UNIVERSITAT DE BARCELONA

Enfermedades mentales y su relación con el microbioma intestinal

Francisco Gutiérrez Romero

Máster Universitario en Bioinformática y Bioestadística

TFM - Bioinformática y Bioestadística Área 3

Andreu Paytuví Gallart

Ferran Padros Carrasco

Junio 2020

” *Pronto será un pecado para los padres tener un hijo que lleve la pesada carga de la enfermedad genética.*”

BOB EDWARDS

” *Las ’leyes del pensamiento’ no solo dependen de las propiedades de las células cerebrales, sino del modo en que están conectadas.*”

MARVIN MINSKY

” *Si en otras ciencias hay que llegar a la certeza, sin duda, y a la verdad sin error, nos corresponde poner las bases de los conocimientos en las matemáticas.*”

ROGER BACON

FICHA DEL TRABAJO FINAL

Título del trabajo:	Enfermedades mentales y su relación con el microbioma intestinal
Nombre del autor:	Francisco Gutiérrez Romero
Nombre del consultor:	Andreu Paytuví Gallart
Nombre del PRA:	Ferran Prados Carrascos
Fecha de entrega (mm/aaaa):	06/2020
Titulación:	Máster Universitario en Bioinformática y Bioestadística
Área del trabajo final:	TFM – Bioinformática y Bioestadística Área 3
Idioma del trabajo:	Castellano
Palabras clave	Disbiosis, microbioma, microbiota, intestino-cerebro, salud, mental, enfermedad, transtorno, comportamiento, predicción, aprendizaje automático, Machine learning

Resumen del trabajo

La relación entre el microbioma intestinal y algunas de las enfermedades mentales más comunes, es una realidad. Es una certeza que no se puede constatar qué conjunto de unidades taxonómicas operativas son responsables de estas afectaciones mentales. Hoy en día se recurre a diferentes tratamientos, en función del trastorno y de la gravedad del mismo. Lo más importante en estos casos es poder llegar a adelantarse a esta situación degenerativa y poder predecir a tiempo cualquiera de estas enfermedades.

Este trabajo se centra en esto mismo, en la predicción de enfermedades mentales que pueda llegar a desarrollar cualquier persona, basándose principalmente en el microbioma intestinal y en ciertos hábitos de comportamiento. Cada individuo tiene su propia microbiota, formada por un determinado tipo de bacterias, algunas con más presencia que otras, lo que se conoce como *enterotipo intestinal*.

A partir del análisis de los datos proporcionados, se ha desarrollado una aplicación de software de predicción para detectar enfermedades mentales. Dicha herramienta supone la parte práctica de un estudio basado en el entrenamiento automático para el ajuste de un modelo predictor. La aplicación permite al usuario realizar ajustes antes del entrenamiento, como la elección de los algoritmos de clasificación, decidir el porcentaje de balanceo de los datos, o incluso, insertar la taxonomía de un grupo de pacientes para ser diagnosticados de un posible trastorno mental.

Durante el trascurso de este documento se describirán los métodos y técnicas utilizadas, y las caracterísiticas de las herramientas necesarias. Finalmente, se presentan los resultados fruto del experimento, que demuestran la eficacia del análisis en su totalidad.

Abstract

The relationship between the gut microbiome and some of the most common mental illnesses, it is a reality. It is a certainty that it cannot be verified which set of operational taxonomic unit are responsible for these mental disorders. Today it is used different treatments, depending on the disorder and its severity. The most important in these cases is to be able to anticipate this degenerative situation and to be able to predict any of these diseases.

This work focuses on this very thing, on predicting mental illnesses that can be developed by any person, mainly based on the intestinal microbiome and in certain behavior habits. Each individual has their own microbiota, formed by a certain type of bacteria, some with more presence than others, what is known as *intestinal enterotype*.

From the analysis of the data provided, a software application has been developed of prediction to detect mental illnesses. This tool is the practical part of a study based on automatic training to fit a predictive model. The application allows the user to make adjustments before training, such as choosing the classification algorithms, decide the percentage of data balancing, or even insert the taxonomy of a group of patients to be diagnosed with a possible mental disorder.

This document shows the methods and techniques used and the characteristics of the necessary tools. Finally, the results of the experiment are presented, that demonstrate the effectiveness of the analysis in its entirety.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Contexto y justificación del Trabajo

Con frecuencia, las bacterias, virus y parásitos son considerados como una gran amenaza que debe ser eliminada. Es cierto que bacterias como *Yersinia pestis* o *Vibrio cholerae* han truncado un gran número de vidas humanas. Sin embargo, también cabe decir que otras como *Bacillus subtilis*, *Escherichia coli* o *Streptomyces venezuelae* han ayudado en la detección de enfermedades metabólicas e incluso son utilizadas para elaborar antibióticos, salvando de este modo muchísimas otras vidas. [3]

La microbiota, que es el conjunto de microorganismos que mantiene una relación de asociación con otro organismo de diferente especie, se aloja en diferentes partes del cuerpo humano. Cabe destacar por su importancia el microbioma intestinal, formado por el conjunto de genes de estos microorganismos.

Se sabe que existe una interacción entre el microbioma intestinal y el huésped, por lo tanto, de esta relación se podría obtener información relevante sobre el estado de la salud mental de éste. Así pues, algunas enfermedades pueden ser relacionadas con los desequilibrios de la microbiota intestinal o disbiosis. Por el momento, sólo se han dado los primeros pasos en establecer una correlación directa entre la disbiosis y el creciente número de enfermedades mentales, síndromes y/o alteraciones funcionales ligadas a estos desequilibrios.

Así pues, la detección precoz de cualquier enfermedad mental es de vital importancia para poder tomar medidas mientras aún se está a tiempo, proporcionando un tratamiento o incluso indicando qué hábitos deben ser corregidos para que la enfermedad no llegue a manifestarse y/o se retrase lo máximo posible.

Se estima que, actualmente, hay más de 1.100 millones de personas en todo el mundo que padecen algún tipo de transtorno mental. Entre las más comunes están los trastornos de ansiedad con 275 millones, depresión con 268 millones, trastorno bipolar con 40 millones o esquizofrenia con 21 millones, entre otras enfermedades [11]. Estos transtornos no solo incluyen características individuales como la capacidad de gestión de los pensamientos o la sociabilidad, sino que también hay factores culturales, condiciones laborales y/o económicas. Otros factores influyentes que pueden causar trastornos mentales pueden ser la alimentación, el estrés o la propia herencia genética. [12].

El objetivo principal de este proyecto es el de poder determinar, en función de la taxonomía de una población y ciertas características de un grupo de individuos, si éstos podrían llegar a desarrollar o no una o varias enfermedades mentales y de ser así, cuál es la probabilidad de que esto ocurra y qué géneros taxonómicos son los más relevantes.

A título personal, considero que lo más importante de la vida es ser consciente de la existencia de uno mismo. Si se pierde esa certeza, todo lo demás carecerá de importancia.

1.2. Objetivos del Trabajo

El objetivo principal de esta investigación es el de generar un modelo, a través del aprendizaje automático *Machine learning*, capaz de determinar la probabilidad de desarrollar ciertos tipos de enfermedades mentales en función del perfil taxonómico del microbioma intestinal de un sujeto.

