivo o entrada en modelado. En este caso, todos los campos son de entrada a excepción del campo *Purpose_Objective_yes_no* que es el campo objetivo, ya que es, el que se quiere predecir con el modelo.
    - Un nodo modelado que genera un nugget de modelo cuando se ejecuta la ruta.

Figura 3.6. *Proceso de modelado*

4. Exploración del modelo

Cuando se finaliza la ejecución del modelo, se añade el nugget de modelo (*Figura 3.7*) a la pestaña de Modelos y al lienzo. Este es el recipiente de un modelo, es decir, es el conjunto de reglas, fórmulas o ecuaciones que representan los resultados de las operaciones de generación de modelos, y cuya finalidad es puntuar datos para generar predicciones o permitir análisis adicionales de propiedades de modelos.

Por tanto, esta fase, consiste en abrir el nugget de modelo en la pantalla, y analizar los resultados obtenidos: la importancia relativa de los campos de entrada en la creación del modelo, el conjunto de reglas, patrones, etc.

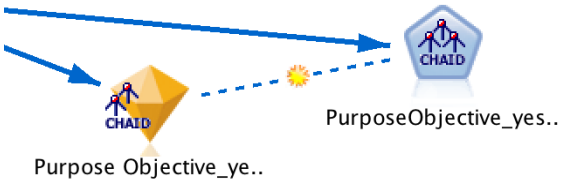


Figura 3.7*. Nugget del modelo*

5. Evaluación del modelo

En esta etapa se evalúa, con que precisión trabaja el modelo, para ello se debe de puntuar, con el nugget de modelo generado, un conjunto de registros diferentes de los utilizados para crear el modelo, y comparar las respuestas predichas por el modelo, con los resultados reales.

Esta comparación se ha realizado con el nodo análisis (*Figura 3.8*), que efectúa comparaciones entre los valores predichos por el modelo y los reales, permitiendo evaluar la capacidad de un modelo para generar predicciones precisas, y determinando así las variables más significativas.

También, se conecta al nugget del modelo a un nodo tabla (*Figura 3.8*), que muestra los valores predichos, en este caso en un campo denominado $R-*Purpose_Objective_yes_no*, generados por el model

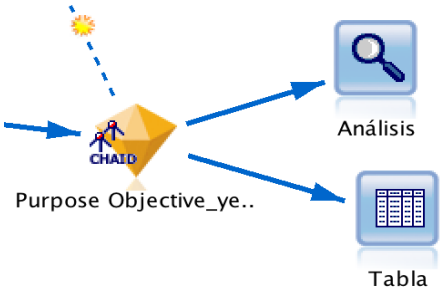


***Figura 3.8**. Conexión con el nodo análisis y el nodo análisis.*

Partición de los datos

Por último, para dividir los datos en muestras separadas para el entrenamiento y la comprobación del modelo, se ha utilizado un nodo Partición, que separa los datos en subconjuntos para cada una de las fases.

Por tanto, a pesar de que se dispone de una base de datos con pocos registros, en cada una de las tablas a analizar, se ha empleado una muestra para generar el modelo y otra muestra distinta para probarlo, a fin de obtener una primera indicación de la bondad del modelo a la hora de generalizarlo a otros conjuntos de datos.

La configuración empleada para cada una de las tablas es la que se muestra en la *Figura 3.9*.

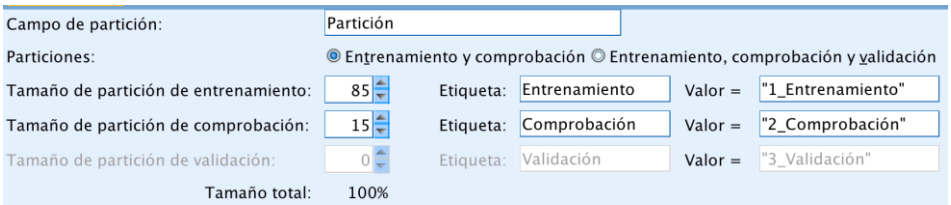


***Figura 3.9.** Configuración para partición de los datos*

Finalmente, la ruta generada para cada una de las tablas de nutrición y ejercicio, es la que se observa en la *Figura 3.10*:

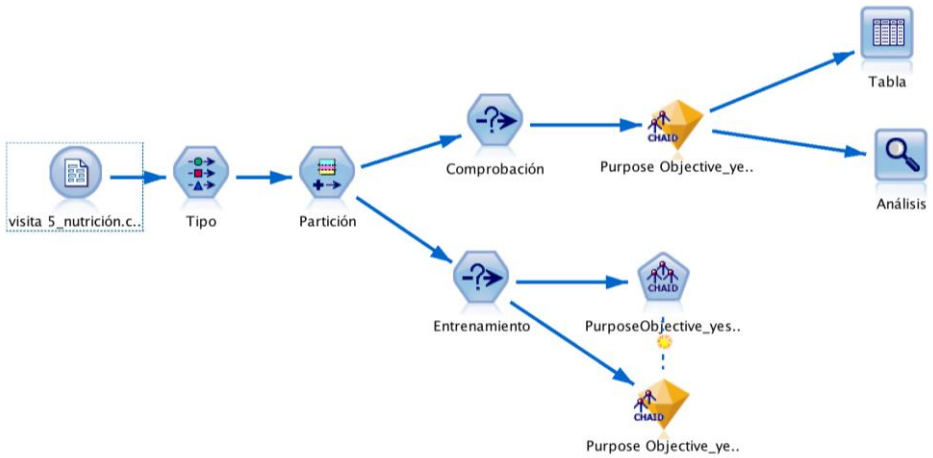


***Figura 3.10**. Proceso de modelado con partición de los datos*

4. MODELOS

En este apartado, se exponen los diferentes procesos de modelado que se han realizado a partir de cada una de las tablas obtenidas del apartado 3, con el propósito, de determinar las variables más significativas de nutrición y ejercicio, a fin de conseguir éxito en el programa de pérdida de peso. Y, tal y como se ha comentado con anterioridad, los sujetos a analizar son 43, pero, se ha de tener en cuenta que debido a que no todos los usuarios han registrado nutrición y ejercicio todas las semanas del tratamiento, muchas de las tablas a examinar se encuentran con una menor cantidad de registros que los correspondientes.

Tras probar y valorar diferentes modelos, se han seleccionado aquellos que mejor se ajustaban a los datos y a partir de los cuales se pretende generar un conjunto de reglas.

Los modelos que se han utilizado para el desarrollo del análisis son los siguientes:

Modelos de clasificación

Estos modelos se han utilizado para **predecir**, ya que usan el valor de uno o más campos de entrada (valores de nutrición y ejercicio) para predecir el valor del resultado o campo de destino (*Purpose_Objective_yes_no*).

CHAID

Este método se ha seleccionado para generar árboles de decisión, los cuales permiten una descripción y comprensión sencilla, e identificar divisiones óptimas, mediante estadísticos de chi-cuadrado. En primer lugar, se examinan, las tablas de tabulación cruzada entre los campos de entrada y los resultados, para, a continuación, comprobar la significación mediante una comprobación de independencia de chi-cuadrado. Si varias de estas relaciones son estadísticamente importantes, entonces, se selecciona el campo de entrada de mayor relevancia (el valor P más pequeño).

C5.0

El algoritmo C5.0 genera árboles de decisión o conjunto de reglas, y dispone de la opción de validación cruzada, que resulta de gran utilidad cuando el conjunto de datos es demasiado pequeño. Es por ello que se ha utilizado este algoritmo, debido a que esta opción, ha evitado dividir la muestra en dos subconjuntos, lo que ha permitido obtener modelos con una alta predicción.

Lista de decisiones

El modelo de lista de decisiones se ha empleado para identificar los subgrupos, o segmentos, que muestran una mayor o menor posibilidad de conseguir el éxito en el programa.

Selección de características

La aplicación de la selección de características, ha permitido filtrar y eliminar los campos de entrada en función de un conjunto de criterios, así como, se emplea para clasificar el grado de importancia del resto de entradas de acuerdo con el objetivo específico (*Purpose_Objective_yes_no)*

La regresión logística

Esta técnica de estadístico se ha utilizado para clasificar los registros a partir de los valores de los campos de entrada, y se basa en crear un conjunto de ecuaciones que relacionan los valores de los campos de entrada con las probabilidades, a cada una de las categorías de los campos de salida.

Red bayesiana

La red bayesiana se ha ejecutado a fin de crear un modelo de probabilidad que combina las pruebas observadas y registradas con conocimiento del mundo real para así, establecer la probabilidad de instancias. Existen dos tipos: las redes de Naïve Bayes para aumentar a ábol (TAN) y de manto de Markov que se utilizan para clasificar. Es un modelo gráfico y se ha utilizado a fin de observar variables (nodos) en un conjunto de datos y las independencias probabilísticas o condicionales entre ellas.

Modelos de segmentación.

Estos modelos, se han empleado a fin de **dividir** los datos en segmentos o clústeres de registros que tienen patrones similares de campos de entrada.

Bietápico

Es un método que se ha empleado en la agrupación de clústeres, debido a que tienen la ventaja de estimar automáticamente el número óptimo de clústeres para los datos de entrenamiento. Y se basa en agrupar los registros de manera que los de un mismo grupo o clúster tiendan a ser similares entre ellos, y que los de otros grupos sean distintos.

5.1. Modelos de nutrición

A partir de las tres tablas de nutrición obtenidas en la fase de preparación de los datos, correspondientes a cada una de las visitas: 5,6,7, se realiza un proceso de modelado para cada una de ellas. Asimismo, se realizan 3 estudios diferentes para determinar que variables son más significativas el fin de semana y cuáles son de forma generalizada.

Modelo 1.

Para la generación del modelo 1, se ha utilizado la tabla de nutrición correspondiente a la **visita 5 con valores del fin de semana**, que abarca los registros comprendidos entre la visita 5 y 6, periodo de dos semanas.

1. 1. Árbol de decisión: Algoritmo C5.0

Tras aplicar el algoritmo C5.0, se obtiene el gráfico Importancia del predictor *(Figura 4.1)*, en el que se observa que el porcentaje de grasas buenas consumidas en el fin de semana, es lo más significativo, seguidamente del grupo de edad y de si el sujeto posee alguna enfermedad en tratamiento (hipertensión arterial y/o dislipidemia en tratamiento con dieta, diabetes mellitus y /o neoplasias tipo 2 en tratamiento u otras enfermedades bajo tratamiento).

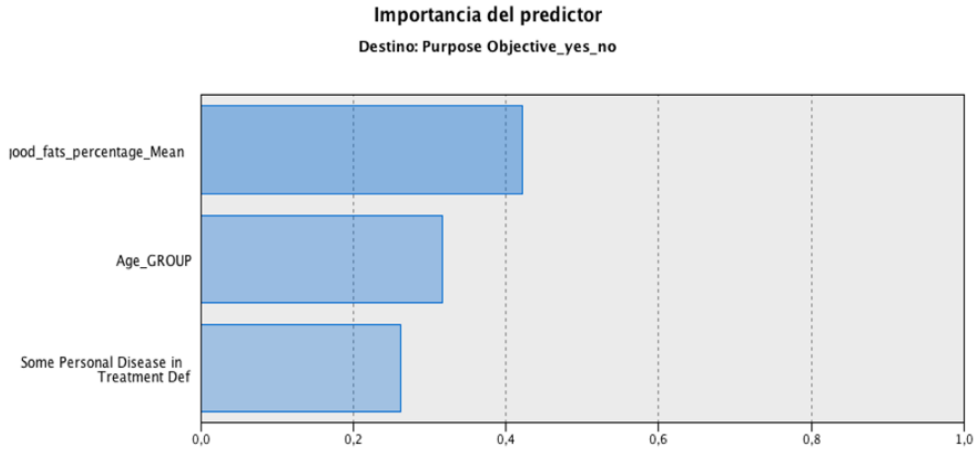


Figura 4.1. *Modelo 1.1. Importancia del predictor*

También, se adquieren los detalles en forma de conjunto de reglas (*Figura 4.2*) del modelo junto con sus porcentajes de confianza:

- Regla 1 – precisión estimada 75,76% [aumentar 50%]
  - good_fats_percentage_Mean <= 61 [Moda: 0] ⇨ **0**
  - good_fats_percentage_Mean > 61 [Moda: 1] ⇨ **1**
- Regla 2 – precisión estimada 63,05% [aumentar 50%]
  - Some Personal Disease in Treatment Def = 1 [Moda: 0] ⇨ **0**
  - Some Personal Disease in Treatment Def = 0 [Moda: 1] ⇨ **1**
- Regla 3 – precisión estimada 64,72% [aumentar 50%]
  - Age_GROUP <= 3 [Moda: 0] ⇨ **0**
  - Age_GROUP > 3 [Moda: 1] ⇨ **1**

Figura 4.2*. Modelo 1.1. Conjunto de reglas*

Asimismo, se muestra el modelo en forma de árbol para cada una de estas reglas, en el que se contempla la cantidad de registros que conforman cada una de ellas. A partir de las *Figura 4.3, Figura 4.4 y Figura 4.5,* se contempla que aquellos sujetos que consiguen el éxito, mayoritariamente, ingieren más del 61% de grasas saludables, no tienen ninguna enfermedad personal y pertenecen al grupo de edad 4 o 5.

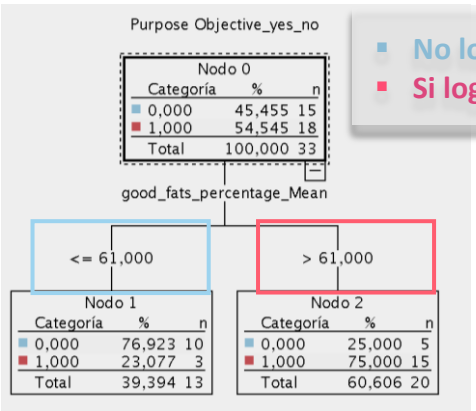


Figura 4.3*. Modelo 1.1. Regla 1.*

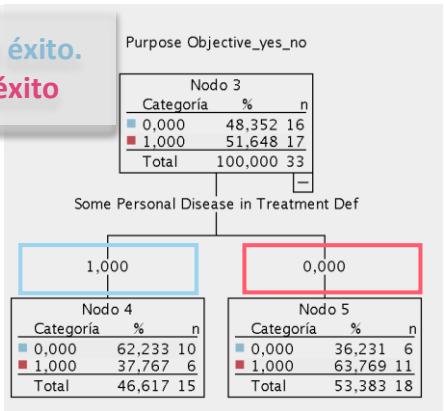


Figura 4.4. *Modelo 1.1. Regla 2.*

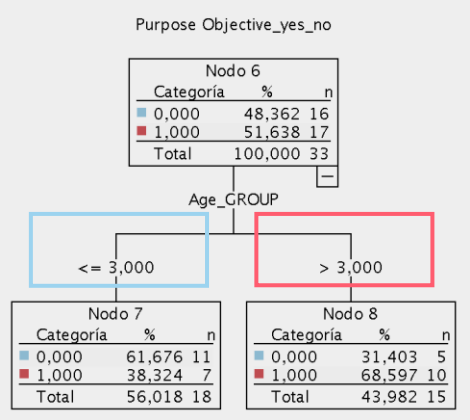


Figura 4.5. *Modelo 1.1. Regla 3.*

Los valores del campo *Age_GROUP,* abarcan las siguientes edades:

1 = $24 \leq edad < 30$

2 = $30 \leq edad < 40$

3 = $40 \leq edad < 50$

4 = $50 \leq edad < 56$

5 = $56 \leq edad \leq 65$

Figura 4.6*. Modelo 1.1. Clasificación grupos de edad.*

Evaluación

Al emplear el algoritmo C.5, con la opción de validación cruzada, no se dividen los datos en subconjuntos de entrenamiento y comprobación, lo que permite obtener un modelo con una precisión alta. Así pues, el análisis de este modelo muestra que, para 27 de un total de 33 registros (el 81.82 %), coinciden los valores predichos con los reales.

Por lo tanto, este modelo predecirá el éxito, basándose en las reglas anteriores, con un porcentaje del 81.82 % (*Tabla 4.1*).

Tabla 4.1. *Modelo 1.1. Evaluación.*

Resultados para el campo de resultado Purpose Objective_yes_no

Comparando $C-Purpose Objective_yes_no con Purpose Objective_yes_no

Correctos	27	81,82%
Erróneos	6	18,18%
Total	33	

1.2. Segmentación: Bietápico

A fin de determinar la relación entre las variables obtenidas en el modelo 1.1, se realiza un método de análisis de clústeres a partir de los campos: *Purpose Objective_yes_no, good_fats_percentage_Mean y Age_GROUP*, y se obtienen dos clústeres con una medida de silueta de cohesión y separación de 0.5 (*Figura 4.7*). El campo *Some Personal Disease in Treatment Def*, se ha excluido, ya que al introducirlo el valor de medida de silueta era muy bajo.

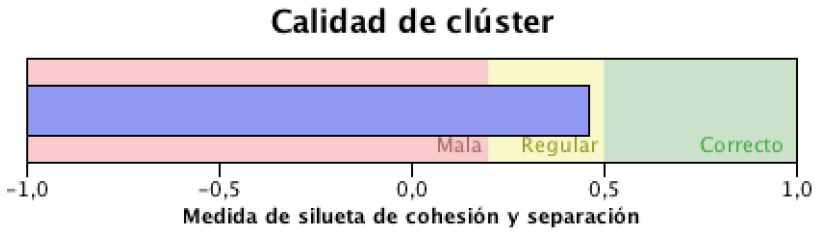


***Figura 4.7**. Modelo 1.2. Medida de silueta de cohesión y separación.*

A pesar de que el valor obtenido 0.5, se encuentra en la ventana regular, al examinar los clústeres, se comprueba que existe una estructura de conglomerados muy evidente. Concretamente, tal y como se contempla en la *Figura 4.8*, la variable objetivo y porcentaje de grasas buenas consumidas, son importantes predictores. Así pues, en la *Figura 4.10*, se observa que aquellos sujetos que no han logrado el éxito son los que han ingerido menor cantidad de grasas saludables y pertenecen al grupo de edad de los más jóvenes. Mientras que los que si han logrado, sucede lo contrario, tal y como se aprecia en la *Figura 4.9*.

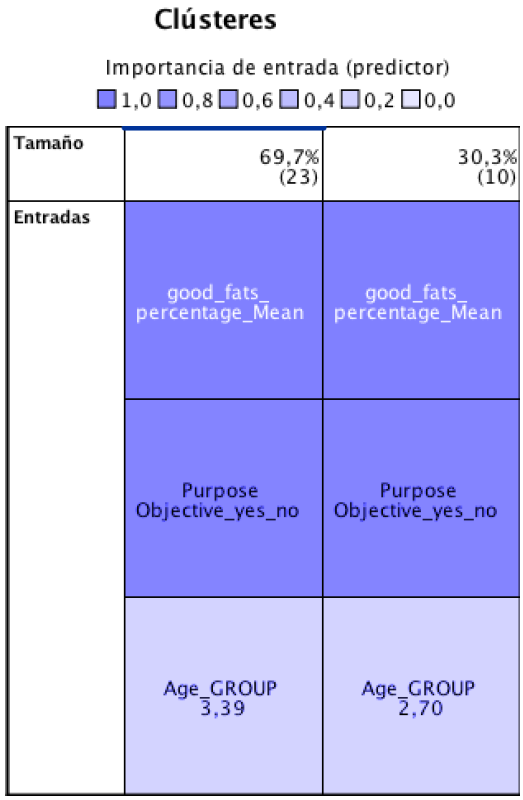


***Figura 4.8**. Modelo 1.2. Vista de conglomerados.*

Y a continuación, se visualizan los detalles de cada uno de los clústeres generados:

Clúster 1 (*Figura 4.9*): Está formado por aquellos sujetos que ingieren mayor porcentaje de grasas buenas, la mayoría de ellos han conseguido el éxito, y pertenecen a los grupos de mayor edad.