A su vez, este modelo será parte de una herramienta informática para el *diagnóstico* e *investigación*, que facilitará el análisis a aquellos profesionales del sector que necesiten saber si un paciente puede llegar a desarrollar o no alguna enfermedad mental, permitiendo de esta forma el poder prevenir y analizar el posible desarrollo de la misma.

1.2.1. Objetivos generales

Los objetivos generales de este trabajo, constan de dos partes bien diferenciadas, tanto en su cometido como en la tecnología en la que se van a desarrollar.

La primera parte se centra, principalmente, en la obtención de un modelo predictor a través del aprendizaje automático. Se utilizará el lenguaje **R** en su versión *3.6.1*, para realizar todos aquellos pasos necesarios en el entrenamiento del modelo, así como en las comprobaciones de los resultados obtenidos en el mismo.

La segunda parte consiste en el desarrollo de una aplicación, a través del lenguaje de programación **Python** en su versión *3.6*, que permitirá al usuario introducir los datos a analizar y visualizar los resultados después de aplicar el modelo de predicción.

1.2.2. Objetivos específicos

Los objetivos específicos de este trabajo se dividen en tres bloques dependientes entre ellos, ya que requieren de una compenetración total entre sí.

El primer bloque se basa en la *preparación de los datos* para el análisis. Éste se centra, principalmente en la lectura, filtrado y selección de información proporcionada a través de diferentes ficheros, que serán utilizados posteriormente en el estudio.

El segundo bloque trata del *análisis de los datos* sin procesar y en la aplicación de técnicas de normalización y/o transformación en función de las necesidades del ajuste del modelo. Además de esto, mediante diferentes algoritmos de clasificación, se podrá determinar cuál o cuáles son más óptimos para este fin.

Por último, el tercer bloque se centra en el desarrollo de una aplicación de software, que mediante una interfaz de usuario permitirá facilitar la tarea de análisis e interpretación de los resultados obtenidos al usuario final.

1.3. Enfoque y método seguido

El procedimiento a seguir se basa en dos pilares principales: la preparación de los datos para ajustar el modelo y la predicción.

Ajuste del modelo

Inicialmente, se dispone de dos grupos de datos:

Un primer grupo de *Metadatos*, formado por un conjunto de ficheros en formato *JSON* que contienen las características de $\approx$ *16.000* individuos. Estas características son, entre otras, *edad, peso, hábito referente a la ingesta de bebidas alcohólicas, alergias, etc...* La fundamental es la información referente a si éste ha llegado a desarrollar o no alguna enfermedad mental, como *anorexia nerviosa, desorden bipolar, bulimia nerviosa, depresión, estrés post-traumático, esquizofrenia y/o abuso de sustancias*.

Un segundo grupo de datos almacena información *taxonómica*, perteneciente al mismo conjunto de individuos del grupo anterior. Dicha información contiene el porcentaje y la cantidad de *OTUs* que están presentes en cada paciente.

A partir de estos ficheros, se realizará un filtrado y una agrupación que formarán el conjunto final de datos a utilizar para realizar el ajuste del modelo. Una vez el conjunto de datos ha sido transformado y/o normalizado, se procederá a probar algunos de los algoritmos de aprendizaje automático más conocidos, como por ejemplo: *Boost Decision Tree*, *Support Vector Machine* y *Random Forest* que a partir de ahora, cuando se haga referencia a alguno de ellos, se utilizarán las siglas BDT, SVM y RF respectivamente, cuando se quiera hacer mención a estos algoritmos.

Seguidamente, se realizará una comparativa de rendimiento y se tomará aquel que mejor se adapte al conjunto de datos. En el caso de haber más de un clasificador óptimo, se trabajará con todos ellos. [1]

Predicción

El proyecto está basado en un tándem de tecnologías, donde el lenguaje *Python* se encargará de mostrar la interfaz de usuario y de entrada/salida de ficheros, y la herra-

mienta estadística *R* actuará como motor de cómputo.

Para obtener la predicción, se necesitará un fichero de taxonomía de la población a la que se le va a realizar dicha predicción, además de los correspondientes metadatos de cada uno de los pacientes. Es importante señalar que se deben tener en cuenta los mismos criterios de filtrado y selección de los pacientes que cuando se ha ajustado el modelo. Por último, el juego de datos resultante será utilizado con el modelo de entrenamiento para que éste realice la predicción final. En esta predicción se conocerá si un individuo padecerá o no una de estas enfermedades mentales, con una cierta probabilidad de que así sea. Además, se mostrarán aquellas OTUs más significativas en la obtención de esta predicción.

1.4. Planificación del Trabajo

La tarea de planificación del trabajo, se podría decir que ha sido tan complicada como el desarrollo del trabajo en si. Han habido algunas subtareas, sobre todo en la conexión de las dos tecnologías, donde se sabía de antemano que podrían surgir problemas, pero era realmente muy difícil poder realizar una valoración en horas del coste del desarrollo.

Nombre de tarea	Comienzo	Fin	Nombre de tarea	Comienzo	Fin
PEC0 - Definición de los contenidos del trabajo	mié 19/02/20	lun 02/03/20	PEC3 - Desarrollo del trabajo Fase 2	jue 23/04/20	lun 18/05/20
PEC1 - Plan de trabajo	mar 03/03/20	lun 16/03/20	Interfaz de usuario	jue 23/04/20	lun 04/05/20
PEC2 - Desarrollo del trabajo Fase 1	mar 17/03/20	mié 22/04/20	Comunicación Python/R	mar 05/05/20	mar 12/05/20
Preparación de los datos	mar 17/03/20	lun 23/03/20	Pruebas unitarias	mié 13/05/20	lun 18/05/20
Estadística Descriptiva	mar 24/03/20	jue 02/04/20	PEC4 - Cierre de la memoria	mar 19/05/20	mié 10/06/20
Algoritmos y generación de modelos	vie 03/04/20	mié 22/04/20	PEC5A - Elaboración de la presentación	jue 11/06/20	dom 14/06/20
			PEC5B - Defensa pública	mié 17/06/20	mié 24/06/20

Figura 1.1: Distribución del tiempo para el desarrollo de las tareas

Por ese motivo, la estimación de alguna de las partes del proyecto han necesitado de una pequeña bolsa de horas adicional, por si llegaba a complicarse mucho más de lo previsto inicialmente.

1.4.1. Preparación de los datos

Inicio tarea *17/03/2020* - Final tarea *23/03/2020*

Consiste en filtrar y preparar toda la información suministrada para poder realizar el análisis. El primer filtrado se basa en **descartar** aquellos individuos del conjunto de *Metadatos* que no cumplan unos ciertos requisitos. De esta forma se intenta eliminar el ruido lo máximo posible en los resultados. Estos requisitos de descarte son: *estar fuera de un rango de índice de masa corporal, haber padecido alguna enfermedad inflamatoria intestinal, ser diabético y haber tomado antibióticos en los últimos seis meses*.

El segundo filtrado se realiza sobre la matriz de taxonomía, y consiste en descartar aquellos individuos con un porcentaje de *Pseudomonas* superior al 3 % y un porcentaje de *(unkn.) Enterobacteriaceae(f)* también superior al 3 %. Una vez realizados los dos filtrados, aquellos individuos que aparezcan en ambos grupos a la vez, serán parte del conjunto final de datos que será analizado.

Por último, a la hora de preparar los datos, se observa que la información de cada uno de los pacientes es muy extensa y no toda es importante para este estudio. Por lo tanto, para no elevar los tiempos de cálculo o incluso para que algunas características no relevantes generen ruido, solo algunas serán incluidas en el conjunto final.

1.4.2. Estadística descriptiva y normalización

Inicio tarea *24/03/2020* - Final tarea *02/04/2020*

Se trata de saber cómo están distribuidos los datos obtenidos para determinar si es necesario o no realizar algún tipo de normalización y/o transformación. También permite tener una primera toma de contacto de cómo puede ir el proceso de entrenamiento y pruebas.

1.4.3. Algoritmos Machine Learning

Inicio tarea *03/04/2020* - Final tarea *17/04/2020*

Se debe decidir qué algoritmo de aprendizaje automático se va a utilizar, ya que sin conocer cómo están distribuidos los datos o qué tipo de correlación existe entre ellos,

sería aventurarse por un camino del que se desconoce si hay salida o no.

Es cierto que algoritmos como *Random Forest*, que se utiliza habitualmente y del que se obtienen buenos resultados si el conjunto de datos es lo suficientemente grande, podría ser a priori, la única propuesta. Pero también hay otros como *Support Vector Machine*, que en el caso de llegar a conseguir una buena distribución de los datos, se podrían obtener unos muy esperanzadores resultados.

Así pues, se realizarán pruebas con distintos clasificadores, y aquellos mejor adaptados al conjunto de datos serán los disponibles desde la interfaz de usuario.

1.4.4. Generación del modelo

Inicio tarea *18/04/2020* - Final tarea *22/04/2020*

Para cada uno de los algoritmos de clasificación se realizará los mismos pasos con el fin de completar el entrenamiento y generar el modelo. Estos pasos están formados por la preparación del grupo de entrenamiento y del grupo de pruebas, la predicción y evaluación del rendimiento del modelo y por último, una mejora del mismo si fuera posible.

1.4.5. Interfaz de usuario

Inicio tarea *23/04/2020* - Final tarea *04/05/2020*

Una vez se ha tomado la decisión de qué algoritmo o algoritmos se van a utilizar para generar el modelo, el siguiente paso es desarrollar una aplicación de software que sea capaz de facilitar todo este proceso al usuario final.

Esta aplicación ha de permitir realizar un número de simulaciones con distintos modelos entrenados, y que aquel con mejor resultado sea el que calcule la probabilidad de que un paciente pueda desarrollar o no una enfermedad mental.

1.4.6. Comunicación Python/R

Inicio tarea *05/05/2020* - Final tarea *12/05/2020*

El proceso de comunicación entre estas dos tecnologías permitirá repartir de una manera ordenada los pesos de cada una de las tareas.

Desde *Python* se enviarán los datos sin procesar, previamente filtrados, a *R* para que éste se encargue de todo el proceso de la generación del modelo. Una vez terminado, *R* deberá enviar los resultados a *Python* con la finalidad de poder informar al usuario de cómo ha ido todo el proceso de análisis.

1.4.7. Pruebas unitarias

Inicio tarea *13/05/2020* - Final tarea *18/05/2020*

Este tipo de pruebas permiten asegurar que cada uno de los pasos atomizados realizan aquello que se espera de ellos, que no es más que el cometido para lo que fueron diseñados.

1.4.8. Hitos

HITO 1: Una vez completado el primero de los objetivos, se realizará un informe para explicar los resultados provisionales obtenidos y cuál o cuáles de los algoritmos de aprendizaje automático pueden ser candidatos a ser utilizados en la aplicación de software.

HITO 2: Al finalizar el desarrollo de la aplicación de software, se documentará el funcionamiento de ésta y se decidirá si se añade alguna nueva funcionalidad o se elimina alguna de las ya existentes.

HITO 3: Entrega de la memoria final y preparación de la presentación virtual.

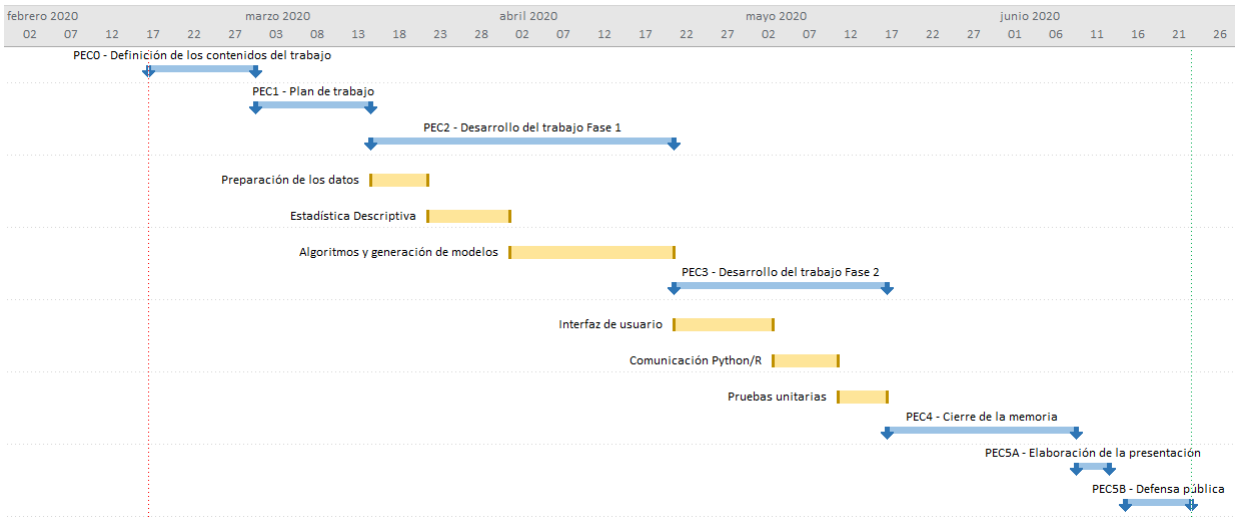

Figura 1.2: Diagrama de Gantt de las tareas a realizar con los tiempos de dedicación para cada una de las diferentes tareas previstas durante el desarrollo del proyecto

1.5. Breve sumario de productos obtenidos

1.5.1. Memoria

Se trata del presente documento, donde se explica cómo se ha desarrollado todo el trabajo final del máster en cada una de las diferentes fases.

1.5.2. Productos

Se entrega un producto final que consiste en una aplicación de software que sintetiza todo el estudio realizado. Además, a través de un repositorio se adjunta tanto el código fuente de la aplicación en lenguaje de programación *Python*, como el código fuente del análisis de los datos y la evaluación de los algoritmos en el lenguaje estadístico *R*.

1.5.3. Presentación

Mediante un documento en formato *power point*, se realizará una presentación y al mismo tiempo una explicación oral del desarrollo del trabajo.

1.5.4. Autoevaluación

Consiste en la defensa personal del trabajo ante un tribunal, que formulará preguntas o dudas sobre el mismo.

1.6. Breve descripción de los otros capítulos de la memoria

En el **segundo capítulo** se encuentra toda la metodología utilizada en el desarrollo del proyecto y los resultados de cada una de las fases.

En el **tercer capítulo** se exponen las conclusiones generadas a partir de los resultados del capítulo anterior.

En el **cuarto capítulo** se enumeran las palabras y expresiones del documento que pueden necesitar una definición más concreta.