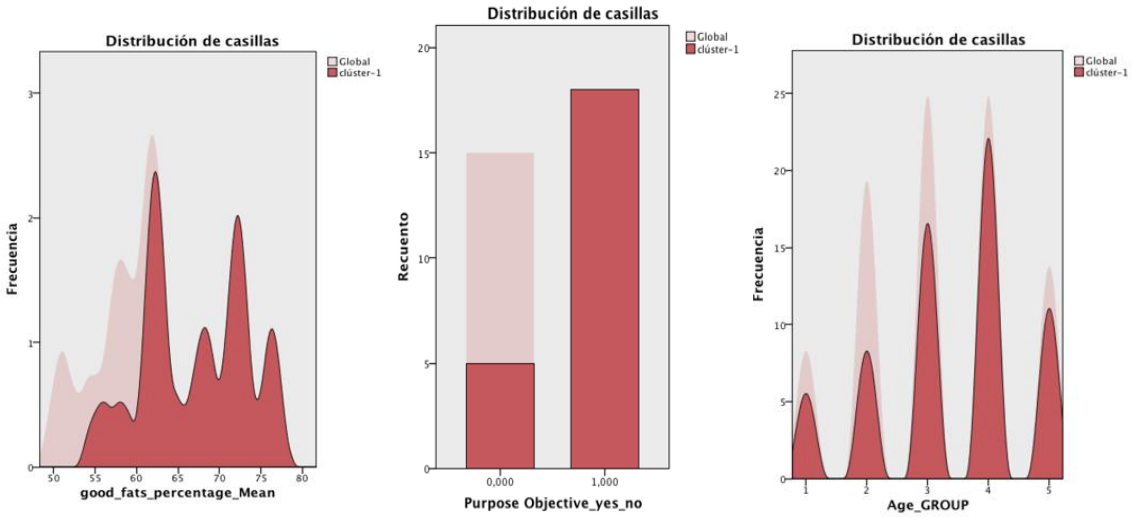


***Figura 4.9**. Modelo 1.2. Clúster 1.*

Clúster 2 (*Figura 4.10*): Está formado por aquellos sujetos que ingieren menor porcentaje de grasas buenas, de los cuáles ninguno ha conseguido el éxito, y la mayoría pertenece al grupo de edad 2 y 3.

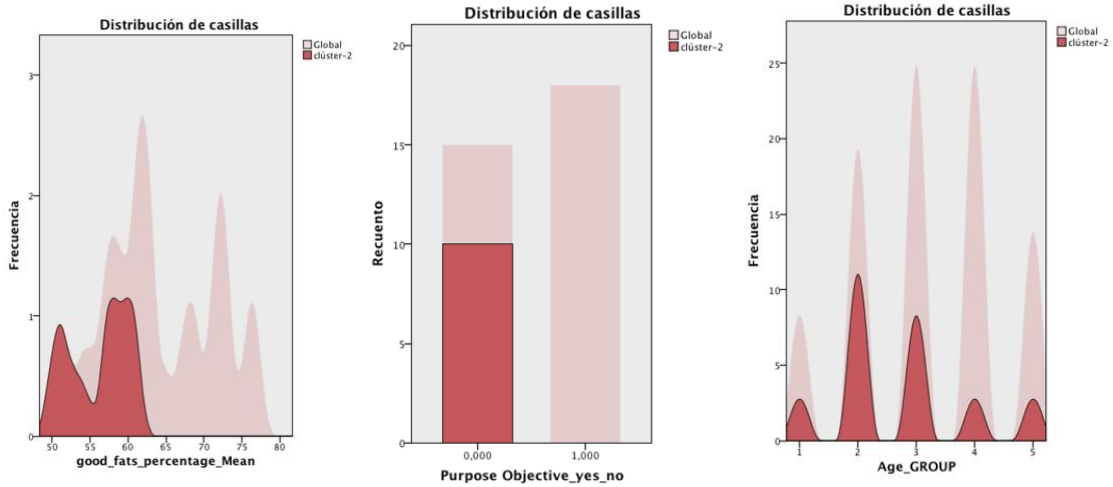


***Figura 4.10.** Modelo 1.2. Clúster 2.*

1.3. Comparación gráfica

Por último, para determinar si las variables obtenidas del modelo, son específicas del fin de semana, o bien, ocurren durante toda la semana, se realizan dos gráficos de comparación a partir de la tabla que engloba todos los registros. Y a partir de estos (*Figura 4.11 y Figura 4.12*), junto con lo obtenido anteriormente, se concluye que el hábito de consumir menos del 61% de grasas buenas, es una práctica propia del fin de semana, y esta rutina ocurre en pacientes que poseen alguna enfermedad en tratamiento.

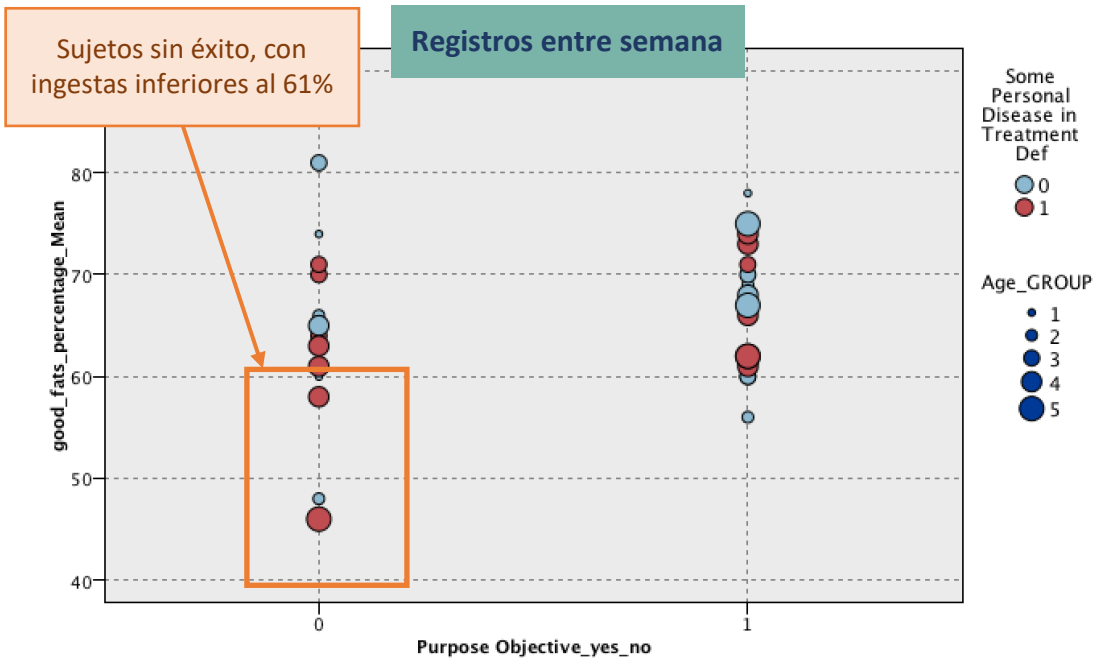


***Figura 4.11**. Modelo 1.3. Gráfico de grasas saludables entre semana.*

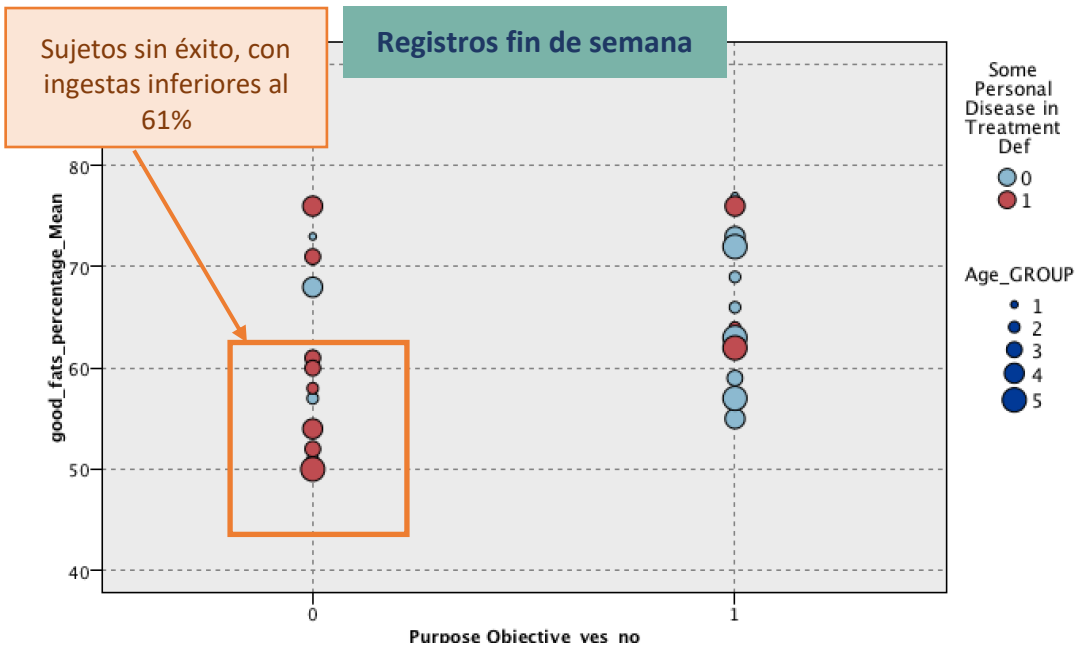


***Figura 4.12**. Modelo 1.3 Gráfico de grasas saludables en fin de semana.*

Modelo 2.

Para la generación del modelo 2, se ha utilizado la tabla de nutrición correspondiente a la **visita 5 con valores de entre semana**, que abarca los registros comprendidos entre la visita 5 y 6, periodo de dos semanas.

2.1. Clasificación Red Bayesiana.

Se realiza un modelo de Red Bayesiana con el propósito de visualizar la relación entre el campo objetivo: *Purpose Objective_yes_no* y los campos de entrada: de los principales grupos de alimentos (*cereals_kcal_Mean, meats_kcal_Mean y vegetables_kcal_Mean)* y de los macronutrientes (carbohydrates_Mean, *proteins_Mean* y total_fats_gr_Mean). Así como, obtener las independencias probabilísticas o condicionales entre ellas.

De modo que, se utiliza la estructura de red bayesiana de manto de Markov, a fin de identificar, de estas variables que forman la red, cuáles son las principales para predecir el éxito. Así pues, como se aprecian en la *Figura 4.13* y *Figura 4.14*, las variables más significativas son: los gramos de proteínas y las kcal de vegetales consumidas al día.

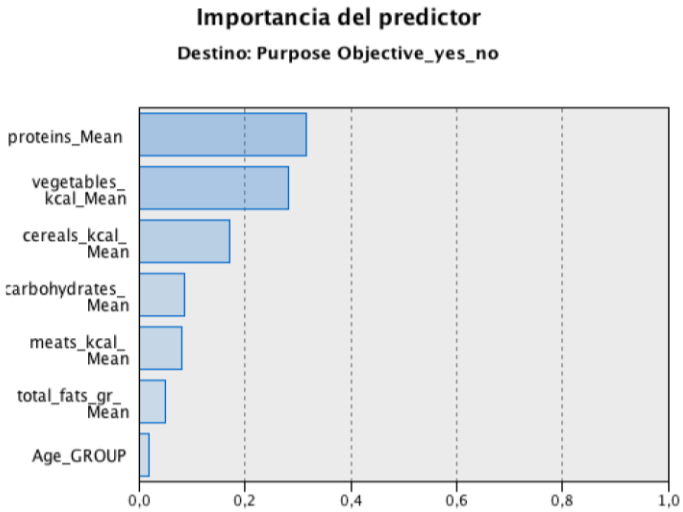


***Figura 4.13**: Modelo 2.1. Importancia del predictor*

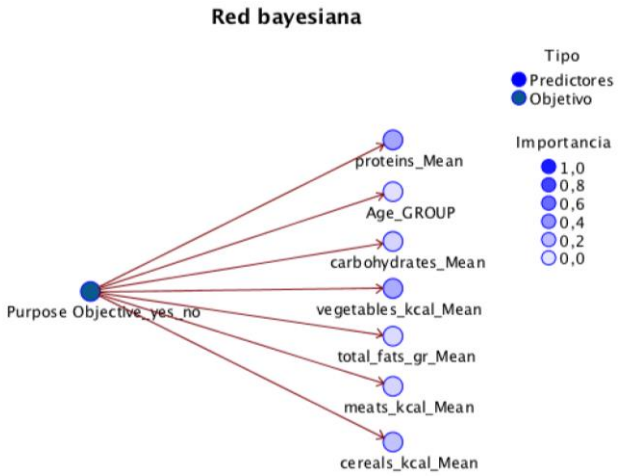


***Figura 4.14:** Modelo 2.1. Red bayesiana*

Por último, se analizan las tablas de probabilidades condicionadas de estas variables.

- A partir de la *Tabla 4.2*, se obtiene que de los sujetos que han conseguido el éxito un 51 % ha consumido cantidades de proteínas entre 78.2g/día y 95.8g/día, este rango, coincide con la prescripción médica óptima. Por otro lado, se observa que el 60 % de los que no, han consumido menos de 78.2 g /día. Por tanto, se considera que cumplir con el consumo de proteínas/día recomendadas es esencial a fin de conseguir el logro del programa.

De los pacientes que si han obtenido éxito:

El **31%** de los pacientes, ha consumido: cantidades de proteínas < 78.2 g.

El **51 %** de los pacientes, han ingerido cantidades de proteínas entre 78.2 g y 95.8 g

Tabla 4.2. *Modelo 2.1. Probabilidades condicionales de los gramos de proteínas.*

Probabilidades condicionales de proteins_Mean

Padres	Probabilidad				
Purpose Objective_yes_no	<= 60,6	60,6 ~ 78,2	78,2 ~ 95,8	95,8 ~ 113,4	> 113,4
1	0,01	0,30	0,51	0,16	0,01
0	0,23	0,44	0,16	0,09	0,09

De los pacientes que no han obtenido éxito:

El **67%** de los pacientes, ha consumido: gramos de proteínas < 78.2 g.

El **16 %** de los pacientes, han ingerido: gramos de proteínas entre 78.2g- 95.8 g

Aquellos sujetos que han consumido entre 78.2 y 95.8 gramos de proteínas, el **76.11 %** ha logrado el éxito mientras que el **23.88 %** no.

- A partir de la *Tabla 4.3*, se obtiene que los sujetos que han consumido cantidades de vegetales superiores a 98.8 kcal/diarios, el 68.42% ha logrado el éxito, mientras que el 31.58% no. Por tanto, se concluye, que la ingesta de vegetales debería ser superior al 98.8 kcal/diarias a fin de lograr el éxito.

De los pacientes que si han obtenido éxito:

El **32 %** de los pacientes, ha consumido: kcal vegetales < 70.2 kcal.

El **39 %** de los pacientes, ha consumido: kcal vegetales > 98.8 kcal

Tabla 4.3. *Modelo 2.1. Probabilidades condicionales de las kcal de vegetales.*

Probabilidades condicionales de vegetables_kcal_Mean

Padres	Probabilidad				
Purpose Objective_yes_no	<= 41,6	41,6 ~ 70,2	70,2 ~ 98,8	98,8 ~ 127,4	> 127,4
1	0,09	0,23	0,30	0,23	0,16
0	0,16	0,44	0,23	0,09	0,09

De los pacientes que no han obtenido éxito:

El **60%** de los pacientes, ha consumido: kcal vegetales < 98.8 kcal

El **18%** de los pacientes, ha consumido: kcal vegetales > 98.8 kcal

Evaluación

El análisis de este modelo muestra que predecirá con un porcentaje del 30 % (*Tabla 4.4*).

Tabla 4.4. *Modelo 2.1. Evaluación.*

Resultados para el campo de resultado Purpose Objective_yes_no

Comparando $B-Purpose Objective_yes_no con Purpose Objective_yes_no

'Partición'	2_Comprobación	
Correctos	3	30%
Erróneos	7	70%
Total	10	

A pesar que el porcentaje de predicción del modelo es bajo, se ha de tener en cuenta que se han introducido muchas variables para generarlo, las cuales tienen valores muy similares entre los sujetos que logran el éxito y los que no, por lo tanto, no son predictoras.

2.2. Clasificación: árbol de decisión CHAID

Con el propósito de evaluar la relación entre el consumo de proteínas y vegetales se elabora un árbol de decisión interactivo CHAID (*Figura 4.15*), en el que se observa que la variable de proteínas es una variable muy significativa, debido a que si un sujeto ingiere más de 81g/día de proteínas tienen una probabilidad del 100% de lograr el propósito (p= 0.001).

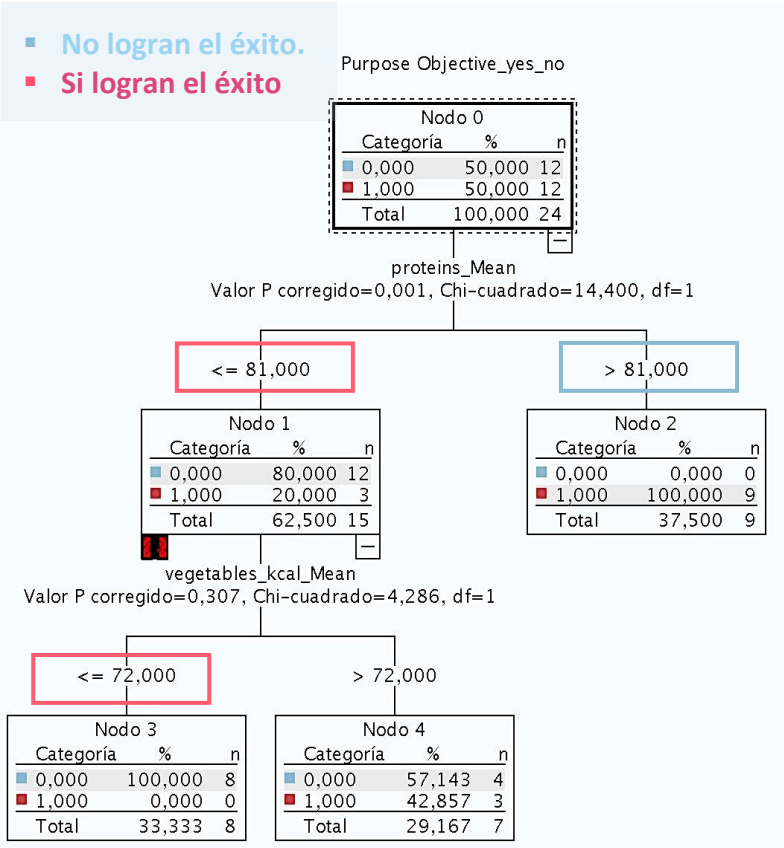


Figura 4.15. *Modelo 2.2. Árbol de decisión CHAID.*

Evaluación

Tal y como se muestra en la *Tabla 4.5,* este modelo predecirá con un porcentaje del 60%.