En el **quinto capítulo** se hace referencia a las publicaciones de los autores mencionadas en alguna parte de la memoria.

En el **sexto capítulo** aparece información adicional como soporte para la mejor comprensión de alguna de las fases del trabajo.

Capítulo 2

Metodología

2.1. Introducción

Tiempo atrás, había la creencia de que los trastornos y las enfermedades mentales eran consecuencia de posesiones demoníacas o fruto de acontecimientos sobrenaturales. Más adelante, aparecieron centros de reclusión donde los enfermos, en lugar de ser curados, simplemente eran torturados con la excusa de la sanación. Incluso, a mediados del siglo pasado se utilizaron técnicas más avanzadas, comparándolas con la trepanación craneal, donde finalmente los afectados terminaban con daños cerebrales irreversibles [4].

Desde hace algo más de una década, la ciencia intenta comprender el complejo funcionamiento de la comunicación entre el intestino y el cerebro. Ya en 2011 se publicaron los primeros resultados que reflejaban la influencia de la microbiota en el comportamiento en ratones, simplemente alterando la composición de ésta. Ahora, un estudio ha podido demostrar por primera vez, que la microbiota intestinal en humanos tiene una relación directa con la salud mental. [5]

Actualmente, se pueden enumerar y clasificar un gran número de trastornos y enfermedades mentales, algunos muy antiguos y otros no tanto a consecuencia de la vida moderna. Por ejemplo: Alzheimer, demencia vascular, Parkinson, pensamientos suicidas, depresión, ansiedad, fobia social, trastorno por déficit de atención, trastorno obsesivo compulsivo, trastorno bipolar, trastorno de la conducta alimentaria, esquizofrenia, distor-

siones cognitivas, y muchas más que seguramente aún no está catalogadas y otras que aparecerán. Por lo tanto, el poder anticiparse a este hecho es de vital importancia para el buen funcionamiento de nuestras mentes. Disponer de herramientas capaces de predecir qué tipo de enfermedad o trastorno mental podríamos llegar a desarrollar, en función de un cierto escenario microbiómico, serían la punta de lanza ante esta eventualidad.

La disbiosis, hecho muy común actualmente, está vinculada a un gran número de enfermedades humanas, de las que cabe destacar por su importancia la salud mental y el comportamiento. Este hecho puede explicar, en gran medida, el papel de la conexión intestino-cerebro y su vital importancia.

Con el fin de relacionar los trastornos depresivos con una cierta bacteria, se ha observado una fuerte asociación entre las bacterias *Faecalibacterium* y *Coprococcus* y los indicadores de mayor calidad de vida. La merma de las bacterias *Dialister* y *Coprococcus* agudizan el trastorno depresivo, incluso después de corregir los efectos con antidepresivos. [5]

Otro tipo de relación son los problemas gastrointestinales, que se ven asociados con la mayoría de los casos de autismo, sugiriendo que no se trata únicamente de un trastorno psiquiátrico, sino que también tiene una base fisiológica. Un trastorno intestinal similar a la enfermedad de Crohn es, a veces, reportado en niños autistas. Se llegó a la conclusión de que los niños con autismo tienen niveles más bajos de bacterias intestinales *Veillonellaceae*, *Coprococcus* y *Prevotella* que aquellos sin la condición. [6]

La base del problema reside en que no se conoce la totalidad de la relaciones entre las enfermedades mentales y la carencia o merma de uno o varias bacterias que forman la microbiota intestinal. Una de las dificultades añadidas a esta investigación con seres humanos, es que éstos a diferencia de los animales por norma general, cambian de patrones de alimentación, toman antibióticos, pueden estar sujetos a estrés contínuo, lo que puede alterar la composición y/o concentración de la microbiota. Además, cabe añadir que ésta es, en cada individuo, única. [7]

2.2. Análisis de los datos

2.2.1. Recolección de los datos

Los datos recopilados provienen del proyecto *American Gut Project* o AGP. Este, junto a otros proyectos, están basados en el análisis del microbioma en todo el mundo, estableciendo los cimientos de la comprensión de los billones de microbios que moran en cada uno de nuestros cuerpos. Ahora, este proyecto, brinda la oportunidad de unirse a la investigación y comparar microbios de nuestro intestino con los de los intestinos de miles de personas en los Estados Unidos y en todo el mundo.

La información taxonónica recopilada para la realización de este trabajo, consta de una población de $\approx$ 16.000 participantes humanos y de $\approx$ 2.900 OTUs formando, de esta manera, una matriz taxonómica. Además, por cada paciente se dispone de información sobre ciertas características, pero sobre todo si padece algún tipo de transtorno de salud mental. Cuando se hizo la selección de pacientes se tuvo en cuenta este hecho, que fue confirmado por cada uno de ellos. Todos los datos de secuencia y las respuestas de los participantes anónimos se pueden encontrar en EBI bajo el proyecto *PRJEB11419* y el ID de estudio *Qiita 10317*. [14]

Taxonomía

La información **taxonómica** está formada por dos ficheros de texto, donde se incluye la relación entre cada uno de los pacientes y un conjunto de OTUs. En el fichero *genus.count.txt* se tiene en cuenta la relación entre el número de apariciones de cada taxón y cada paciente.

Por otro lado, el fichero *genus.perc.txt* muestra el porcentaje de apariciones de cada taxón en cada paciente respecto al total de la población. Este mismo fichero se ha utilizado en este proyecto para el filtrado de pacientes en función de un porcentaje, ya que precisamente contiene estos valores ya calculados y permite un mejor manejo de los datos.

Metadatos

Para cada paciente existe un conjunto de **metadatos** que se almacenan en ficheros con formato *JSON*. En éstos se encuentra información sobre ciertas características y hábitos relevantes, como por ejemplo si el paciente ha sufrido alguna enfermedad mental del tipo *anorexia nerviosa*, *desorden bipolar*, *bulimia nerviosa*, *depresión*, desorden debido a *estrés post traumático*, *esquizofrenia* o *abuso de sustancias*.

También se tiene en cuenta otro tipo de características tales como: *sexo*, *edad*, *altura*, *peso*, *frecuencia en la ingesta de vegetales*, *marisco*, *fruta*, *carne roja*, *si toma al menos un litro de agua diario*, *si consume bebidas azucaradas*, *alcohol*, *si fuma*, *cuánto tiempo duerme* y *si padece migraña*.

Se ha hecho una extensión de las características y se ha ampliado la lista a: si padece *reflujo estomacal*, alguna *enfermedad cardiovascular*, si toma *anticonceptivos*, con qué frecuencia *usa productos cosméticos*, cuánto tiempo hace que *se vacunó contra la gripe*, si ha padecido alguna *enfermedad de riñón*, de *hígado*, de *pulmón*, frecuencia en la *ingesta de huevos*, *queso*, *multivitaminas*, *probióticos*, *bocadillos salados*, *azúcar*, *vitamina b*, *vitamina d* y si se está *embarazada*.

2.2.2. Extracción y filtrado de los datos

El objetivo principal de este apartado es el de filtrar y preparar toda la información suministrada para poder realizar el estudio. Ésta incluye datos sobre la taxonomía de una población e información relevante sobre cada uno de los individuos, constituida por un conjunto de metadatos con características sobre cada uno.

El primer criterio de filtrado es taxonómico y éste se basa en la exclusión de aquellos pacientes que dispongan del género *Pseudomonas* superior al 3 % y/o un porcentaje del género *(unkn.) Enterobacteiaceae(f)* también superior al 3 %. Este filtrado reduce la lista de pacientes, aproximadamente, a la mitad.

Cada paciente dispone de un identificador único que se relaciona con un fichero de metadatos. A partir del contenido de ciertos campos incluídos en estos ficheros, se realiza el segundo filtrado. Aquellos pacientes que **cumplan** los siguientes requisitos, serán aquellos que formarán parte del juego de datos para el ajuste del modelo.

A continuación se enumeran:

- Tener el índice de masa corporal **BMI** comprendido entre los valores *[18, 30]*.
- No haber padecido alguna enfermedad inflamatoria intestinal **IBD**.
- No tener *diabetes*.
- No tener antecedentes de haber tomado antibióticos en los últimos *seis meses*.

Después de realizar este filtrado, el número de pacientes se reduce a unos 2.500, aproximadamente.

Por último, se ha realizado otro filtrado pensando en la optimización de los cálculos, basado en la eliminación de aquellas unidades taxonómicas sin relevancia, debido a la falta de resultados o incluso por carecer de ellos. Por lo tanto, ya que no van a realizar ninguna aportación al estudio y además van a incrementar el consumo de los recursos del sistema, se procederá a eliminar todas aquellas sin resultados.

2.2.3. Estadística descriptiva

Multivariante vs Univariante

En un primer momento, sin conocer las distribución de las variables de salida o dependientes, que son todas las enfermedades mentales que están englobadas en el conjunto de los datos, parecía factible utilizar las herramientas y métodos del análisis multivariante. El problema reside en que hay pacientes que manifiestan más de un trastorno al mismo tiempo.

	Anorexia	Bipolar	Bulimia	Depression	Ptsd	Schizophrenia	Subs. abuse
False	2655	2645	2644	2367	2631	2662	2657
True	17	27	28	305	41	10	15

Cuadro 2.1: Distribución de las enfermedades mentales dentro del juego de datos

En la tabla anterior, se muestra la relación entre padecer una enfermedad y de qué tipo o no padecerla.

Además, después de someter los datos a los distintos filtrados estadísticos, la cantidad de pacientes con múltiples enfermedades mentales no ha sido suficientemente significativa en cada uno de estos casos, lo que dificultaría en gran medida el ajuste del modelo. [2]

	Healthy	One	Two	Three	Four	Five
Counter	2334	258	65	7	6	2

Cuadro 2.2: Relación de múltiples enfermedades mentales en pacientes

En la tabla anterior, se puede observar que hay 2334 pacientes sin ningún tipo de trastorno, 258 que manifiestan un único trastorno, 65 pacientes que manifiestan dos trastornos al mismo tiempo, 7 pacientes que manifiestan tres trastornos a la vez, 6 pacientes que manifiestan cuatro trastornos, y por último, 2 pacientes que manifiestan cinco trastornos de los siete analizados.

De realizar un análisis multivariante, además de tener en cuenta los 7 tipos de enfermedades mentales, el sistema también debería valorar las distintas combinaciones posibles entre ellas. Esto sumado al bajo índice de pacientes que podrían presentar esta casuística, haría muy difícil encontrar patrones significativos y el sistema tendría una sensibilidad muy baja o casi nula.

Por ese motivo, se ha optado por el análisis univariante con una única variable como respuesta, siendo cada una de las siete enfermedades mentales existentes en los metadatos.

Distribución de los datos

Toda la información contenida en el juego de datos, después de aplicar los diferentes filtrados, se puede dividir en tres grupos: datos taxonómicos, características de los pacientes y enfermedades mentales.

Los datos **taxonómicos** se encuentran en forma de matriz donde las filas son los identificadores de los pacientes y las columnas cada una de los OTUs con sus correspondientes valores.

	unkn..Actinomycetaceae.f.	Varibaculum	unkn..Actinomycetales.o.	Actinopolyspora
ERR1596984	0.00000	0.00000	0.00555	0
ERR1596990	0.00000	0.00000	0.00000	0
ERR1596999	0.00000	0.00000	0.00000	0
ERR1597001	0.00000	0.00000	0.00000	0
ERR1597002	0.00000	0.00000	0.00000	0
ERR1597006	0.00000	0.18181	0.00000	0
ERR1597011	0.00000	0.00000	0.00000	0

Figura 2.1: Representación de la matriz taxonómica

En primer lugar, se calcula la variabilidad entre las variables que hacen referencia a la taxonomía y se concluye que existe una alta variabilidad entre la primera, que es *Bacteroides* con $287{,}6967$ y la última variable, que es *Ammoniibacillus* con $3{,}628886e^{-11}$. Por lo tanto, se requiere normalizar los datos para mitigar esta gran diferencia. Como apunte, indicar que en algunos análisis ha sido demostrado que el suministro de pectina o inulina de manzana da como resultado el enriquecimiento altamente específico de unidades taxonómicas operativas del género *Bacteroides*. [15]

Se comprueba cuáles son las bacterias con mayor valor de media con los datos sin normalizar.

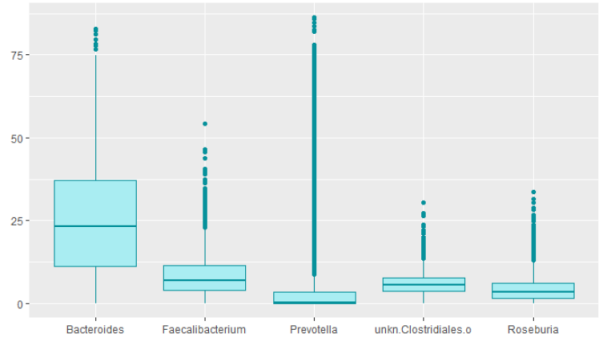

Figura 2.2: Boxplot OTUs sin normalizar con más apariciones

Unidad Taxonómica	Valor medio aparición
Bacteroides	25.379843
Faecalibacterium	8.608991
Prevotella	7.703624
unkn..Clostridiales.o.	6.059902
Roseburia	4.509156

Cuadro 2.3: Valor medio OTUs sin normalizar con más apariciones

Se comprueba cuáles son las bacterias con mayor valor de media con los datos después de la normalización.

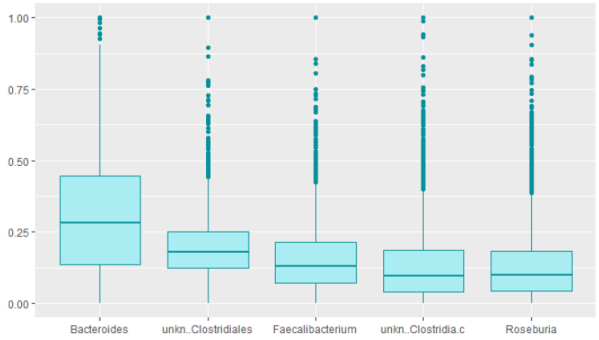

Figura 2.3: Boxplot OTUs normalizados con más apariciones

Unidad Taxonómica	Valor medio aparición
Bacteroides	0.3064847
unkn..Clostridiales.o.	0.1987499
Faecalibacterium	0.1585255
unkn..Clostridia.c.	0.1372455
Roseburia	0.1338626

Cuadro 2.4: Valor medio OTUs normalizados con más apariciones

Ya que estos datos son cuantitativos pueden ser normalizados para poder unificar el rango de todos los valores. Antes de la normalización, por ejemplo, *Bacteroides* tenía una varianza de $25{,}379$ y después de $0{,}3064$ o *Faecalibacterium* con una varianza de $8{,}6089$ antes de la normalización y con un valor de $0{,}1585$ después de ésta.