Tabla 4.5. *Modelo 2.2. Evaluación*

Resultados para el campo de resultado Purpose Objective_yes_no

Comparando $R-Purpose Objective_yes_no con Purpose Objective_yes_no

'Partición'	2_Comprobación	
Correctos	6	60%
Erróneos	4	40%
Total	10	

Finalmente, a partir de las tablas de probabilidades condicionadas (*Tabla 4.1 y 4.2*) y el árbol de CHAID (*Figura 4.16*), se concluye, que la ingesta de los vegetales es una variable característica en el programa de pérdida de peso. Así cómo, el cumplir con la recomendación médica de un consumo comprendido (75 - 95) g/día según las kcal prescritas, es importante a fin de conseguir el éxito.

Modelo 3.

Para la generación del modelo 3, se ha utilizado la tabla de nutrición correspondiente a la **visita 6 con valores de fin de semana**, que abarca los registros comprendidos entre la visita 6 y 7, periodo de cuatro semanas.

3.1. Lista de decisiones

Primeramente, se realiza el modelo: lista de decisiones, para identificar los subgrupos o segmentos de los datos que muestran mayor posibilidad de no lograr el éxito. De modo que, se configura el modelo, para el valor 0 del campo objetivo, con una alta probabilidad, es decir, se buscan aquellas variables con mayor probabilidad para obtener, en este caso 0.

Y, tal y como se contempla en la *Figura 4.16,* se obtiene, que la variable más relevante es el porcentaje de grasas buenas consumidas, debido a que de entre los 24 registros totales, 6 se corresponden a sujetos que han ingerido menos de 60 g/diarios y de esos 6, el 100% no ha tenido éxito.

Por tanto, al igual que en el periodo de las dos primeras semanas (Modelo 1), la variable más significativa durante el fin de semana, es el porcentaje de grasas buenas consumidas.

id	Reglas de segmentación	Puntuación	Cobertura (n)	Frecuencia	Probabilidad
	Todos los segmentos incluido Resto		24	11	45,83%
1	**good_fats_percentage_Mean** good_fats_percentage_Mean <= 60.000	0	6	6	100,00%
	Resto		18	5	27,78%

***Figura 4.16**. Modelo 3.1. Lista de decisiones.*

3.2. Regresión logística binomial

A fin de clasificar los registros a partir de los valores de los campos de entrada, se emplea el modelo de regresión logística, el cuál se configura para desarrollar un modelo binomial mediante un método por pasos hacia delante.

Y, tras ejecutar este modelo, se obtienen diferentes tablas estadísticas que ofrecen información detallada sobre el modelo estimado y su rendimiento. De las cuáles, cabe destacar:

- **Variables en la ecuación** (*Tabla 4.6*), en esta tabla se muestran las variables que han sido seleccionadas para la realización del modelo, junto el valor p y el odd ratio correspondiente a cada una de ellas. Estos valores me indican si existe o no una, relación relevante entre la variable objetivo y cada una de ellas.

***Tabla 4.6**. Modelo 3.2. Variables en la ecuación.*

		gl	Sig.	Exp(B)
Paso 1[a]	proteins_Mean	1	,022	1,116
	Constante	1	,030	,001
Paso 2[b]	proteins_Mean	1	,027	1,140
	fruits_units_Mean	1	,057	3,341
	Constante	1	,025	,000

- **Pruebas ómnibus de coeficientes de modelo** *(Tabla 4.7)* una prueba Chi Cuadrado que evalúa la hipótesis nula de que los coeficientes (β) de todos los términos (excepto la constante) incluidos en el modelo son cero. En esta tabla se muestra el p-valor del modelo tras el último paso, un valor de p-valor =0.02, que indica una probabilidad del 98% de que haya una relación significativa entre conseguir éxito y las variables empleadas para realizar el modelo.

Tabla 4.7. *Modelo 3.2. Pruebas ómnibus de coeficientes del modelo.*

Pruebas ómnibus de coeficientes de modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	8,043	1	,005
	Bloque	8,043	1	,005
	Modelo	8,043	1	,005
Paso 2	Paso	4,906	1	,027
	Bloque	12,949	2	,002
	Modelo	12,949	2	,002

- **La tabla de clasificación** (*Tabla 4.8*), muestra los resultados del modelo a medida que se añade un predictor en cada paso. Este modelo se ha desarrollado en dos pasos, en el primer paso, con el predictor 1, proteínas ingeridas, el modelo tiene una precisión del 70.4 %, y al añadirle un segundo predictor, unidades de fruta, la precisión de predicción del modelo aumenta hasta obtener un valor del 77.8 %. Siendo este, el paso 2, el último paso realizado por el algoritmo, que decide que no se ha de añadir ningún predictor más.

Tabla 4.8. *Modelo 3.2. Tabla de clasificación.*

Tabla de clasificación

			Pronosticado		
			Purpose Objective_yes_no		Porcentaje correcto
	Observado		0	1	
Paso 1	Purpose Objective_yes_no	0	6	5	54,5
		1	3	13	81,3
	Porcentaje global				70,4
Paso 2	Purpose Objective_yes_no	0	7	4	63,6
		1	2	14	87,5
	Porcentaje global				77,8

- **Resumen del modelo** *(Tabla 4.9)***,** se aportan tres medidas para evaluar de forma global la validez del modelo: la primera es el valor del -2LL, que mide el ajuste del modelo a los datos, y las otras dos son Coeficientes de Determinación (R2), que expresan la proporción (en tanto por uno) de la variación explicada por el modelo. Se observa, como en cada paso el modelo se mejora, ya que el valor de -2LL disminuye mientras que los R cuadrado aumentan.

Tabla 4.9. *Figura 5.22. Modelo 3.2. Resumen.*

Resumen del modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	28,456[a]	,258	,348
2	23,550[b]	,381	,514

Evaluación

El análisis de este modelo muestra que, para 4 de un total de 4 registros (el 100%), coinciden los valores predichos con los reales, por lo tanto, tal y como se muestra en la *Tabla 4.10*, este modelo predecirá con un porcentaje del 100 %.

Tabla 4.10*. Modelo 3.2. Evaluación.*

Resultados para el campo de resultado Purpose Objective_yes_no

Comparando $L-Purpose Objective_yes_no con Purpose Objective_yes_no

'Partición'	2_Comprobación	
Correctos	4	100%
Erróneos	0	0%
Total	4	

Pero, debido a que la base de datos de análisis es pequeña, por tanto, los datos de entrenamiento y comprobación también lo son, el valor obtenido de la *Tabla 5.19*, no es fiable al completo, es por ello que se examinan las tablas adquiridas de este modelo. Se observa que el modelo es significativo en su conjunto, debido al p-valor= 0.02 que se muestra en la *Tabla 4.6*. Así cómo, los p-valores de los parámetros estimados del modelo (columna Sig. de la *Tabla 4.5* con valores pequeños), y, los ods ratio o razones de ventajas (columna Exp (B), con valores mayor que la unidad), son muy aceptables. Sin embargo, los pseudo R-cuadrados han crecido (*Tabla 4.8),* pero no son los más óptimos, aunque esto no es determinante. Y, tras evaluar el modelo, se obtiene una predicción del 100%.

Por tanto, se concluye que el modelo estimado paso a paso que ha seleccionado las variables significativas, muestra que las variables que influyen, en fin, de semana, en la probabilidad de conseguir el éxito son las proteínas ingeridas y las unidades de fruta.

3.3. Árbol de decisión: CHAID

Con objetivo, de una mejor compresión y evaluación de las variables obtenidas con el modelo anterior de regresión logística, se ejecuta un árbol de decisión, en el que se observan los valores límite para conseguir el éxito o no, así como la influencia y relación de estas variables.

En este árbol de decisión (*Figura 4.17*) se contempla con claridad que, si un sujeto no consume ninguna fruta, durante el fin de semana, posee una probabilidad del 100% (p=0.036) de no conseguir el éxito. Sin embargo, si el sujeto consume fruta y, además, consume más de 73 g/día proteínas tiene una probabilidad del 100% (p=0.049) de conseguir éxito.

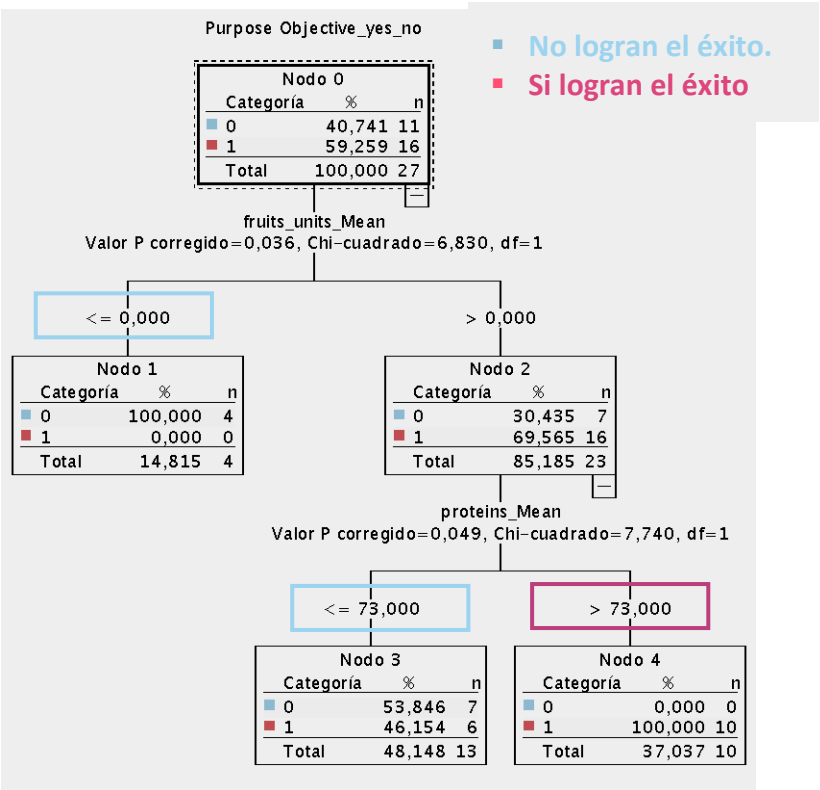


***Figura 4.17**. Modelo 3.3. Árbol de decisión: CHAID.*

Evaluación

Tras la evaluación del modelo (*Tabla 4.11*), que me indica una predicción del 100%, junto con lo obtenido en la *Figura 4.17* y las tablas del modelo de regresión logística binomial se considera que la ingesta de proteínas y cantidad de fruta es determinante para lograr el propósito.

***Tabla 4.11.** Modelo 3.1. Evaluación*

Resultados para el campo de resultado Purpose Objective_yes_no

Comparando $R-Purpose Objective_yes_no con Purpose Objective_yes_no

'Partición'	2_Comprobación	
Correctos	4	100%
Erróneos	0	0%
Total	4	

3.4. Comparación gráfica

Por último, para determinar si las variables obtenidas del modelo, son específicas del fin de semana, o bien, ocurren durante toda la semana, se realizan dos gráficos de comparación a partir de la tabla que engloba todos los registros. Y a partir de estos gráficos (*Figura 4.18 y Figura 4.19*), se concluye que el hábito de disminuir el consumo de frutas, es propio del fin de semana, mientras que el de consumir pocas proteínas es algo que ocurre de forma generalizada, en aquellos sujetos que no logran el éxito.

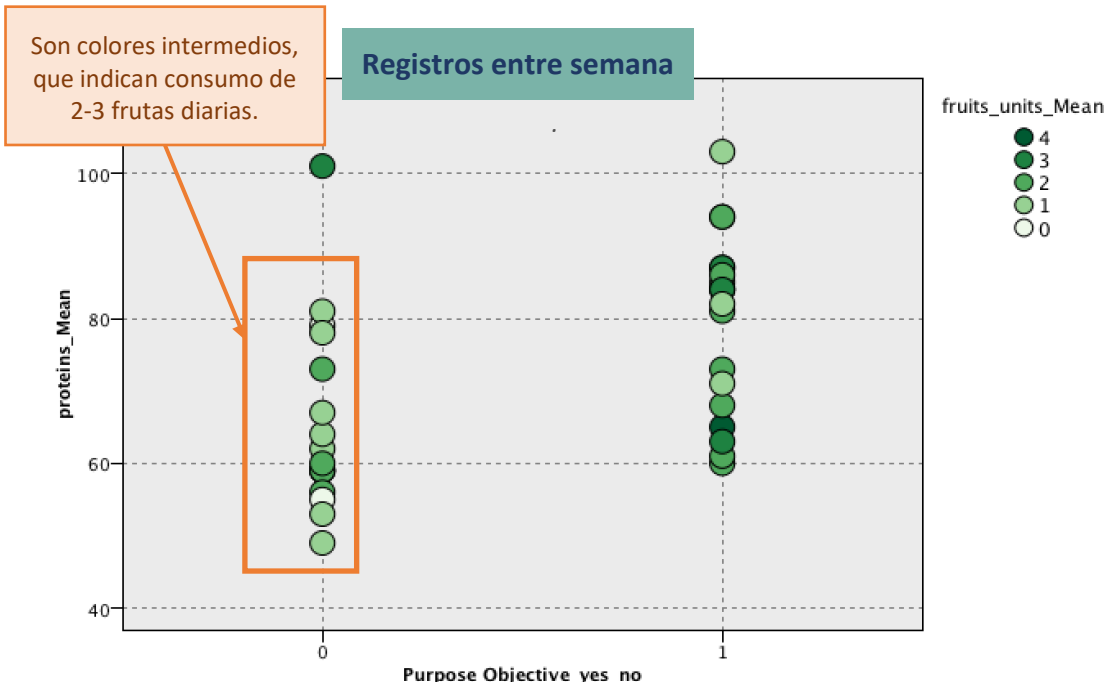


***Figura 4.18**. Modelo 3.4. Gráfico de ingesta de proteínas y frutas entre semana.*

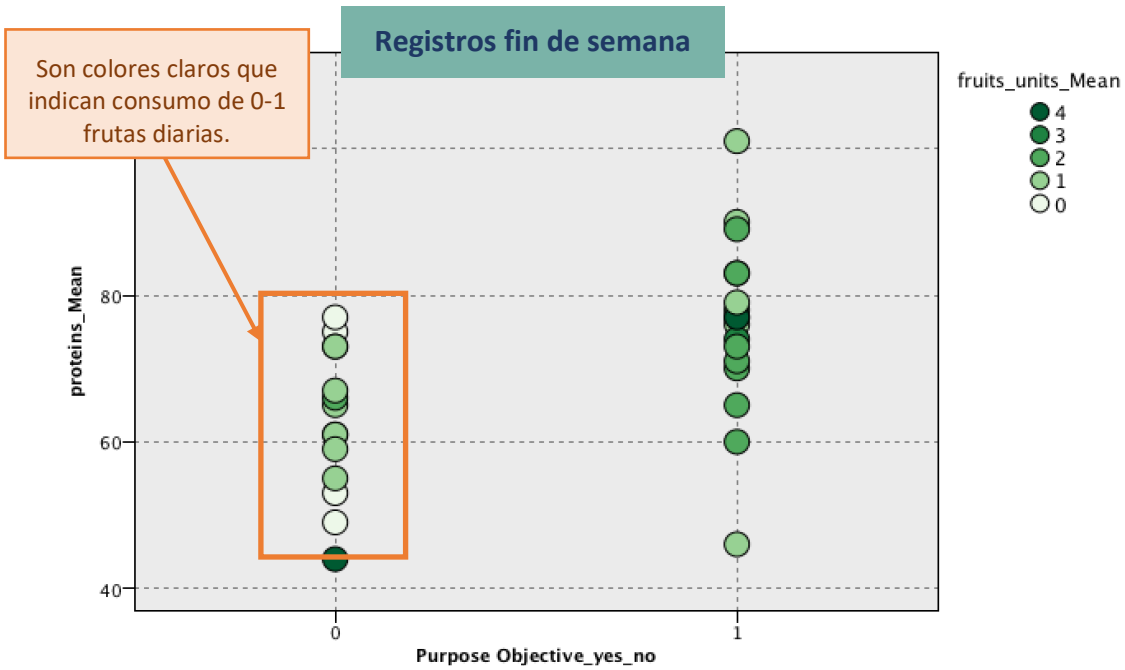


***Figura 4.19**. Modelo 3.4. Gráfico de ingesta de proteínas y frutas en fin de semana*

Modelo 4.

Para la generación del modelo 4, se ha utilizado la tabla de nutrición correspondiente a la **visita 6 con valores de entre semana**, que abarca los registros comprendidos entre la visita 6 y 7, periodo de cuatro semanas.

4.1. Listas de decisiones

A fin de identificar los subgrupos o segmentos de los datos que muestran una mayor o menor posibilidad de conseguir el éxito se emplea el modelo de listas de decisiones.

De modo que, primeramente, se configura el nodo, para el **valor 0** del campo objetivo, con una alta probabilidad, es decir, se buscan aquellas variables con mayor probabilidad para obtener, en este caso 0.

Y, como se observa en la *Figura 4.20,* se obtiene, que las variables más relevantes en el fracaso, son la ingesta de alimentos no saludables (*trash_kcal_Mean*) superior a 133 kcal/diarias y la ingesta inferior de 59 g/día de proteína.

id	Reglas de segmentación	Puntuación	Cobertura (n)	Frecuencia	Probabilidad
	Todos los segmentos incluido Resto		27	11	40,74%
1	**trash_kcal_Mean** trash_kcal_Mean > 133.000	0	5	5	100,00%
2	**proteins_Mean** proteins_Mean <= 59.000	0	3	3	100,00%
	Resto		19	3	15,79%

Figura 4.20. *Modelo 4. Lista de decisiones para valor del Objetivo=0.*

Asimismo, se configura otro nodo, para el **valor 1** del campo objetivo, con una alta probabilidad, es decir, se buscan aquellas variables con mayor probabilidad para obtener, en este caso 1.

Y, como se contempla en la *Figura 4.21*, se obtiene que la variable más relevante para conseguir éxito, es la ingesta de proteínas superior a 82 g/día, pero inferior a 94 g/día. Por tanto, tal y cómo se ha mostrado y comentado en el modelo 2, el cumplimiento de la cantidad en gramos recomendada de proteínas es importante a fin de conseguir el éxito.

id	Reglas de segmentación	Puntuación	Cobertura (n)	Frecuencia	Probabilidad
	Todos los segmentos incluido Resto		27	16	59,26%
1	**proteins_Mean** proteins_Mean > 82.000 y proteins_Mean <= 94.000	1	8	8	100,00%
	Resto		19	8	42,11%

Figura 4.21*. Lista de decisiones para valor del Objetivo=1.*

4.2. Árbol de decisión: CHAID.

Con la finalidad de visualizar más detalladamente lo obtenido con el modelo de lista de decisiones, se ejecuta un árbol de decisión.

En este árbol de decisión (*Figura 4.22*), se observa con claridad que, si un sujeto consume más de 106 kcal de alimentos no saludables, tiene una probabilidad de 87.5 % (p=0.009) de no superar el programa, así como, aquellos que, aunque consuman menos de 106 kcal de esta comida, si ingieren menos de 60 g/día de proteína poseen de una probabilidad del 80% (p=0.001) de no superarlo. Sin embargo, si estos últimos sí que consumen más de 60 g/día de proteína, la probabilidad de superar el programa es del 100% (p=0.001).