Las medidas de dispersión se utilizan para medir la variabilidad de una variable respuesta dentro de una muestra o población. La dispersión promedio que presenta cada dato es evaluado respecto a su valor medio. [18]

Mediante el *test de Shapiro-Wilk* se puso a prueba si los datos estaban basados en una distribución normal. El valor de las probabilidades fue menor al $5\,\%$, por lo que se concluye que los datos no siguen una distribución normal.

Posteriormente, se realiza un análisis de los componentes principales de los datos taxonómicos normalizados, concluyendo que las primeras 608 variables explican el $80\,\%$ de los datos, lo que antes explicaban 2117 variables. Para estimar el número K óptimo de clusters se puede aplicar el algoritmo de *K-means* para un rango de valores de *K* e identificar aquel valor a partir del cual la reducción en la suma total de varianza intra-cluster deja de ser sustancial. Se utiliza la *Distancia de Manhattan* para definir la distancia entre objetos.

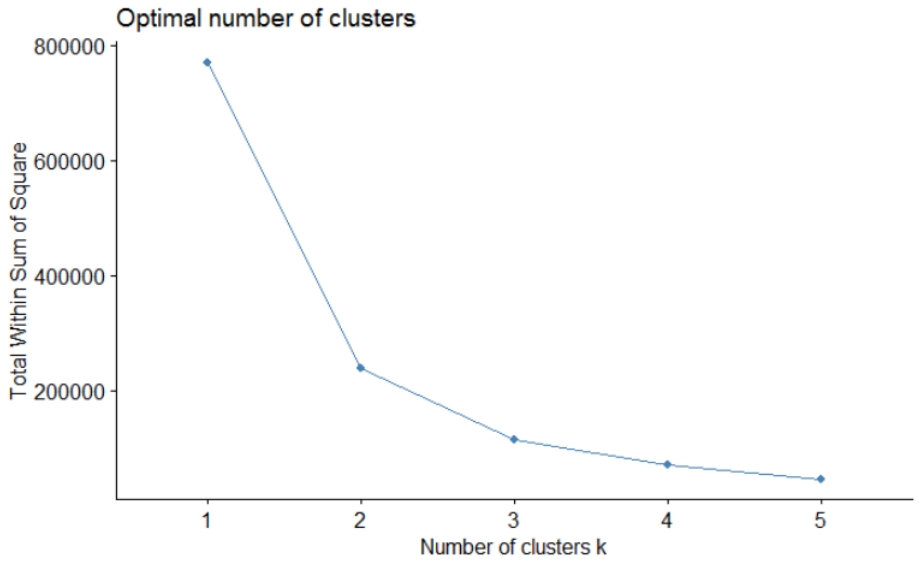

Figura 2.4: Representación gráfica del número óptimo de clusters

Como se puede observar en la figura 2.4, parece que el codo se da con 3 componentes o, como máximo, con 4. De todas maneras, la representación *PCoA* se realizará con 3 componentes, como se muestra a continuación.

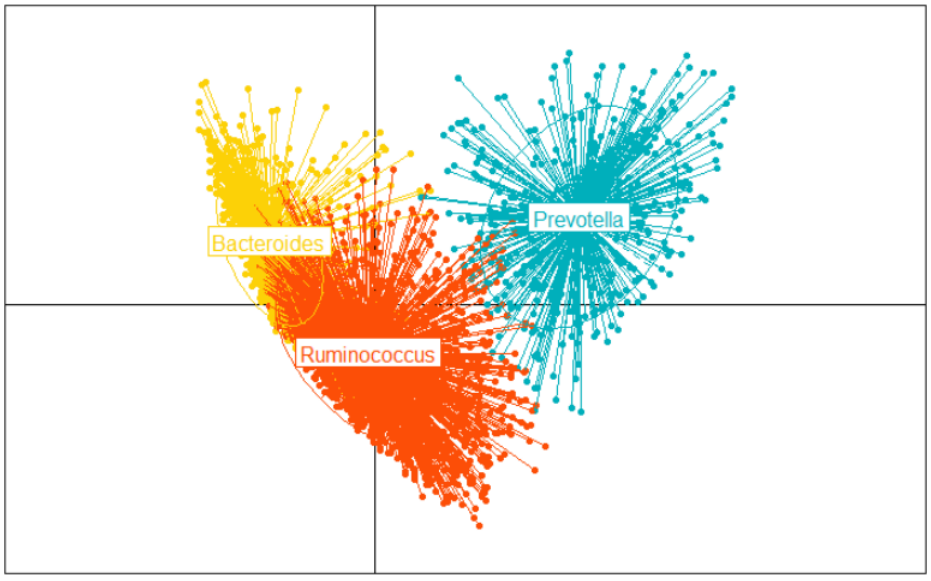

Figura 2.5: Representación gráfica PCoA de la población taxonómica con los filtros sobre los taxones *Pseudomonas* y *(unkn.) Enterobacteiaceae(f)* aplicados

El procedimiento de agrupación ha sido aleatorio y las etiquetas corresponden a los enterotipos *Bacteroides (enterotipo tipo 1)*, *Prevotella (enterotipo tipo 2)* y *Ruminococcus (enterotipo tipo 3)* [17].

Los datos que hacen referencia a las **características** de los pacientes, se encuentran en forma de matriz dónde las filas son los identificadores de los pacientes y las columnas cada una de estas características.

	Sex	Age	Height	Weight	Vegetable.freq	Seafood.freq	Fruit.freq	Red.meat.freq
ERR1596984	female	48	175	74	Daily	Occasionally	Daily	Occasionally
ERR1596990	female	62	162	76	Daily	Regularly	Regularly	Never
ERR1596999	female	59	175	57	Daily	Occasionally	Daily	Rarely
ERR1597001	male	36	182	75	Occasionally	Occasionally	Rarely	Occasionally
ERR1597002	male	37	174	81	Occasionally	Occasionally	Occasionally	Occasionally
ERR1597006	female	55	167	55	Daily	Rarely	Regularly	Rarely
ERR1597011	male	37	176	82	Regularly	Occasionally	Rarely	Occasionally

Figura 2.6: Representación de la matriz de características

Los datos que hacen referencia a las **enfermedades mentales** de los pacientes, se encuentran en forma de matriz dónde las filas son los identificadores de los pacientes y las columnas cada de los trastornos mentales analizados en este estudio.

	Anorexia.nervosa	Bipolar.disorder	Bulimia.nervosa	Depression	Ptsd	Schizophrenia	Substance.abuse
ERR1596984	False	False	False	False	False	False	False
ERR1596990	False	False	False	False	False	False	False
ERR1596999	False	False	False	True	False	False	False
ERR1597001	False	False	False	False	False	False	False
ERR1597002	False	False	False	False	False	False	False
ERR1597006	False	False	False	False	False	False	False
ERR1597011	False	False	False	False	False	False	False

Figura 2.7: Representación de la matriz de enfermedades mentales.

Como se puede observar en la siguiente figura, la distribución de las enfermedades mentales no es equitativa entre todas ellas.

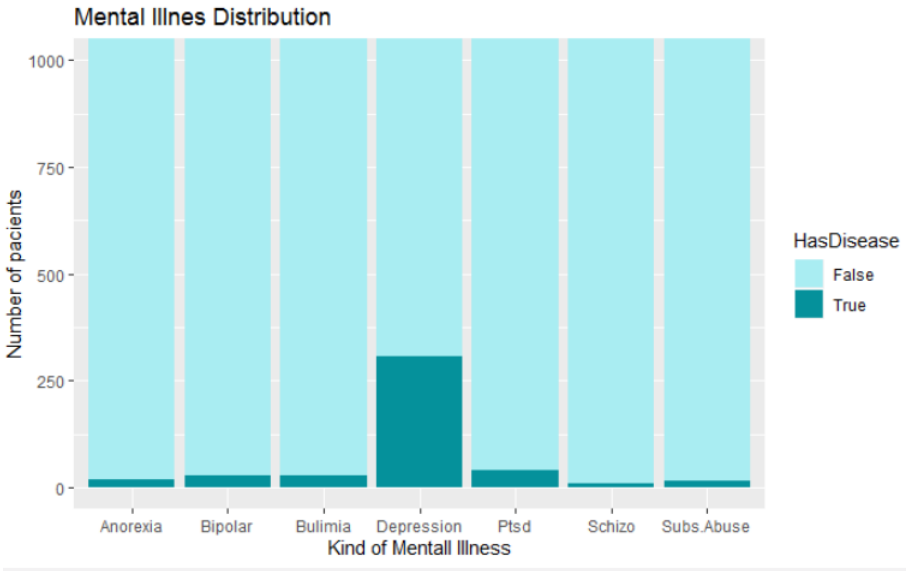

Figura 2.8: Distribución de las distintas enfermedades mentales

2.3. Evaluación de los algoritmos de clasificación

2.3.1. Planteamiento

Una vez finalizada la fase de análisis de los datos, queda determinar qué algoritmo o algoritmos serán utilizados en el proceso de clasificación dentro de la aplicación. Los candidatos son *Boost Decision Trees*, *Support Vector Machine* y *Random Forest*. El criterio para decidir cuáles serán incluídos en la aplicación final, dependerá en gran medida de cómo estos algoritmos son capaces de adaptarse a las necesidades de los datos. Recordemos que no todas las enfermedades mentales están distribuidas por igual dentro del juego de datos, lo que puede afectar muy significativamente en la eficiencia y el rendimiento de los clasificadores.

2.3.2. Re-sampling

Antes de entrar en materia con el funcionamiento y los resultados obtenidos de cada uno de los algoritmos, es importante explicar en qué consiste este apartado. Para realizar un correcto entrenamiento, es necesario que los datos dispongan de un buen balanceo o equilibrio entre la clase positiva, que es la que hace referencia a la existencia de la enfermedad mental, y la clase negativa que indicará que no se tiene tal enfermedad mental.

Se sabe que existe una gran descompensación entre estas dos clases. Para poner en práctica el *Re-sampling*, existen algunas técnicas que han sido puestas en marcha, como son *Under sampling* y *Over sampling* [10]. Con el fin de mitigar este hecho, se han aplicado ambas técnicas.

Under sampling

En este caso, se selecciona solo algunos de los datos de la clase mayoritaria, manteniendo la distribución de probabilidad de la clase. Por lo tanto, se realiza un sesgado para igualar, en función de un porcentaje, los resultados positivos con los negativos.

Cabe indicar que este tipo de balanceo tiene ventajas, como la ayuda al algoritmo de clasificación a distinguir mejor la clase minoritaria de la clase mayoritaria eliminando las

observaciones ruidosas. Pero también tiene desventajas en la eficiencia, ya que aquellos datos eliminados podrían tener información relevante sobre la clase mayoritaria.

Over sampling

Se trata de crear copias de la clase minoritaria para tener la misma cantidad de ejemplos que la clase mayoritaria, manteniendo la distribución de la clase minoritaria. Por lo tanto, se duplican, en función de un porcentaje, los resultados negativos con los positivos.

Del mismo modo que en el caso anterior, este tipo de balanceo tambien tiene ventajas, como la reducción del sesgo introducido por el desequilibrio de clase. Pero también tiene la desventaja de haber un riesgo de tener únicamente un solo ejemplo minoritario para la observación perteneciente a la clase minoritaria con una baja distribución.

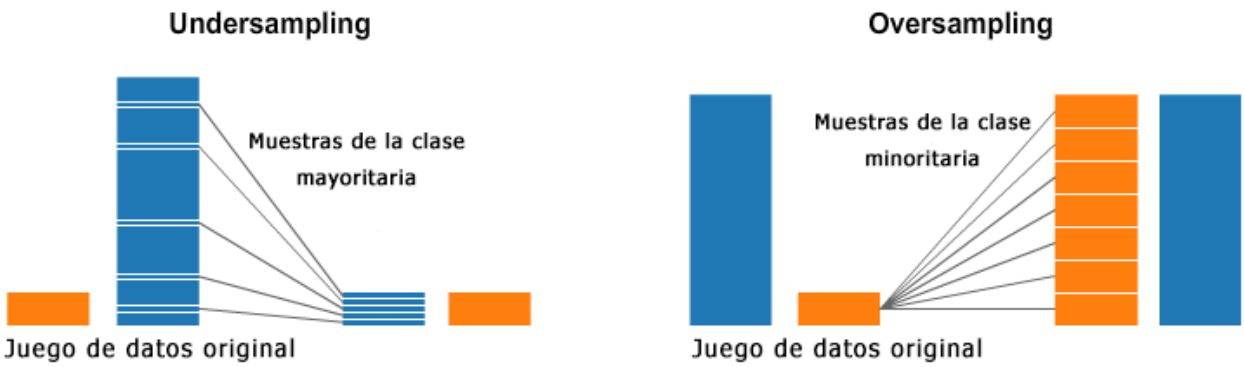

Figura 2.9: Representación gráfica de Re-sampling

2.3.3. Boost Decision Trees

Los *Árboles de Decisión* o *Boost Decision Trees* son modelos de predicción que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva para la resolución de un problema. Estos algoritmos aportan una visión gráfica de la toma de decisión, especifican las variables que son evaluadas, qué acciones deben ser tomadas y el orden en la que la toma de decisiones será efectiva.

Cada vez que se ejecuta un árbol de decisión, solo un camino será seguido dependiendo del valor actual de la variable evaluada.