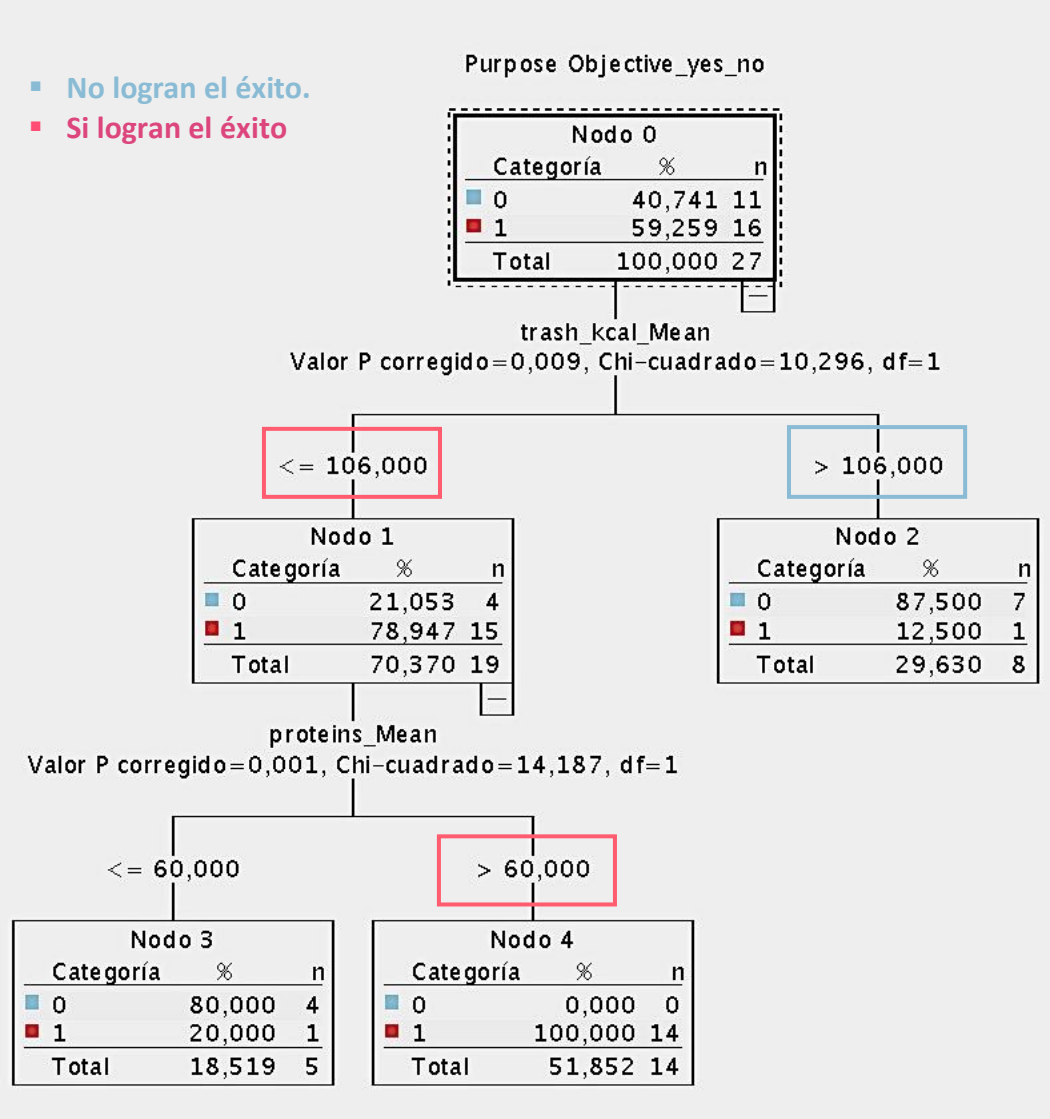


Figura 4.22. Modelo 4.2. Árbol de decisión CHAID.

Evaluación del modelo

A pesar de que, al evaluar este modelo (*Tabla 4.12*), la predicción sea tan sólo del 55,56 %, se ha de tener en cuenta que se trabaja con una base de datos con pocos registros, por lo que los datos de entrenamiento y comprobación son reducidos, es por ello que este valor no es determinante para descartar el modelo.

***Tabla 4.12**. Modelo 4.2. Evaluación*

Resultados para el campo de resultado Purpose Objective_yes_no

Comparando $R-Purpose Objective_yes_no con Purpose Objective_yes_no

'Partición'	2_Comprobación	
Correctos	5	55,56%
Erróneos	4	44,44%
Total	9	

Por tanto, en base a lo examinado con el árbol de decisión y el modelo lista de decisiones, se considera que la ingesta de alimentos no saludables, así como, una ingesta de proteínas inferior a 60 g/día son factores importantes en el fracaso del programa.

Modelo 5.

Para la generación del modelo 5, se ha utilizado la tabla de nutrición correspondiente a la **visita 7 con valores de fin de semana**, que abarca los registros comprendidos entre la visita 7 y 8, periodo de cuatro semanas.

Tras generar diversos modelos a partir de la tabla de nutrición correspondiente a la visita 7, se obtienen prácticamente las mismas variables significativas que las conseguidas con modelos anteriores. Y, al aplicar el modelo lista de decisiones, se aprecia claramente en la *Figura 4.23*, al igual que en el modelo 3, que la fruta consumida durante el fin de semana es una variable determinante.

id	Reglas de segmentación	Puntuación	Cobertura (n)	Frecuencia	Probabilidad
	Todos los segmentos incluido Resto		25	12	48,00%
1	**fruits_kcal_Mean** fruits_kcal_Mean > 22.000 y fruits_kcal_Mean <= 89.000	0	5	5	100,00%
	Resto		20	7	35,00%

***Figura 4.23**. Modelo 5. Lista de decisiones*

5.1. Comparación gráfica

Con propósito de indagar en los datos y adquirir nuevo conocimiento, se genera unos gráficos (*Figura 4.24 y Figura 4.25*), en los que se examinan que el beber alcohol es un hábito del fin de semana, realizado generalmente por los grupos 2 (30-40) y 4 (50-56).

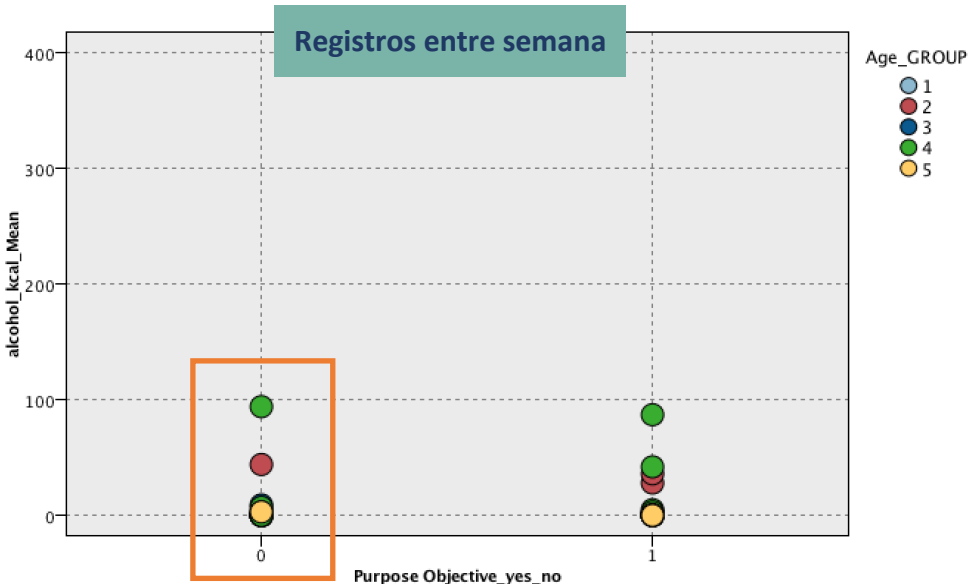


Figura 4.24. *Modelo 5.1. Gráfico alcohol entre semana.*

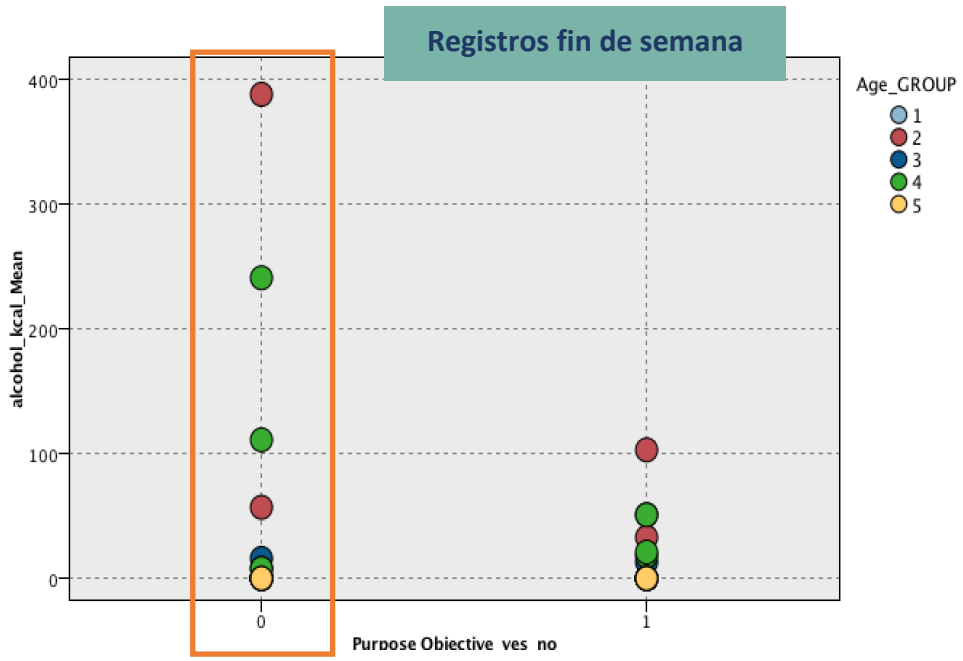


Figura 4.25. *Modelo 5.1. Gráfico alcohol en fin de semana*

5.2. Red bayesiana

A fin de determinar, qué valor límite establecer de kcal de alcohol, se realiza un modelo de red bayesiana, para así obtener la *Tabla 4.13* de probabilidades condicionales. En esta tabla, se observa, que todos aquellos sujetos que beben más de 128.8 kcal de alcohol no consiguen el propósito.

Tabla 4.13. *Modelo 5.2. Probabilidades condicionales de las kcal alcohol.*

Probabilidades condicionales de alcohol_kcal_Mean

Padres	Probabilidad			
Purpose Objective_yes_no	<= 128,8	128,8 ~ 322	> 322	3
1	1,00	0,00	0,00	0,00
0	0,83	0,08	0,08	0,00

Evaluación del modelo

Del análisis de este modelo se obtiene que predecirá el éxito, con un porcentaje del 42.86 % (*Tabla 4.14*), este valor es debido a que, para la realización del modelo, tan sólo se está evaluando la ingesta de alcohol, y que en términos generales los sujetos no ingieren alcohol. Por tanto, junto a la *Tabla 4.13* y las *Figuras 4.24 y 4.25*, se examina que, aunque el alcohol no es una variable determinante, se ha de tener en cuenta, concretamente durante el fin de semana.

Tabla 4.14*. Modelo 5.2. Evaluación.*

Resultados para el campo de resultado Purpose Objective_yes_no

Comparando $B-Purpose Objective_yes_no con Purpose Objective_yes_no

'Partición'	2_Comprobación	
Correctos	3	42,86%
Erróneos	4	57,14%
Total	7	

Modelo 6.

Para la generación del modelo 6, se ha utilizado la tabla de nutrición correspondiente a la **visita 7 con valores entre semana**, que abarca los registros comprendidos entre la visita 7 y 8, periodo de cuatro semanas.

6. 1. Árbol de decisión: Algoritmo C5.0

Tras aplicar el algoritmo C5.0, se obtiene:

- **Conjunto de reglas** (*Figura 4.26*), a partir de las cuáles se construye el modelo. Se observa que vuelven aparecer muchas variables obtenidas anteriormente como son: unidades de frutas, proteínas y kilocalorías de comida no saludable.

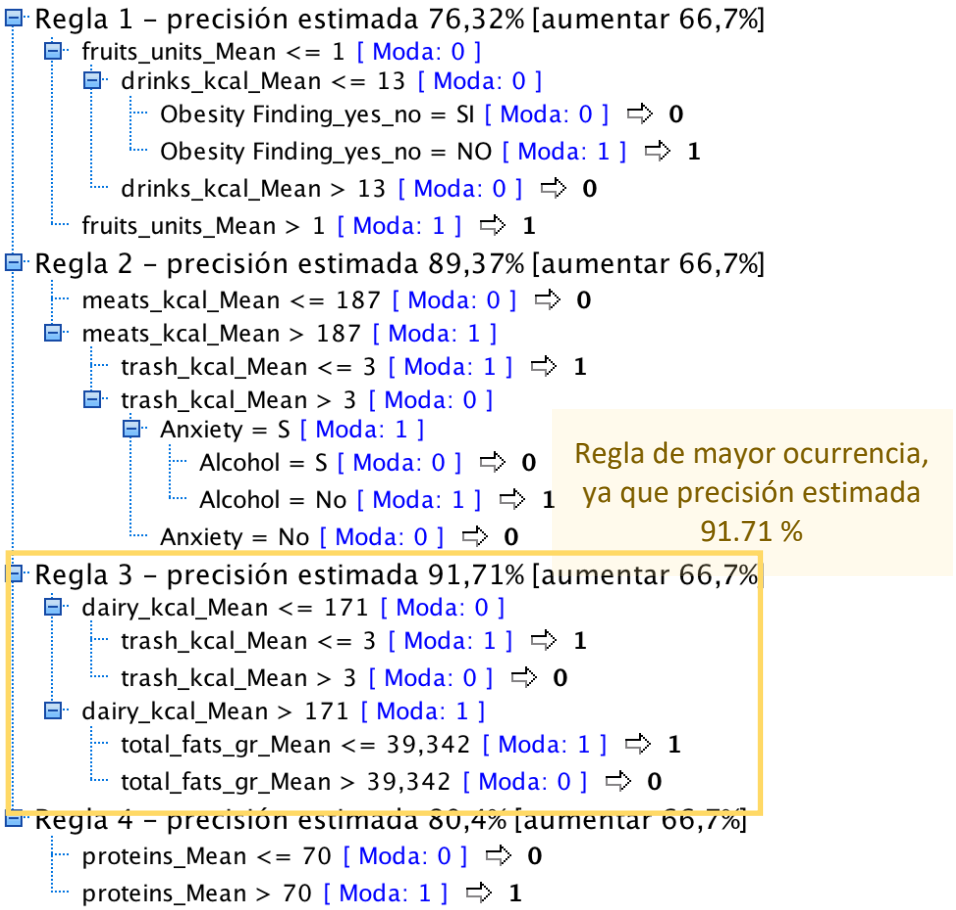


***Figura 4.26**. Modelo 6.1. Conjunto de reglas del algoritmo C.5.*

- **Gráfico importancia del predictor** (*Figura 4.27*), indica la importancia de cada una de las variables que se han utilizado para la generación del modelo.

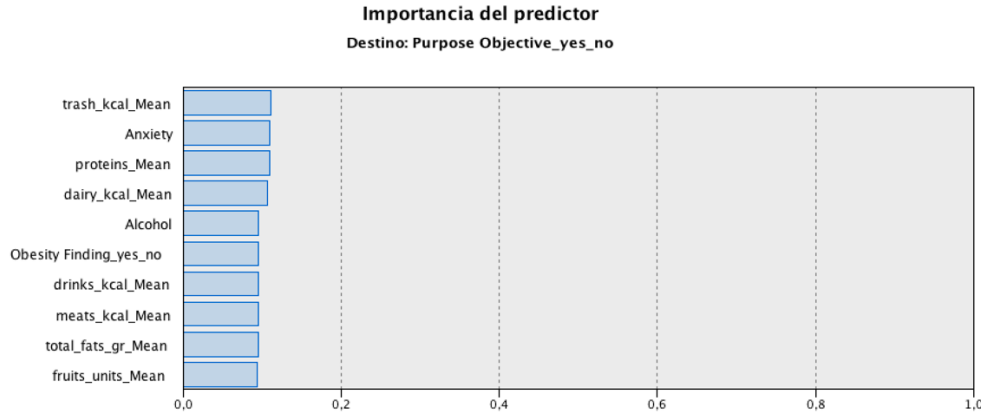


Figura 4.27. *Modelo 6.1. Importancia del predictor.*

- **Árbol de decisión de la regla** (*Figura 4.28*)**,** con mayor porcentaje de ocurrencia. De las 4 reglas obtenidas, se visualiza el gráfico de la regla 3 que es la de mayor porcentaje de ocurrencia, con una precisión estimada del 91.71%. En el árbol asociado, se muestra que, si un sujeto toma al día más de 171 kcal/diarias de lácteos posee un 78.464 % de probabilidad de conseguir el éxito, pero si además consume un total de grasas inferior o igual a 39.342 g, entonces según los datos obtenidos, seguro que alcanza la meta fijada. Por otra parte, si un paciente no consume 171 kcal/diarias de lácteos y como añadido, consume alimentos no saludables, en el 95.596% de los casos no cumplirá.

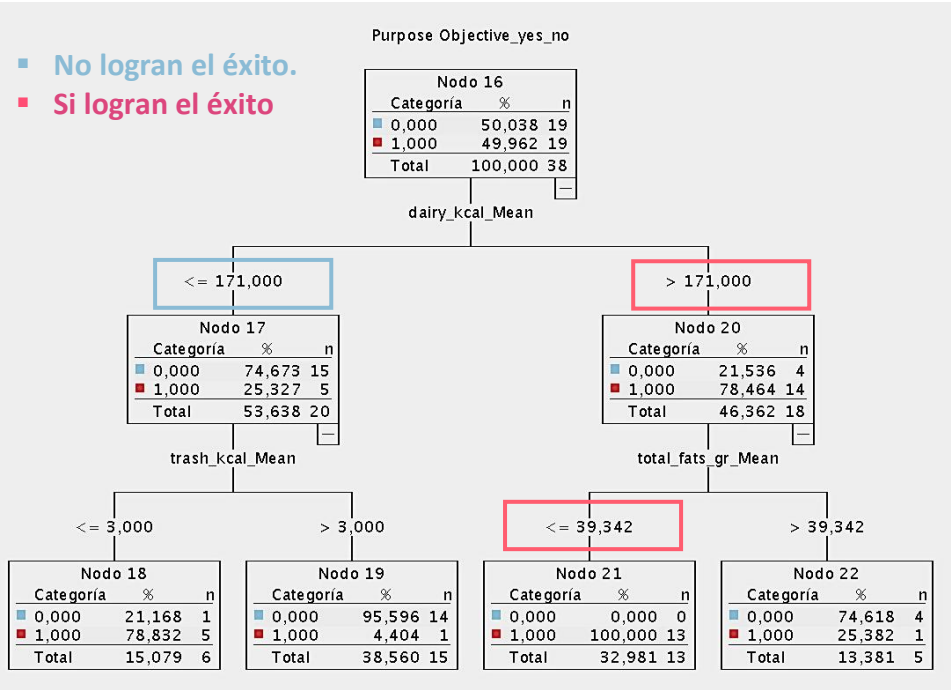


Figura 4.28*. Modelo 6.1. Árbol de decisión.*

Evaluación

Al emplear el algoritmo de C.5, este realiza una validación cruzada, por tanto, en este caso no se dividen los datos en subconjuntos de entrenamiento y comprobación. Esta opción, permite obtener un modelo con una predicción 94.74 % (*Tabla 4.15*).

Por tanto, a partir de este valor y de lo obtenido en la *Figura 4.28*, se valora que para conseguir el propósito se han de consumir más de 172 kcal/día de lácteos. Sin embargo, se descarta, el valor de consumir menos de 39.342 g de grasas totales, ya que es inferior a lo prescripto por los médicos, y esto no es recomendable.

***Tabla 4.15**. Modelo 6.1. Evaluación*

Resultados para el campo de resultado Purpose Objective_yes_no

Comparando $C-Purpose Objective_yes_no con Purpose Objective_yes_no

Correctos	36	94,74%
Erróneos	2	5,26%
Total	38	

5.2. Modelos de ejercicio

A partir de las tres tablas de ejercicio obtenidas de la fase de preparación de los datos, correspondientes a cada una de las visitas: 5,6,7, se ha realizado un proceso de modelado para cada una de ellas. Pero debido a que tienen pocos campos referidos a la actividad física, se han obtenido las mismas variables significativas. Esto sucede, ya que, a lo largo de todas las visitas, tal y como se observa en la *Figura 4.30 y Figura 4.31*, las variables características son las mismas. Es por ello que a partir de estas 3 tablas se ha realizado un único modelo.

- En la *Figura 4.29*, se contempla que la mayoría de los sujetos que no han conseguido éxito han quemado una media de 1500 kcal/semana, y en la visita 7, muchos no han registrado ninguna actividad física.

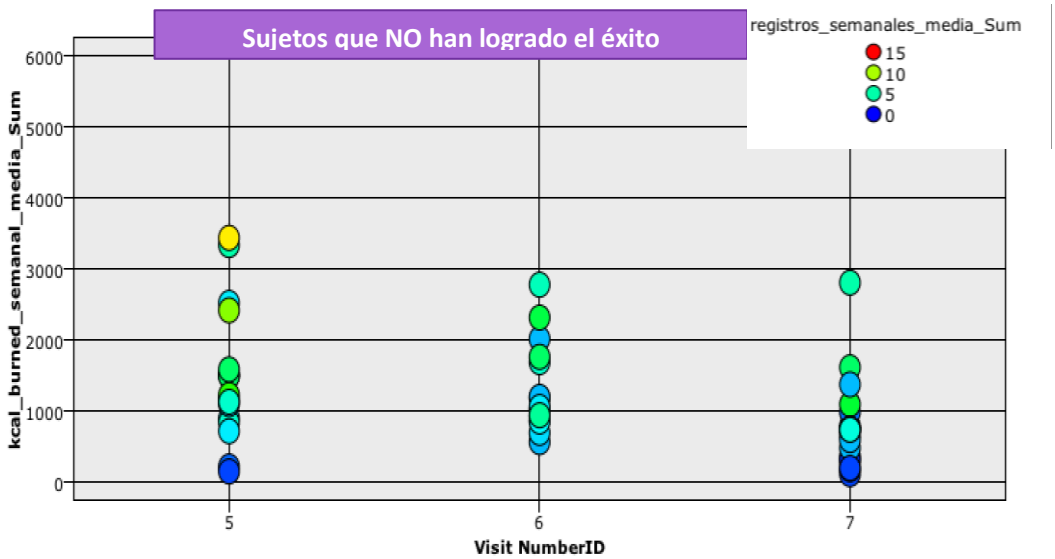


***Figura 4.29**. Kcal quemadas y registros de ejercicio a lo largo de las visitas 5,6 y 7.*

- En la *Figura 4.30*, se contempla que los sujetos que, si han conseguido éxito, han quemado una media de 2000 kcal/semana, y cada vez, van registrando mayor número de actividad física semanalmente.

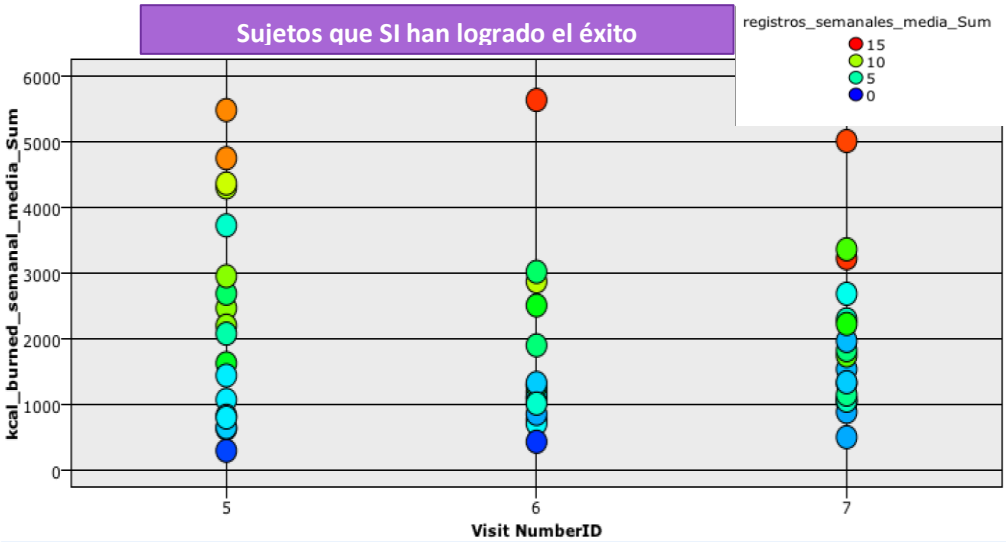


***Figura 4.30**. Kcal quemadas y registros de ejercicio a lo largo de las visitas 5,6 y 7.*

Modelo 7. Ejercicio

7.1. Comparación gráfica

Primeramente, se examina cada una de las tablas de ejercicio, y se obtiene que en la tabla que corresponde a la visita 7, se aprecia claramente la relación entre la cantidad de registros semanales y las kcal quemadas, respecto si ha conseguido o no el éxito (*Figura 4.31*).

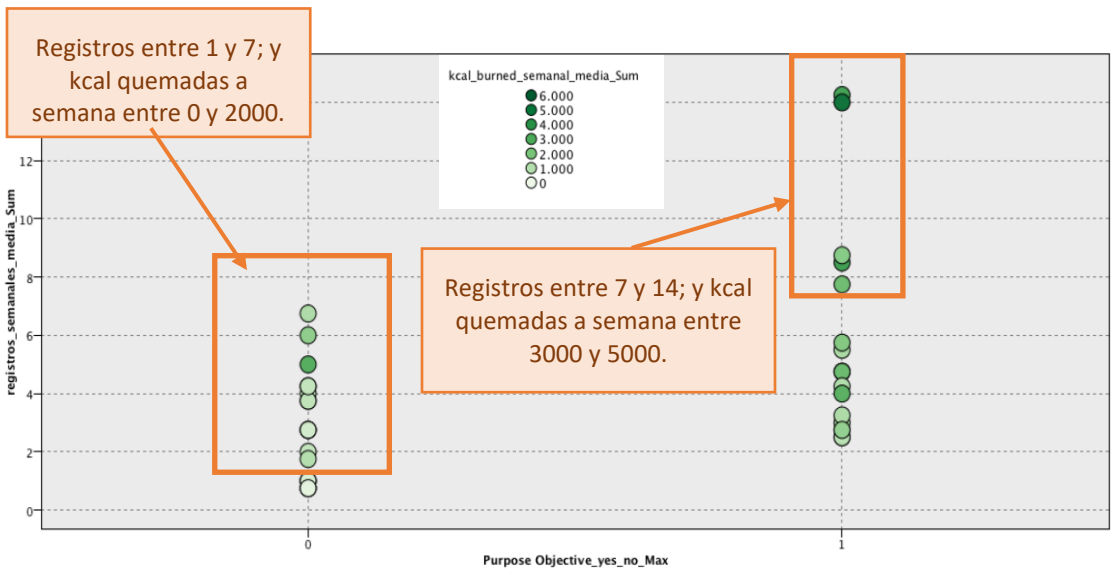


***Figura 4.31**. Modelo 7.1. Gráfico registros, calorías quemadas y objetivo.*

7.2. Red bayesiana.

A partir de la *Figura 4.31*, se genera un modelo de red bayesiana a fin de indagar la relación entre estas variables, y se obtienen así, las probabilidades condicionales de cada una de ellas:

- En la *Tabla 4.16*, que relaciona el objetivo con las kcal quemadas a la semana, se obtiene que el 81.57% de los sujetos que han quemado más de 2284.669 kcal, han adquirido éxito, mientras que el 18.42 % no. Así como se observa, que de entre los que no han conseguido el éxito el 87% han quemado menos de 1214.897 kcal.

Tabla 4.16. *Probabilidades condicionales de las kcal quemadas.*

Probabilidades condicionales de kcal_burned_semanal_media_Sum

Padres	Probabilidad				
Purpose Objective_yes_no_Max	<= 1.214,897	1.214,897 ~ 2.284,669	2.284,669 ~ 3.354,441	3.354,441 ~ 4.424,213	> 4.424,213
1	0,38	0,31	0,15	0,08	0,08
0	0,87	0,07	0,07	0,00	0,00

- En la *Tabla 4.17* que relaciona el objetivo con los registros realizados a la semana, se adquiere que el 84.44 % de los sujetos que registra más de 6 veces/semana, han alcanzado el objetivo mientras que el 15.55 % no.

Tabla 4.17. *Probabilidades condicionales de los registros semanales.*

Probabilidades condicionales de registros_semanales_media_Sum

Padres	Probabilidad				
Purpose Objective_yes_no_Max	<= 3,4	3,4 ~ 5,8	5,8 ~ 8,2	8,2 ~ 10,6	> 10,6
1	0,15	0,46	0,08	0,15	0,15
0	0,53	0,40	0,07	0,00	0,00

Evaluación

El análisis de este modelo muestra que, para 7 de un total de 9 registros (el 77.77%), coinciden los valores predichos con los reales. Por lo tanto, este modelo predecirá con un porcentaje del 77.77 % (*Tabla 4.18*).

Tabla 4.18*. Modelo 7.2. Evaluación*

Resultados para el campo de resultado Purpose Objective_yes_no_Max

Comparando $B-Purpose Objective_yes_no_Max con Purpose Objective_yes_no_Max

'Partición'	2_Comprobación	
Correctos	7	77,78%
Erróneos	2	22,22%
Total	9	

Entonces, en base a la *Figura 4.32* y la evaluación de la red bayesiana se concluye que los registros y las kcal quemadas semanalmente son variables características para lograr el éxito.

5. RESULTADOS

En el punto anterior se han comentado y especificado los modelos obtenidos para cada una de las dimensiones analizadas, a fin de determinar variables significativas de nutrición y ejercicio, en el tratamiento de pérdida de peso.

A partir de estas variables y de los modelos realizados, en este apartado, se van a exponer los resultados obtenidos que se configuran en un conjunto de reglas en las que se basará el diseño del sistema de recomendaciones personalizado.

5.1. Tablas resultantes

A fin de examinar, algunas de las variables de nutrición y actividad física consideradas como significativas en el apartado anterior, a lo largo de las visitas 5, 6 y 7, se han realizado las figuras resultantes que se exponen a continuación. En estas, se muestran los registros realizados entre semana (tabla izquierda) y los que se realizan en fin de semana (tabla derecha). Y en cada una de ellas, el eje x representa la visita a la que corresponden estos registros, mientras que el eje y representa la variable evaluada. Por último, el color rojo indica aquellos sujetos que sí han logrado el éxito, mientras que el color azul son los que no lo han hecho.

Tablas resultantes de nutrición

- Variables del Modelo 1

En la *Figura 5.1*, se observa que aquellos sujetos que no han logrado el éxito son, los que, en los 3 primeros meses de tratamiento, han consumido menor porcentaje de grasas saludables. Y conjuntamente con lo obtenido en el modelo 1, se concluye que la disminución de estas grasas es un hábito de fin semana, tan sólo en la visita 5. Así pues, se considera que se deberían de ingerir más del 60% de grasas saludables al día.

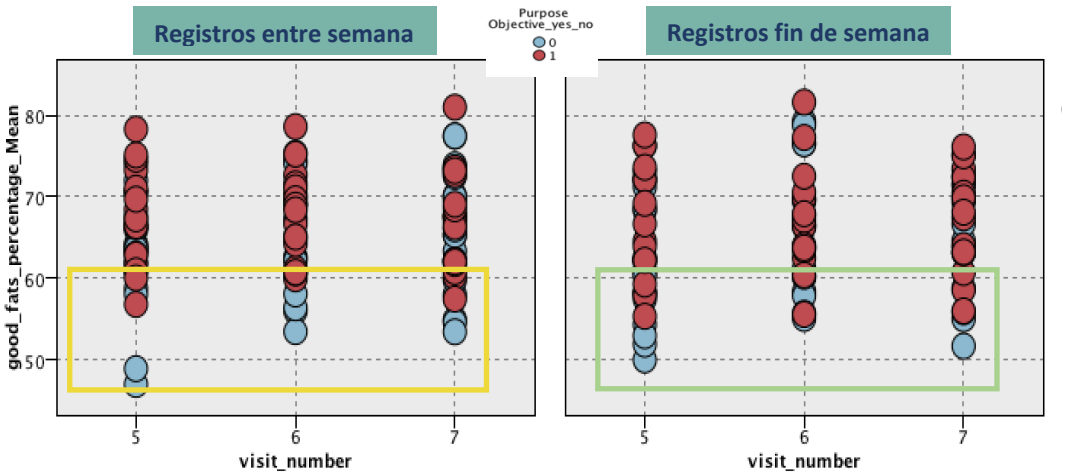


***Figura 5.1**. Grasas saludables en los 3 primeros meses.*

- Variables del Modelo 2 y 4

En la *Figura 5.2*, se muestra que aquellos sujetos que han logrado el éxito son, los que, en los 3 primeros meses de tratamiento, han consumido mayor cantidad de proteínas. Y, conjuntamente con lo obtenido en el modelo 2 y 4, se considera que es importante cumplir con la prescripción médica de una ingesta adecuada de proteínas, así como, que un consumo inferior a 60 g/día es arriesgado si se quiere lograr el tratamiento.

Figura 5.2. *Proteínas consumidas en los 3 primeros meses.*

- Variables del Modelo 5

En la *Figura 5.3*, se contempla que aquellos sujetos que no han logrado el éxito son, los que, en los 3 primeros meses de tratamiento, han consumido grandes cantidades de alimentos no saludables. Aunque en la prescripción médica, se recomienda que no se deben tomar este tipo de comida, se considera, en base a lo obtenido en el modelo 4, el consumo de alimentos no saludables no debería ser superior a 100 kcal/día. Así como, se observa que en el fin de semana se incrementa la ingesta de este tipo de alimentos.

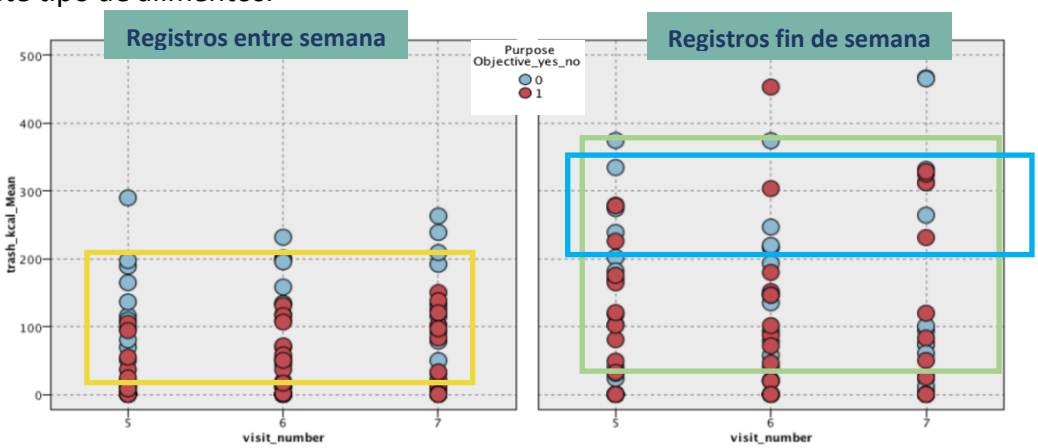


Figura 5.3. *Alimentos no saludables consumidos en los 3 primeros meses.*

- Variables del Modelo 6

En la *Figura 5.4*, se examina que aquellos sujetos que no han logrado el éxito son, los que, en los 3 primeros meses de tratamiento, han consumido mayor cantidad de alcohol, concretamente los fines de semana. Y conjuntamente con lo obtenido en el modelo 6, se considera que es perjudicial la ingesta superior a 128.8 kcal a fin de conseguir el propósito del programa.

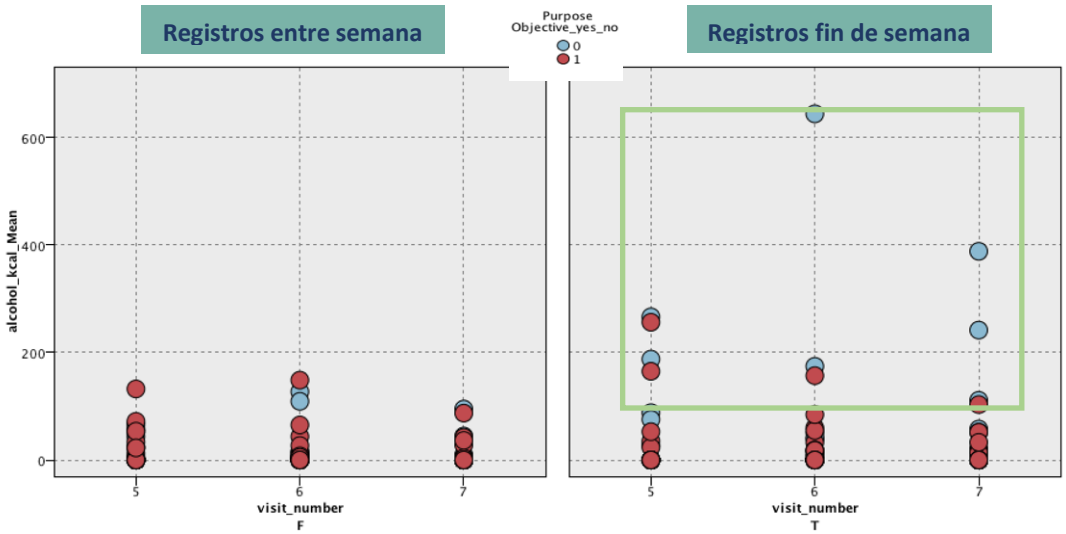


Figura 5.4. *Alcohol ingerido en los 3 primeros meses.*

Tabla resultante de ejercicio

En la *Figura 5.5*, se examina que aquellos sujetos que no han logrado el éxito, son los que menos kcal han quemado en el transcurso de su progreso. Por tanto, se razona que cumplir con la prescripción de ejercicio, es necesario a fin de alcanzar la finalidad del programa.

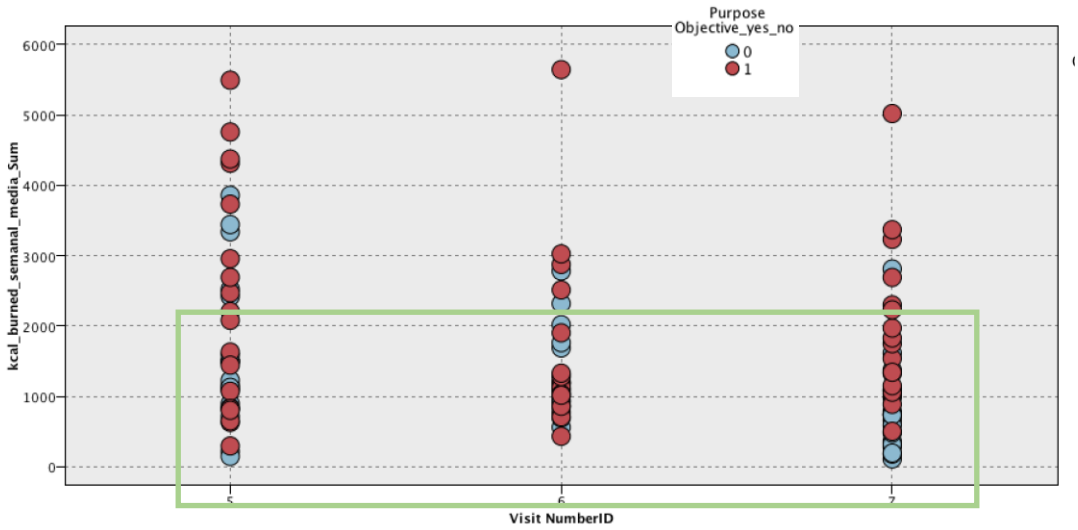


Figura 5.5*. Kcal quemadas en los tres primeros meses.*

5.2. Conjunto de reglas de decisión

Posteriormente, respetando las pautas actuales de nutrición y actividad física, avaladas por la comunidad médica, se han evaluado e interpretado los resultados obtenidos en el apartado 4 y las tablas resultantes anteriores. Y, junto con el equipo experto se han determinado el siguiente conjunto de reglas de decisión:

Reglas de nutrición

En la *Tabla 5.1*, se especifican las reglas de decisión, a partir de las cuáles se enviaría una recomendación de nutrición personalizada **entre semana**.

Tabla 5.1. *Conjunto de reglas de nutrición entre semana.*

Modelo	Regla
2	✓ *vegetables_kcal_Mean* < 72 kcal
4	✓ *trash_kcal_Mean* > 106 kcal ✓ *proteins_Mean* <= 60 g
6	✓ *dairy_kcal_Mean* <= 171kcal

En la *Tabla 5.2*, se especifican las reglas de decisión a partir de las cuáles se enviaría una recomendación de nutrición personalizada **en fin de semana**.

Tabla 5.2*. Conjunto de reglas de nutrición en fin de semana.*

Modelo	Regla
1	✓ *good_fats_percentage_Mean* < 61 % y *Age_GROUP*<=3 ✓ *good_fats_percentage_Mean* < 61 % y *Some Personal disease in treatment = 1*
3	✓ *fruits_units_Mean* < =0
5	✓ *Age_group* =2 o *Age_group* =4 y *Alcohol_kcal_Mean* >128.8 kcal
Figura 5.3	✓ *Trash_kcal_Mean* > 150 kcal

Reglas de ejercicio

En la *Tabla 5.3*, se especifican las reglas de decisión a partir de las cuáles se les enviaría una recomendación de nutrición personalizada en fin de semana.

***Tabla 5.3**. Conjunto de reglas de ejercicio.*

Modelo	Regla
7	✓ *kcal_burned_semanal_media_Sum* <= 1215 kcal ✓ *registros_semanales_media_Sum* <6

5.3. Sistema de recomendaciones personalizadas

Finalmente, se propone un sistema de recomendaciones basado en conocimiento. Este sistema, mediante una función de similitud calcularía en que cantidad, los valores de nutrición, ejercicio y datos personales de cada paciente coinciden con alguna de las reglas de decisión detalladas en las *Tablas 5.1, 5.2 y 5.3,* a fin de proporcionar recomendaciones individualizas y generales, tal y como se muestra en la *Figura 5.6.* Así cómo, se necesitaría disponer de técnicas de aprendizaje para asegurar la eficacia del sistema.

Estas recomendaciones, se adaptarían a las necesidades de cada paciente, y serían de tipo informativas, con el fin de aconsejarles platos, alimentos, actividades, así como aportarles información del ámbito de nutrición, el deporte y estilos de vida saludables, y en caso conveniente, advertencias referidas a un estilo de vida insano, que siempre irían acompañadas de una recomendación para contrarrestar.

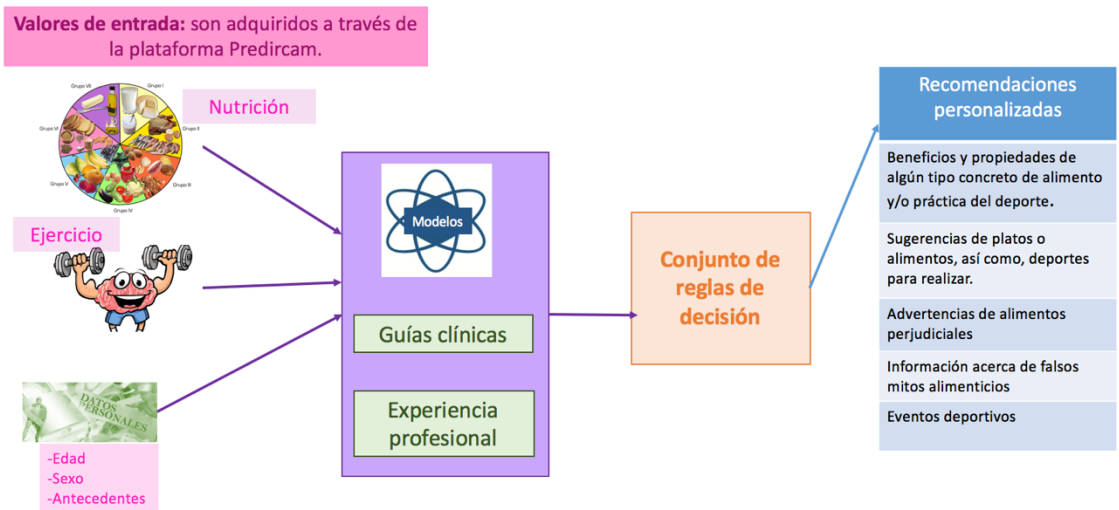


***Figura 5.6**. Sistema de recomendación personalizado.*

Existirían dos momentos para el envío de las recomendaciones, unas proporcionadas entre semana y otras específicas de fin de semana, pero ambas tendrían el propósito de que el sujeto pueda aprender más sobre la nutrición y el deporte, se motivara, mantuviera las ganas de mejorar sus costumbres día a día y fuese adquiriendo unos hábitos de vida saludable.

Las recomendaciones estarían compuestas por diferentes tipos de información, sujetas a unos determinados requisitos, y para ello se contaría con una base de datos, generada por expertos en nutrición y actividad física para personas con obesidad, que contendría un amplio repositorio de recomendaciones diferentes a fin de asegurar que, aunque un sujeto repitiera los mismos requisitos durante todo el progreso, no recibiera la misma recomendación en ningún momento. Así cómo, esta base de datos se iría actualizando cada cierto periodo de tiempo con el propósito de proporcionar a los usuarios conocimiento o recomendaciones lo más actualizadas posibles.

Recomendaciones entre semana

Los tipos de recomendaciones proporcionadas entre semana, serían 3 y se encontrarían sujetas a los requisitos que se muestran en la *Tabla 5.4*

Tabla 5.4. *Recomendaciones personalizadas entre semana.*

Tipo de recomendación	Requisitos	Categorías de información enviada
Específica	Cumple alguna de las **reglas de decisión**, especificadas en las *Tablas 5.1 y 5.3.*	✓ Beneficios y propiedades de algún tipo concreto de alimento y/o práctica del deporte. ✓ Sugerencias de platos o alimentos, así como, deportes para realizar. ✓ Advertencias de alimentos perjudiciales que ha consumido y/o el peligro de la inactividad en base al incumplimiento con los límites, junto con recomendaciones a fin de contrarrestar.

Generalizada I	No cumple ninguna regla de decisión, pero tiene algún valor de nutrición y/o ejercicio que incumple los límites recomendados.	✓ Beneficios y propiedades de aquellos alimentos y/o ejercicio incumplido. ✓ Sugerencia de algún plato o alimento y/o alguna determinada práctica de deporte para reducir o aumentar aquellos valores incumplidos. ✓ Advertencias de alimentos perjudiciales que ha consumido y/o el peligro de la inactividad en base al incumplimiento con los límites, junto con recomendaciones a fin de contrarrestar.
Generalizada II	No cumple ninguna regla de decisión y tiene todos los valores dentro de los límites recomendados.	✓ Información/recomendaciones generalizadas de nutrición y deporte: - Falsos mitos alimenticios - Errores alimenticios frecuentes - Beneficios del deporte

Por último, para las notificaciones del tipo generalizadas I, en caso que el usuario presentase varios valores de incumplimiento, se insertaría una sola recomendación referida a la nutrición y una referida al deporte.

Al contar con numerosos y diferentes valores en el apartado de nutrición, se establecería un orden de prioridad, para valores incumplidos por exceso y otro para los valores incumplidos por carencia. Así pues, la recomendación enviada sería referida al valor más prioritario, pero en caso que la semana siguiente se volviese a repetir este valor, si hubiese otros, se confeccionaría la recomendación a partir del segundo valor más prioritario, y así sucesivamente.

Recomendaciones de fin de semana

Los tipos de recomendaciones establecidas en el fin de semana, serían 2 y se encontrarían sujetas a los requisitos que se muestran en la *Tabla 5.5*.

***Tabla 5.5**. Recomendaciones personalizadas fin de semana.*

Tipo de recomendación	Requisitos	Categorías de información enviada
Específica	Cumple alguna de **las reglas de decisión de fin de semana** (*Tabla 5.2*) y/o no ha realizado ninguna actividad durante la semana.	✓ Beneficios y propiedades de algún tipo concreto de alimento y/o práctica del deporte. ✓ Sugerencias de platos o alimentos y/o deportes para realizar. ✓ Advertencias de alimentos perjudiciales que ha consumido y/o el peligro de la inactividad en base, junto con recomendaciones de con que alimento y/o deporte se podría contrarrestar.
Generalizada	No cumple ninguna regla.	✓ Información/recomendaciones generalizadas de nutrición y deporte: - Deportes que realizar en fin de semana. - Eventos deportivos acontecidos en el fin de semana. - Advertencias generalizadas de los excesos producidos en fin de semana.

En el caso del tipo de recomendación específica referida al deporte, ante la repetición del requisito de no se ha realizado ninguna actividad durante la semana, el orden de las categorías de información que se enviaría al paciente sería el siguiente: la primera vez que se llegara a viernes por la tarde y hasta ese momento, no se hubiese introducido ningún registro de ejercicio, se le enviaría una sugerencia de deporte a realizar en fin de semana, la segunda vez que volviera a ocurrir este requisito, se le informaría de los beneficios y propiedades de la práctica del deporte y finalmente la tercera vez se le enviaría una advertencia de la inactividad junto con una recomendación a fin de contrarrestar este efecto. Este orden se produciría consecutivamente, teniendo en cuenta que ninguna de las recomendaciones se repetiría, debido al repositorio con que el que se contaría, que aseguraría que cada una de ellas fuera diferente.

Propuesta

En base a todo lo expuesto con anterioridad, se expone modificar el sistema de recomendaciones actual de la plataforma Predircam, por un **informe personalizado** en formato PDF (en el que se incluirían las recomendaciones personalizadas), unos **consejos de fin de semana individualizados** y un **conjunto de alertas.**

Informe personalizado

El informe semanal sería enviado cada semana, al correo personal de cada usuario y constaría de 4 partes bien diferenciadas:

Nutrición *(Figura 5.7)*

En este apartado se insertaría un resumen de su consumo medio semanal mediante los iconos que se utilizan en la plataforma de PREDRICAM, acompañados de los valores medios de ingesta. Por tanto, se mostrarían los valores de los porcentajes de cada tipo de alimento y los porcentajes de cada macronutriente consumido en formato de barras. Así cómo, la cantidad media consumida de agua, bebidas saludables, no saludables, alcohol, leche, frutas, grasas, aceite, frutos secos y alimentos no saludables. Por último, también se incluiría un tacómetro calórico dónde se indicaría el porcentaje de kcal consumidas respecto las kcal prescriptas por el médico.

En esta misma sección, se encontraría un recuadro en el que aparecerían las felicitaciones y/o recomendaciones en función de los valores medios consumidos. Estas felicitaciones y recomendaciones son las que se encuentran en el sistema de recomendación actual.

Ejercicio físico *(Figura 5.8)*

En este apartado, se indicarían los valores medios de kcal quemadas, sesiones realizadas y minutos empleados en actividad física, junto con los valores prescritos para cada uno de los apartados. Así cómo, para hacer saber de manera rápida al usuario, si está cumpliendo o no con el objetivo fijado, se utilizaría un icono visual en forma de cara, y cuya expresión dependerá de la cantidad de cumplimiento con respecto los valores prescritos.

En esta misma sección, se encontraría un recuadro en el que aparecerían las felicitaciones y/o recomendaciones en función de kcal quemadas, sesiones y tiempo de ejercicio realizado. Estas felicitaciones y recomendaciones son las que se encuentran en el sistema de recomendación actual.

Peso, IMC e ICC *(Figura 5.9)*

Este apartado aparecería cada dos/tres semanas y en él se mostraría el valor del peso junto con una flecha indicando si se ha aumentado, disminuido o mantenido respecto a otro tomado anteriormente. Por otra parte, los valores de BMI, e ICC aparecerían junto con un punto: naranja, amarillo o verde. Y estos puntos, en el caso del BMI se referirían al estado nutricional del sujeto. Mientras que en el caso del ICC, estos puntos se referirían al riesgo de padecer enfermedades cardiovasculares.

Recomendaciones personalizadas

Estas recomendaciones serían individualizadas, aparecerían en un recuadro y estarían condicionadas al conjunto de requisitos detallados en la *Tabla 5.4.*

Consejos individualizados de fin de semana

Los consejos individualizados de fin de semana, estarían sujetos al conjunto de requisitos detallados en la *Tabla 5.5*, y estos, se enviarían cada viernes por la tarde a través de la plataforma. Los mensajes de consejo de fin de semana estarían basados en aportar advertencias y sugerencias a los usuarios, con intención de prever excesos y/o carencias de, nutrición y/o actividad física, que son propensos a ocurrir en el fin de semana. Además, en ellos se incluiría una frase motivadora y otra en la que se les desearía un buen fin de semana.

Alertas

Se mantendrían las alertas que existen actualmente, sin ninguna modificación, respetando los periodos y condiciones establecidas.

Procedimiento

El primer **informe personalizado** se enviaría en la semana 3, en este informe aparecerían todas las secciones, a excepción de la del peso, ya que la primera vez que se introduciría el apartado de peso, IMC e ICC, sería en la semana 5, y en él, el valor de peso que se mostraría sería el peso más reciente después de la visita 6 y en caso que no hubiese, se consideraría el peso de la visita 6 como el actual. Así pues, para la determinación, de la orientación de la flecha indicando si el sujeto ha subido, mantenido o bajado de peso, se compararía este peso con el peso de la visita 4, sabiendo que para considerar que ha disminuido de peso la diferencia ha de ser positiva y se ha reducido el peso en al menos un 1% a la semana (sin considerar la semana 1 ya que es una semana de prueba de la plataforma). Mientras que si la diferencia es negativa y se ha aumentado en al menos un 1% a la semana se consideraría aumento, en caso contrario, ha mantenido su peso.

Y el valor de ICC mostrado, sería el último registrado después del obtenido en la visita 6, en caso de que no hubiera ninguno posterior, se mostraría el ICC de la visita 6 como el actual. Para los demás días en los que el apartado peso quede incluido en el informe, se consideraría el peso registrado más reciente como peso actual y se compararía, siempre y cuando se cumpliesen las siguientes condiciones: hubiese un peso espaciado de al menos 1 semana, y que no fuese un peso precedente al informe previo. En caso que no se cumplieran estas dos condiciones, el apartado peso quedaría excluido del informe.

El informe sería enviado semanalmente, concretamente cada lunes a lo largo de la mañana, al correo personal de cada usuario. Y el apartado del peso sería incluido de forma fija la semana siguiente a la realización de una visita y después, en caso posible, cada dos/tres semanas.

Por otra parte, desde la semana 3, cada viernes por la tarde, en base a los requisitos detallados en la *Tabla 5.5*, se enviaría mediante la plataforma, unos **consejos de cara al fin de semana.**

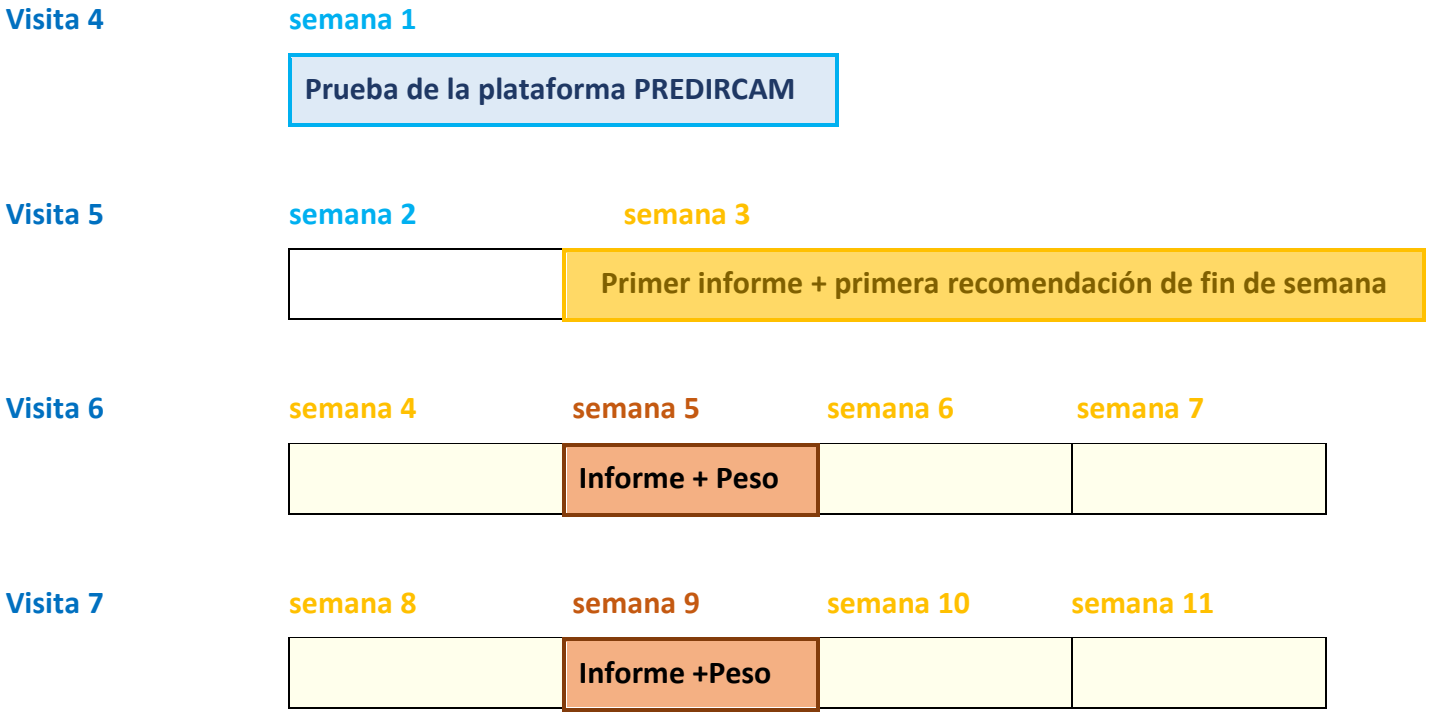
Finalmente, las **alertas** se enviarían con la frecuencia establecida actualmente, mediante la plataforma.

Con objetivo de resumir lo comentado, se ha elaborado una planificación en la que se muestra cuándo sería enviado cada uno de los puntos anteriores, mediante colores Identificativos siendo:

Amarillo claro: Indica cada una de las semanas en las que se les enviaría el informe personalizado y, las indicaciones de fin de semana.

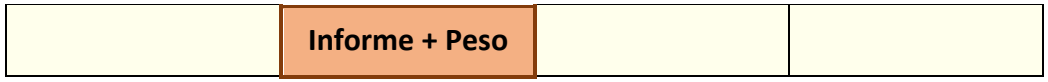
Naranja: Indica cada una de las semanas en las que el informe iría acompañado, de forma obligatoria, del apartado peso, ya que se compararía el peso de la visita realizada la semana anterior con el peso de la visita anterior en caso que no hubiese ningún peso posterior, es decir, el peso de la visita 7 se compararía con el de la 6, suponiendo que entre la 6 y la 7 no hubiese registrado ningún peso entre medias, ya que en caso que existiese, se tomaría este, siempre y cuando cumpliese con las condiciones expuestas. Así como, se les enviarían los consejos de fin de semana.

Rosa claro: Indica aquellas semanas en las que se debería incluir el peso en el informe, siempre y cuando se cumplieran las dos condiciones citadas respecto al peso.

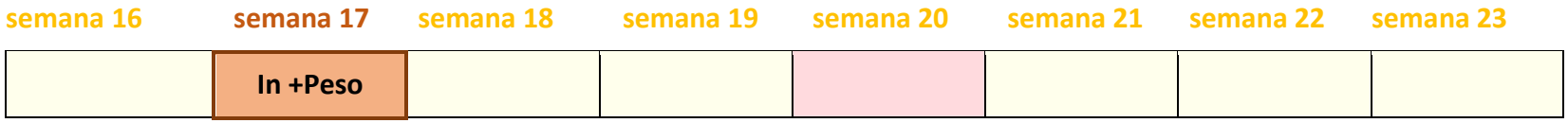


Visita 8

3 meses

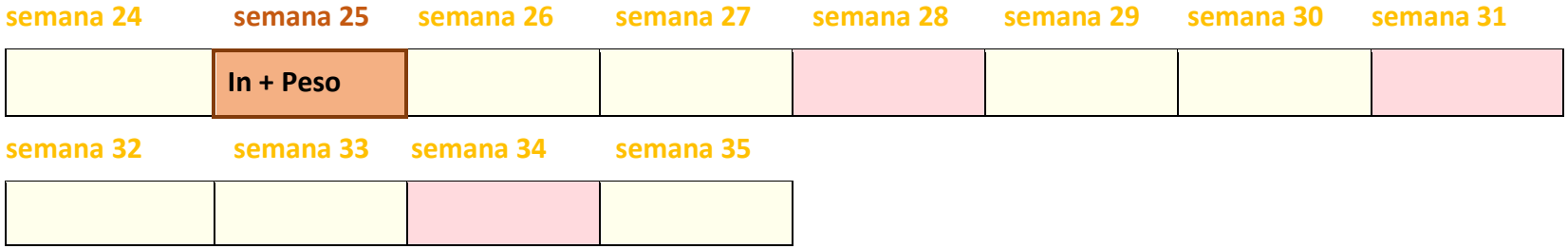


Visita 9



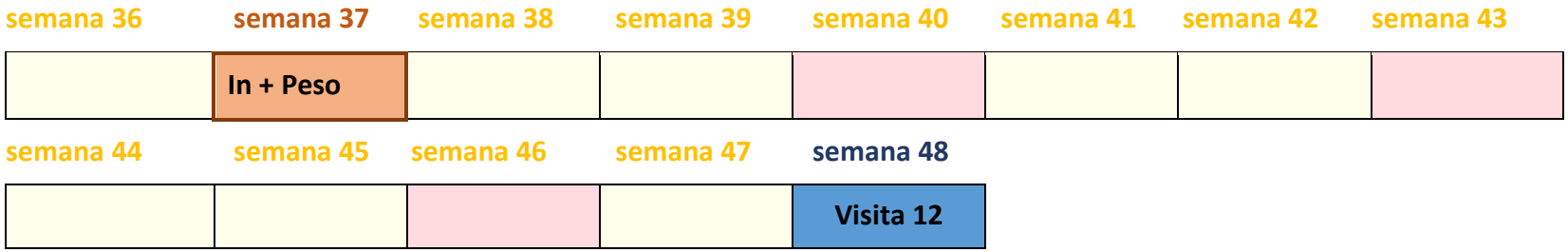
Visita 10

6 meses



Visita 11

9 meses



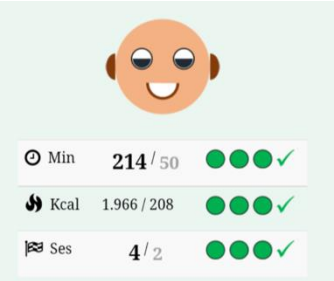
Informe semanal

Nutrición

Figura 5.7. *Resumen de nutrición del informe semanal [19].*

Felicitaciones y/o recomendaciones de nutrición que se encuentran en el sistema actual basadas en el cumplimiento de los límites establecidos.

Ejercicio



Felicitaciones y/o recomendaciones de ejercicio que se encuentran en el sistema actual, basadas en el cumplimiento de los límites

Figura 5.8*. Resumen de ejercicio del informe semanal [19].*

Peso

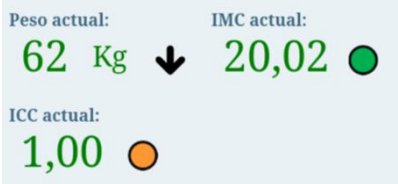


Figura 5.9*. Resumen de peso del informe semanal [19].*

Información

Recomendaciones personalizadas, basadas en la *Tabla 5.4.*

6. CONCLUSIONES

Una vez finalizado el desarrollo del trabajo, es imprescindible valorar y analizar las conclusiones a las que se han llegado. En primer lugar, se analizan las actividades desarrolladas más importantes en dirección hacia la consecución del objetivo de este Trabajo Final de Grado y posteriormente se realiza una valoración global.

Contexto del problema

Este proyecto queda enmarcado en el ecosistema tecnológico del que formaría parte: la plataforma Predircam (Plataforma Inteligente para la Monitorización, Tratamiento y Prevención Personalizados de la Diabetes Mellitus, el Riesgo CardioMetabólico y la Insuficiencia Renal). Así pues, el primer paso consistió en conocer de cerca la problemática, como es el síndrome metabólico, a la que se encuentra asociada, así como, cada uno de sus módulos y su sistema de recomendaciones.

Por tanto, el objetivo ha sido comprender cada una de las funcionalidades de la plataforma, así como, las variables de nutrición y ejercicio que se registran en ella, las cuáles, conforman la base de datos de análisis. Para ello, fue necesario el estudio de trabajos y artículos específicos, además, se contó con la ayuda del equipo experto de esta herramienta.

De esta manera, se ha podido estar en contacto con profesionales médicos y nutricionistas, lo que ha permitido adquirir unos conocimientos y unas pautas en el ámbito de la nutrición y la actividad física que son de gran utilidad, para el posterior análisis de los datos. Así cómo, se ha tomado conciencia de la importancia de estilos de vida saludables y de la gran amenaza que suponen los factores de riesgo del síndrome metabólico.

Proceso de extracción de conocimiento

Este apartado ha supuesto la parte más importante y extensa del trabajo, así pues, un[1]a vez comprendido y estudiado el entorno, se llevó a cabo el análisis de la base de datos disponible. Para ello, fue necesario formarse en el proceso de extracción de conocimiento, por lo que, se examinó bibliografía concreta, así como, se estudiaron las diferentes técnicas predictivas y descriptivas de minería de datos.

Todas los métodos estudiados y revisados fueron empleados en la creación de varios modelos predictivos, a partir de los cuales, se considera que de entre todos los grupos de alimentos, la cantidad de proteínas, vegetales y lácteos ingeridas durante los 3 primeros meses, son variables significativas en el tratamiento de pérdida de peso, así como, el consumir más de un específico valor de alimentos no saludables, es determinante para no conseguir el éxito.

Por otra parte, se ha obtenido que aquellos sujetos que no consiguen el propósito del tratamiento, poseen unos hábitos característicos durante el fin de semana. Concretamente, se ha examinado que estos sujetos pertenecen al grupo de edad más joven, y estos hábitos, se basan en disminuir la cantidad de frutas, de grasas buenas y de aumentar la ingesta de alcohol.

Diseño de un sistema de recomendación personalizado

Finalmente, partiendo del sistema de recomendación actual de la plataforma, y de los criterios obtenidos de los modelos realizados, se han elaborado un conjunto de reglas en las que se fundamenta el diseño de un sistema de recomendación personalizado basado en conocimiento. Asimismo, se ha realizado una propuesta de modificación del sistema de recomendación que dispone la plataforma.

Conclusiones generales

Algunos de los resultados obtenidos coinciden con lineamientos existentes en guías clínicas, y otros se espera que aporten información extra que permita mejorar los resultados de pacientes semejantes a los involucrados en el estudio clínico de Predircam. Así pues, con el apoyo de una plataforma tecnológica, ha sido posible recabar datos y realizar el análisis de los mismos para apoyar la toma de decisiones, tanto a nivel de autogestión de la salud, como en el seguimiento médico. Y, se considera que el uso de esta, en el progreso del tratamiento y el cumplimiento de la prescripción de actividad física y de nutrición, es determinante a fin de conseguir el propósito del programa.

En este trabajo se ha realizado un análisis de un subset de datos, utilizando la herramienta IBM SPSS Modeler y esto representa una primera aproximación hacia la obtención de modelos predictivos que podrán seguir perfeccionándose para su aplicación en un conjunto más amplio de datos, a fin de extraer conocimiento en forma de reglas, que podrán mejorar el sistema de recomendaciones de la plataforma.

Por último, para que la realización de este trabajo haya sido posible, han resultado de gran utilidad todas aquellas asignaturas cursadas durante la carrera que han tenido una especial aplicación durante toda esta etapa, siendo en este caso las asignaturas relacionadas con los sistemas de información y comunicaciones en la sanidad, así como con la biología y la estadística. También, la realización simultánea de las prácticas en la empresa PRONAF, han sido de gran provecho, debido a que permitieron experimentar con el análisis de la base de datos de pacientes con sobrepeso y obesidad del proyecto de investigación que realizó esta empresa. Así pues, gracias a la ejecución de este trabajo y a las prácticas, se ha ampliado el entendimiento en el área de minería de datos, debido al estudio del proceso de extracción de conocimiento.

Además, se han asimilado una gran cantidad de conocimientos en las técnicas de análisis y concretamente en la generación de modelos predictivos con la herramienta de IBM SPSS Modeler, lo que ha supuesto un logro personal el que, sin apenas información y documentación, se ha conseguido llevar a cabo un análisis de los datos, pudiendo obtener unas determinadas reglas y diseñar un sistema de recomendaciones personalizadas.

Finalmente, se puede considerar que se han cubierto los objetivos del trabajo Fin de Grado y, puede afirmarse que, con el desarrollo del presente trabajo, se ha obtenido conciencia de la importancia de diseñar e implementar herramientas que faciliten al usuario adquirir hábitos de vida saludables y le proporcionen información de interés, ayudándole así a mejorar su estado de salud.

7. TRABAJOS FUTUROS

Tal y como se ha comentado, este proyecto se ha basado en realizar una primera aproximación a partir de un subconjunto de la base de datos, debido a que el estudio clínico, mediante el uso de la plataforma Predircam, se encuentra sin finalizar.

Por tanto, es objeto de futuros trabajos el análisis de la base de datos completa, es decir, un análisis en el que se incluya los registros de todas las visitas, así como, se realice el análisis de aquellos campos que no han sido tratados en el desarrollo de este proyecto, como son: campos de texto libre, tabla de notificaciones, progreso de los valores de analíticas, evaluación de los cuestionarios, etc. Todo ello a fin de obtener un modelo predictivo con mayor precisión, así como, entrenar los modelos, obtenidos en la elaboración de este trabajo, con diferentes conjuntos de datos, y comprobar, re-evaluar y, posiblemente, reconstruirlos, a fin de extraer conocimiento.

Asimismo, se propone como posible mejora el diseño y desarrollo de un sistema de recomendaciones híbrido. Un sistema que combinase el conocimiento, es decir, que se base en unas reglas de decisión adquiridas a partir del análisis de los datos, y, además, sea colaborativo, lo que significa que a partir de los valores del usuario activo le recomiende aquello que ha resultado positivo en usuarios similares, es decir, que le recomiende los campos de usuarios semejantes, "correlación entre personas".

8. BIBLIOGRAFÍA

8.1. Referencias bibliográficas

[1] Organización Mundial de la Salud (OMS). (2017). *Enfermedades no transmisibles* [Online]. Disponible en: http://www.who.int/mediacentre/factsheets/fs355/es/

[2] Organización mundial de la salud (OMS), "Informe sobre la situación mundial de las enfermedades no transmisibles", 2011. [Online]. Disponible en: http://www.who.int/nmh/publications/ncd_report_summary_es.pdf

[3] Organización de Naciones Unidas, "Proyecto de resolución presentado por el Presidente de la Asamblea General. Declaración Política de la Reunión de Alto Nivel de la Asamblea General sobre la Prevención y el Control de las Enfermedades No Transmisibles", 16 de septiembre de 2011. [Online]. Disponible en: http://www.who.int/fctc/reporting/party_reports/spain_annex27_political_declaration.pdf

[4] Barrera MP, Pinilla AE, Cortés E, Mora G y Rodríguez MN, "Síndrome metabólico: una mirada interdisciplinaria", *Revista Colombiana de Cardiología*, vol. 15, no. 3, 2008. [Online]. Disponible en: http://www.scielo.org.co/pdf/rcca/v15n3/v15n3a4.pdf

[5] Alberti KGM.M, Zimmet PZ y Shaw JE, "The metabolic syndrome: a new world-wide definition from the International Diabetes Federation consensus", *The Lancet*, 2005. [Online]. Disponible en: http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(05)67402-8/fulltext

[6] K.G.M.M. Alberti, Robert H. Eckel and others, "Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity", *Amer. Heart Assoc.* , 2009. [Online]. Disponible en: http://circ.ahajournals.org/content/120/16/1640

[7] Fundación española del corazón. *Hipertensión y riesgo cardiovascular*. [Online]. Disponible en: http://www.fundaciondelcorazon.com/prevencion/riesgo-cardiovascular/hipertension-tension-alta.html

[8] Cuídate plus .(2016). *Obesidad* [Online]. Disponible en: http://www.diabetes.org/es/vivir-con-diabetes/tratamiento-y-cuidado/el-control-de-la-glucosa-en-la-sangre/hiperglucemia.html

[9] Simmons RK, Alberti KG, Gale AM, Colagiuri S, Tuomilehto J, Qiao Q, et al, "The metabolic syndrome: useful concept or clinical tool?", 2009. [Online]. Disponible en: https://link.springer.com/article/10.1007%2Fs00125-009-1620-4

[10] Salim Yusuf, Srinath Reddy, Stephanie Ôunpuu and Sonia Anand, "Global Burden of Cardiovascular Diseases: Part I: General Considerations, the Epidemiologic Transition, Risk Factors, and Impact of Urbanization", *Amer. Heart Assoc. ,* 2001. [Online]. Disponible en: http://circ.ahajournals.org/content/104/22/2746/tab-figures-data

[11] Sociedad Española de Cardiología (SEC). (2016). *La enfermedad cardiovascular encabeza la mortalidad en España* .[Online]. Disponible en: https://secardiologia.es/comunicacion/notas-de-prensa/notas-de-prensa-sec/7266-la-enfermedad-cardiovascular-encabeza-la-mortalidad-en-espana

[12] Sociedad Española de Arteriosclerosis (SEA), "Las enfermedades cardiovasculares y sus factores de riesgo en España: hechos y cifras", 2007. [Online]. Disponible en: http://www.se-arteriosclerosis.org/assets/informe-sea-2007.pdf

[13] Secchi JD, "Historia de la actividad física y su relación con la salud: La influencia Jeremiah Morris & Ralph Paffenbarger", *Revista Brisas de Salud*, N°2, pp.12-15, 2012.

[14] Miguel A. Aguirre-Urdanetaa, Joselyn J. Rojas-Quinteroa,c, Marcos M. Lima-Martínezb,d, "Actividad física y síndrome metabólico: Citius-Altius-Fortius", *Diabetologia*, Venezuela, 2012. [Online].Disponible en: http://www.elsevier.es/es-revista-avances-diabetologia-326-articulo-actividad-fisica-sindrome-metabolico-citius-altius-fortius-S1134323012001433

[15] World Health Organization, "Global Strategy on Diet, Physical Activity and Health. Obesity and Overweight", 2008. [Online]. Disponible en: http://www.who.int/dietphysicalactivity/M&E-2008-web.pdf

[16] WHO (World Health Organization), "Telemedicine. Opportunities and developments in member states" , 2010. Disponible en: http://www.who.int/goe/publications/goe_telemedicine_2010.pdf

[17] Ministerio de Sanidad y Consumo INSALUD, " Plan de telemedicina del INSALUD", Madrid, 2000. [Online]. Disponible: http://www.ingesa.msssi.gob.es/estadEstudios/documPublica/pdf/telemedicina.pdf

[18]J. Tapia, J. M. Iniesta, V. Alcántara, G. García, G. Navarro, C. González y Mª Elena Hernando, "PREDIRCAM 2. Plataforma Tecnológica para la Prevención de la Diabetes Tipo 2 y el Riesgo CardioMetabólico", en *XXXIII Congreso Anual de la Sociedad Española de Ingeniería Biomédica*, Madrid, 2015.

[18] González C, Herrero P, Cubero JM, Iniesta JM, Hernando ME, García-Sáez G, Serrano AJ, Martinez-Sarriegui I, Perez-Gandia C, Gómez EJ, Rubinat E, Alcantara V, Brugués E, Chico A, Mato E, Bell O, Corcoy R y de Leiva A, "PREDIRCAM eHealth platform for individualized telemedical assistance for lifestyle modification in the treatment of obesity, diabetes, and cardiometabolic risk prevention: a pilot study (PREDIRCAM 1)", Julio, 2013.

[19] J. Tapia, "Diseño y desarrollo de una aplicación móvil multiplataforma para la autogestión de la dieta y la promoción de estilos de vida saludables", Trabajo Final de Grado, Madrid, 2015.

[20] J.M. Rodríguez, *Cómo hacer inteligente su negocio: Business Intelligence a su alcance*, 1ªed., Patria, México, 2014.

[21] J. Hernández, Mª. J. Ramírez, y C. Ferri, *Introducción a la Minería de Datos.* Pearson Prentice Hall, España, 2004, pp. 3-39.

[22] D. Jannach, M. Zanker, A. Alexander y G. Friedrich, *Recommender System*. New York, NY, USA: Cambridge University Press, 2011. Disponible en: https://books.google.es/books?id=eygTJBd_U2cC&printsec=frontcover&dq=recommender+systems&hl=en&ei=kuiNTKRDhc2zBtiR0csB&sa=X&oi=book_result&ct=result&redir_esc=y#v=onepage&q&f=false

8.2. Bibliografía de consulta

-Web oficial de IBM SPSS Modeler. Disponible en: https://www.ibm.com/es-es/marketplace/spss-modeler

-*Manual nodos de origen, proceso y salida de IBM SPSS Modeler 18.0*, 2016.

-*Manual nodos de modelado de IBM SPSS Modeler 18.0*, 2016.

- J. Hernández, Mª. J. Ramírez, y C. Ferri, *Introducción a la Minería de Datos*. Pearson Prentice Hall, España,2004.

-C.Pérez Lopez, *Data Mining con herramientas de IBM SPSS Modeler*.

9. ANEXOS

9.1. IBM SPSS Modeler

Introducción *IBM SPSS Modeler.*

IBM SPSS Modeler, es un conjunto de herramientas de minería de datos que permite desarrollar modelos predictivos precisos de forma rápida y proporcionar inteligencia predictiva al usuario. Proporciona un rango de algoritmos avanzados y técnicas de análisis, para mejorar la toma de decisiones. Con un diseño que sigue el modelo CRISP-DM, estándar del sector, IBM SPSS Modeler admite el proceso completo de minería de datos, desde los propios datos hasta obtener los mejores resultados.

Este software ofrece una gran variedad de métodos de modelado procedentes del aprendizaje automático, la inteligencia artificial y el estadístico. Los métodos disponibles en la paleta de modelado permiten derivar nueva información procedente de los datos y desarrollar modelos predictivos. Cada método tiene ciertos puntos fuertes y es más adecuado para determinados tipos de problemas.

IBM SPSS Modeler es una herramienta integrada de Big Data, business intelligence y minería de datos que incluye diversas fuentes de datos (ASCII, XLS, ODBC…), una interfaz visual basada en procesos/flujos de datos (streams), distintas herramientas de minería de datos (correlación, reglas de asociación, regresión, segmentación, redes neuronales, árboles de decisión…), manipulación de datos (muestreo, combinación y separación…), combinación de modelos, visualización de datos, exportación de modelos a distintos lenguajes (C, SPSS, SAS…), exportación de datos integrada a otros programas (XLS) y generación de informes.

El entorno de Modeler está basado en nodos, iconos y formas que representan operaciones individuales de sus datos, que se van disponiendo y conectando para formar un flujo o stream, también conocido como ruta (*Figura 9.1)*. Los streams pueden almacenarse en ficheros separados o en proyectos que engloban a varios de ellos que se puede cargar, guardar, modificar, reejecutar o reorganizar utilizando las opciones del menú Archivo de la pantalla de entrada de IBM SPSS Modeler y que son independientes de las fuentes de datos.

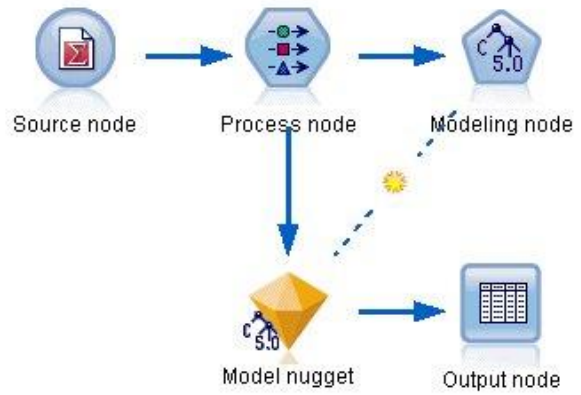


***Figura 9.1**. Ruta básica con 4 nodos interconectados.*

1. Paletas de nodos.

Modeler presenta varias paletas que **clasifican los nodos** (*Figura 9.2*) en seis categorías:

- Favoritos: nodos más utilizados.
- Orígenes: nodos para obtener los datos de trabajo (fuentes de datos).
- Oper. con registros: operadores para modificar o combinar registros (filas) de distintas fuentes. Es decir, selecciones y combinaciones.
- Oper. con campos: operadores para modificar o combinar campos (columnas).
- Gráficos: nodos que muestran gráficamente los datos antes y después del modelado.
- Modelado: tipos de modelos/patrones que puede generar Modeler.
- Resultado: presentación de tablas, análisis de modelos, estadísticas, exportación de datos.
- Exportar: exportación de información a otros formatos y aplicaciones.
- IBM SPSS Statistcs: conexión con otros procedimientos de IBM SPSS Statistics.

***Figura 9.2**. Paletas de nodos en IBM SPSS Modeler.*

2. Gestores.

En la parte superior derecha de la ventana se encuentra el **panel de gestores** (*Figura 10.3*). Este panel cuenta con 3 pestañas:

- Rutas: para abrir, cambiar nombres, guardar o eliminar las rutas creadas en una sesión.
- Resultados: contiene una serie de archivos, como gráficos y tablas, generados mediante operaciones de rutas en Modeler. Puede mostrar, guardar, cambiar el nombre y cerrar las tablas, gráficos e informes que se enumeran en esta pestaña.
- agradecim

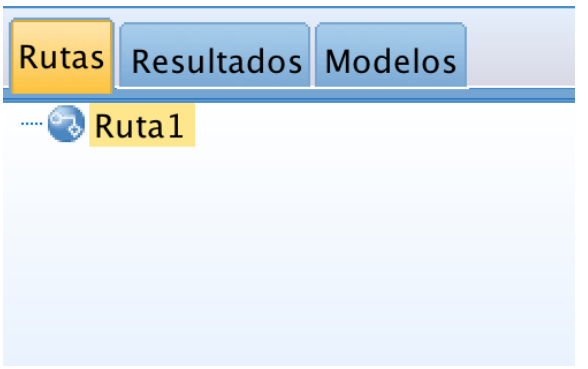


***Figura 9.3**. Gestores de IBM SPSS Modeler.*

3. Proyectos.

En la parte inferior derecha de la ventana se encuentra el **panel de proyecto**, que se utiliza para crear y administrar proyectos de minería de datos. Existen dos formas de ver los proyectos que se crean en IBM SPSS Modeler:

- Pestaña CRISP-DM (*Figura 9.4*): permite organizar los proyectos según el proceso CRISP-DM, una metodología independiente y probada en el sector, para mejorar la organización y comunicación de los esfuerzos.
- Pestaña Clases *(Figura 9.5):* permite organizar el trabajo en IBM SPSS Modeler de forma categórica, por los tipos de los objetos que se hayan creado. Esta vista resulta útil al realizar un inventario de datos, rutas y modelos.

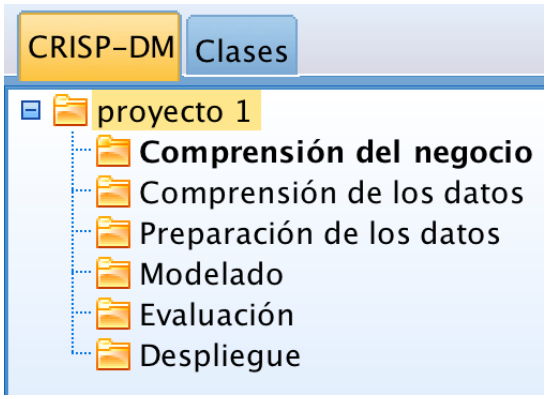


Figura 9.4. *Vista CRISP-DM.*

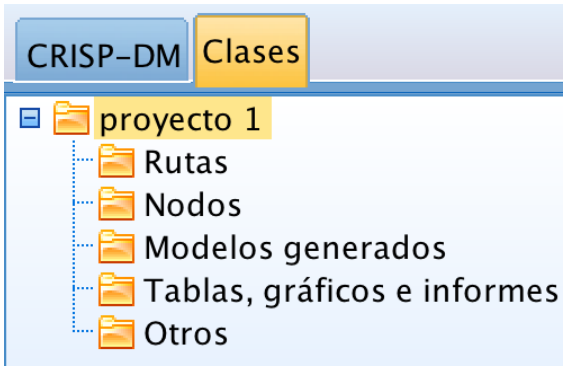


Figura 9.5*. Vista Clases.*

4. Barra de herramientas.

En la parte superior de la ventana de IBM SPSS Modeler hay una barra de herramientas (*Figura 9.6*) con iconos que proporciona una serie de funciones muy útiles.

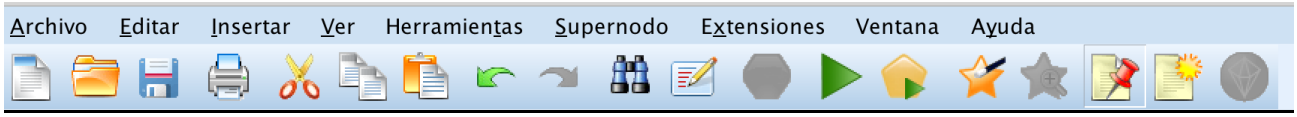


Figura 9.6*. Barra de herramientas de IBM SPSS Modeler..*

Proceso de KDD

El proceso de descubrimiento de conocimiento en bases de datos (KDD) llevado a cabo, para la extracción de conocimiento con IBM SPSS Modeler, es el CRISP-DM (Cross-Industry Standard Process for Data Mining).

El CRISP-DM (*Figura 9.7*) es un proceso cíclico que está formado por 6 fases, estas son:

1. Compresión del negocio.

Determinación de **objetivos** de la minería de datos, evaluación de la situación y la producción de un plan del proyecto.

2. Compresión de los datos.

Comprender cuáles son los **orígenes** de los datos y las **características** de dichos orígenes. Incluye la recopilación de datos iniciales, la descripción, la exploración y verificación de la calidad de datos.

3. Preparación de los datos.

Después de catalogar los orígenes, es necesario **preparar** los datos para el análisis. Esto incluye: selección, limpieza, construcción, integración y asignación de formato de los datos.

4. Modelado o data mining.

Se utilizan sofisticados **métodos de análisis** para extraer información de los datos. Implica la selección de las técnicas de modelado, la generación de diseños de comprobación y la generación de modelos de evaluación.

5. Evaluación.

Una vez elegidos los modelos, ya se puede **evaluar** la forma en que los **resultados** del análisis pueden ayudar a lograr los objetivos. Los elementos principales de esta fase son la evaluación de los resultados, la revisión del proceso de minería de datos y la determinación de los siguientes pasos.

6. Despliegue.

Esta fase incluye el despliegue, el control y el mantenimiento del plan, la producción de un **informe final**, así como la revisión del proyecto.

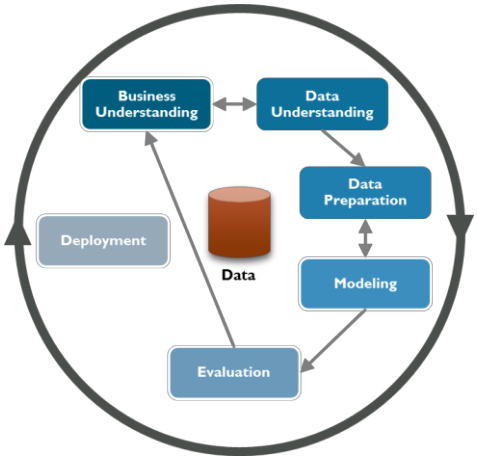


Figura 9.7. *Ciclo del análisis de datos.*

Tanto las decisiones realizadas como la información recogida durante la fase de modelado, generalmente, pueden hacer que se desee configurar de nuevo ciertas partes de la fase de preparación de datos, las cuales podrán, acto seguido, presentar nuevos problemas de modelado. Por tanto, tal y como se observa en la *Figura 10.7,* ambas fases se retroalimentan hasta que ambas se resuelvan de manera adecuada. De igual manera, la fase de evaluación puede hacer que se desee evaluar de nuevo la compresión original y puede hacer caer en la cuenta de que se ha estado intentando responder a la pregunta equivocada. En este punto, se puede revisar, ya con un mejor objetivo en mente, la compresión del negocio e iniciar de nuevo el resto del proceso.

En concordancia con el ciclo del proceso de KDD, el proceso de extracción de conocimiento consta de varias fases, como la preparación de datos (selección, exploración, limpieza y transformación), su análisis mediante técnicas de modelado adecuadas y la evaluación y valoración de los modelos seleccionados para extraer el conocimiento.

Preparación de los datos.

El proceso comienza con la **compresión** de un problema práctico de inteligencia de datos que surge de la actividad cotidiana. Una vez identificado el problema y determinado los objetivos, es necesario acudir a los datos óptimos para su resolución de acuerdo a los posibles modelos aplicables (fase de **comprensión de los datos**).

Posteriormente, los datos necesitan una **preparación** adecuada para que sean utilizables de modo adecuado en la aplicación de los modelos teóricos identificados para la resolución de nuestro problema, esto conlleva:

Selección

En la fase de selección:

- Se integran y recopilan los datos.
- Se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas.
- Se identifican y seleccionan las variables relevantes en los datos.
- Se aplican las técnicas de muestreo adecuadas.

De esta forma se pasa de unos datos brutos inicialmente disponibles a unos datos objetivo adecuados al problema o investigación.

A la fase de selección se le pueden asociar todos los nodos de la paleta *Orígenes* para obtención de datos de diversas fuentes y varios nodos de las paletas *Oper. con registros* y *Oper. con campos*.

Exploración y limpieza

En la fase de exploración:

- Se utilizan técnicas de análisis exploratorio de datos.
- Se deduce la distribución de los datos, simetría y normalidad.
- Se analizan las correlaciones existentes en la información.

Dentro de esta fase se encuentra la fase de limpieza en la que:

- Se detectan y tratan la presencia de valores atípicos (outliers).
- Se imputa la información faltante o valores perdidos (data missing).
- Se eliminan los datos erróneos e irrelevantes.

A la fase de exploración y limpieza se le pueden asociar todos los nodos de la paleta *Gráficos* para visualizar los datos gráficamente, los nodos de la paleta *Resultados* que permiten obtener información acerca de los datos y modelos mediante la presentación de tablas, estadísticas, exploración de datos…, y algunos nodos de la paleta *Oper. con campos:* nodo tipo, nodo filtrar, derivar, rellenar y el nodo de variables globales.

Transformación (modificación).

Después de la fase de exploración, de la que se obtienen los datos procesados, el proceso de extracción del conocimiento contempla la fase de transformación o modificación de los datos, en la que:

- Se utilizan técnicas de reducción y aumento de la dimensión.
- Se aplican técnicas de discretización y numerización.
- Se realiza escalado simple y multidimensional.

IBM SPSS Modeler dispone de nodos con finalidades de modificación de datos dentro de la paleta *Operaciones con registros,* los cuáles se usan para realizar cambios a nivel de registro y dentro de la paleta Operaciones con campos, que contiene muchos nodos útiles para la transformación de los datos.

Minería de datos: Análisis de datos o modelización.

Una vez preparados los datos, la siguiente fase será la de modelado,, que consiste en la utilización de técnicas predictivas y descriptivas de minería de datos.

Las tareas de esta fase son las siguientes:

1. Se selecciona la técnica de modelado.
2. Se construye el modelo de pruebas
3. Se implementa el modelo.
4. Se evalúa el modelo.

IBM SPSS Modeler ofrece una gran variedad de métodos de modelado procedentes del aprendizaje automático, la inteligencia artificial y el estadístico. Los métodos disponibles en la paleta de modelado permiten derivar información procedente de los datos y desarrollar modelos predictivos. Cada método tiene ciertos puntos fuertes y es más adecuado para determinados tipos de problemas.

Los métodos de modelado se dividen en cuatro categorías (*Figura 9.8*):

Figura 9.8. *Categorías de la paleta Modelado.*

Servidor de análisis.

Al seleccionar esta categoría, sólo se muestran los nodos (*Figura 9.9*) que se pueden ejecutar en IBM SPSS Analytic Server.

Figura 9.9*. Nodos de Analytic Server.*

Modelos de clasificación.

Los modelos de clasificación (*Figura 9.10*) usan el valor de uno o más campos de entrada para predecir el valor de uno o más campos de destino. Las técnicas de modelado incluyen aprendizaje automático de las máquinas, inducción de reglas, identificación de subgrupos, métodos estadísticos y generación de varios modelos.

Figura 9.10*. Nodos de clasificación.*

Modelos de asociación.

Los modelos de asociación (*Figura 9.11)*, encuentran patrones en los datos en los que una o más entidades se asocian con una o más entidades, es decir, los modelos construyen conjuntos de reglas que definen estas relaciones. Aquí los campos de los datos pueden funcionar como entradas y destinos. Se podrían encontrar estas asociaciones manualmente, pero los algoritmos de asociaciones lo hacen mucho más rápido, y pueden explorar patrones más complejos.

***Figura 9.11**. Nodos de asociación.*

Modelos de segmentación.

Los modelos de segmentación (*Figura 9.12)*, dividen los datos en segmentos o clústeres de registros que tienen patrones similares de campos de entrada. Cómo sólo se interesan por los campos de entrada, los modelos de segmentación no contemplan el concepto de campos de salida o destino.

Estos modelos de agrupación en clústeres se centran en la identificación de grupos similares y en el etiquetado de registros según el grupo al que pertenecen. Y su valor viene determinado por su capacidad de capturar agrupaciones interesantes en los datos y proporcionar descripciones útiles de dichas agrupaciones.

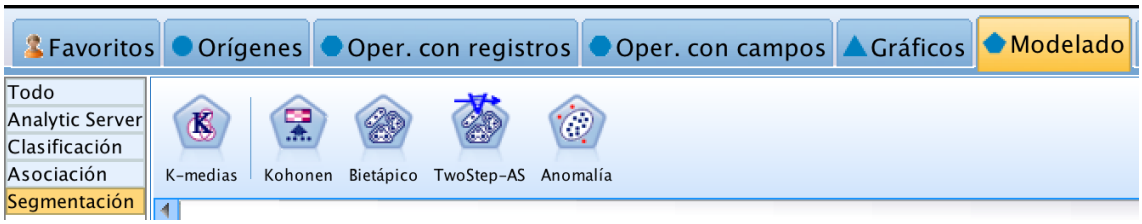


***Figura 9.12**. Nodos de segmentación.*

Evaluación.

En la quinta fase, la fase de evaluación, se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la explotación del modelo.

Normalmente los proyectos de Data Mining no terminan en la implantación del modelo (sexta fase), sino que se deben documentar y presentar los resultados de manera comprensible en orden a lograr un incremento del conocimiento. Además, en la fase de explotación se debe de asegurar el mantenimiento de la aplicación y la posible difusión de los resultados.