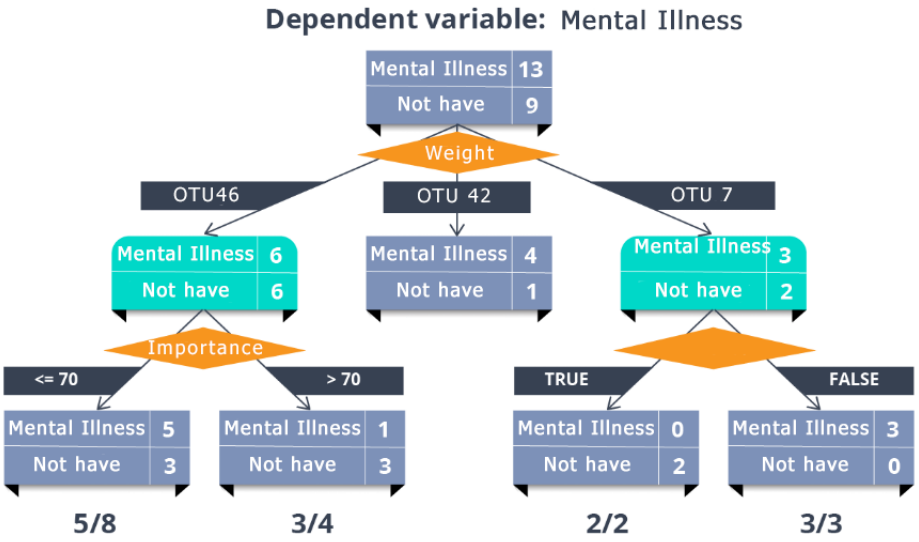

Figura 2.10: El árbol de clasificación lo compone una serie de nodos internos y externos, así como arcos que los unen

A los nodos externos se les conoce como hojas del árbol y se marcan con una clase o una distribución de probabilidad sobre las clases. Aunque los resultados que se obtienen dependen de la calidad de los ejemplos introducidos, sin lugar a dudas, este mecanismo de descubrimiento de patrones se ha convertido, en los últimos años, en una de las fuentes más fiables de predicción y su uso se ha extendido velozmente. [1]

Para evitar el overfitting, debido a que este tipo de árboles tienden a reducir rápidamente el error de entrenamiento ajustándose bien a las observaciones utilizadas para la fase de entrenamiento, es importante controlar bien el tamaño del árbol. Para el proceso de entrenamiento del modelo se ha limitado a 50 iteraciones.

En la siguiente tabla se enumeran las fortalezas y debilidades generales de los modelos basados en este algoritmo.

Fortalezas	Debilidades
Clasificador de uso múltiple con buenos resultados en la mayoría de los problemas.	Frecuentemente son sesgados hacia las divisiones, si tienen una gran cantidad de niveles.
Proceso de aprendizaje altamente automático que puede permitir valores numéricos o nominales, como valores faltantes.	Es fácil que se produzca *overfitting* o *underfitting*.
Excluye características sin importancia.	Puede tener problemas al modelar algunas relaciones debido a la confianza en las divisiones del eje-paralelo.
Se puede utilizar tanto en pequeños como en grandes conjuntos de datos.	Pequeños cambios en los datos del entrenamiento, pueden resultar ser grandes cambios en la toma de decisiones.
La interpretación de los datos no requierede experiencia matemática por parte del usuario.	Grandes árboles pueden dificultar la interpretación y la toma de decisión.
Más eficiente que otros modelos complejos.	

Cuadro 2.5: Fortalezas y debilidades del algoritmo Boost Decision Tree

2.3.4. Boost Decision Trees Resultados

A continuación, se muestran los diferentes resultados obtenidos durante el proceso de validación de este algoritmo, con los distintos tipos de técnicas de balanceo de los datos. Las enfermedades mentales mostradas en estos resultados son la *Depression* y la *Schizophrenia*, ya que la primera es, de las siete, la que más resultados positivos tiene y la segunda es la que menos casos positivos tiene. De esta manera se muestra la importancia de tener los datos balancelados.

No sampling

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   774  104
##               True      0    4
##
##                Accuracy : 0.8821
##                  95% CI : (0.8589, 0.9026)
##     No Information Rate : 0.8776
##     P-Value [Acc > NIR] : 0.3639
##
##                   Kappa : 0.0632
```

Figura 2.11: No sampling Depression

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   876    6
##           True      0    0
##
##                Accuracy : 0.9932
##                  95% CI : (0.9853, 0.9975)
##     No Information Rate : 0.9932
##     P-Value [Acc > NIR] : 0.60630
##
##                   Kappa : 0
```

Figura 2.12: No sampling Schizophrenia

Tanto en la figura 2.11 como en la figura 2.12 se observan las dificultades que ha tenido el clasificador para ajustar el modelo. En el caso de la *Depression* ha detectado muchos falsos positivos o «Error de tipo I »y el valor del coeficiente *kappa* tiende a cero, lo que implica un sesgo muy grande y por lo tanto, se puede decir que estos resultados son fruto del ”azar” más que de una buena clasificación. En el caso de la *Schizophrenia* los resultados fueron peores, ya que fue incapaz de detectar un resultado positivo y el valor de *kappa* es cero.

Under sampling

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   130   68
##               True     27   32
##
##                Accuracy : 0.6304
##                  95% CI : (0.5682, 0.6895)
##     No Information Rate : 0.6109
##     P-Value [Acc > NIR] : 0.2835
##
##                   Kappa : 0.1599
```

Figura 2.13: Under sampling Depression

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   174    4
##           True      0    1
##
##                Accuracy : 0.9777
##                  95% CI : (0.9438, 0.9939)
##     No Information Rate : 0.9721
##     P-Value [Acc > NIR] : 0.4380
##
##                   Kappa : 0.3271
```

Figura 2.14: Under sampling Schizophrenia

Al aplicar Under sampling, los resultados han sido mucho mejores que en el caso anterior, pero sin llegar a ser satisfactorios. En el caso de la *Depression*, además de los

falsos positivos de antes, también han aparecido falsos negativos o «Error de tipo II »y el valor del coeficiente *kappa* apenas ha mejorado significativamente. Sin embargo, donde si ha habido una mejora sustancial, ha sido en el caso de la *Schizophrenia* donde el «Error de tipo I »se mantiene, pero ha sido capaz de detectar un solo caso positivo. Aquí el valor de *kappa* ha mejorado, pero todavía con resultados que aportan poca seguridad de que éstos no sean fruto del ”azar”.

Over sampling

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   775   74
##               True      3  110
##
##                Accuracy : 0.92
##                  95% CI : (0.901, 0.9363)
##     No Information Rate : 0.8087
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6966
```

Figura 2.15: Over sampling Depression

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   874    3
##           True      4    3
##
##                Accuracy : 0.9921
##                  95% CI : (0.9838, 0.9968)
##     No Information Rate : 0.9932
##     P-Value [Acc > NIR] : 0.7444
##
##                   Kappa : 0.4576
```

Figura 2.16: Over sampling Schizophrenia

Al aplicar *Over sampling*, los resultados han mejorado lo suficiente como para tenerlos en cuenta. Los valores de *kappa* en ambos casos toman más relevancia y a pesar de que cuando se tienen pocos resultados positivos en los datos, como ocurre con la *Schizophrenia*, los resultados son más equilibrados aunque sin ser del todo satisfactorio. En el caso de la *Depression*, son algo mejores, también a consecuencia de que existen más positivos en el juego de datos, aunque el «Error de tipo I »sigue siendo bastante elevado.

La valoración sobre este algoritmo de clasificación no es muy positiva, porque incluso después de balancear los datos de diferentes maneras, el modelo no ha sido ajustado correctamente y deja muchas lagunas en la veracidad de los resultados. De todas maneras, no ha quedado descartado y se incluye en la aplicación de software.

2.3.5. Support Vector Machine

Support Vector Machines son un conjunto de algoritmos de aprendizaje supervisado relacionados con problemas de clasificación y regresión. Se pueden adaptar a casi cualquier tipo de tarea de aprendizaje, incluido la predicción numérica. La clave del éxito del algoritmo tiene que ver con el reconocimiento de patrones.

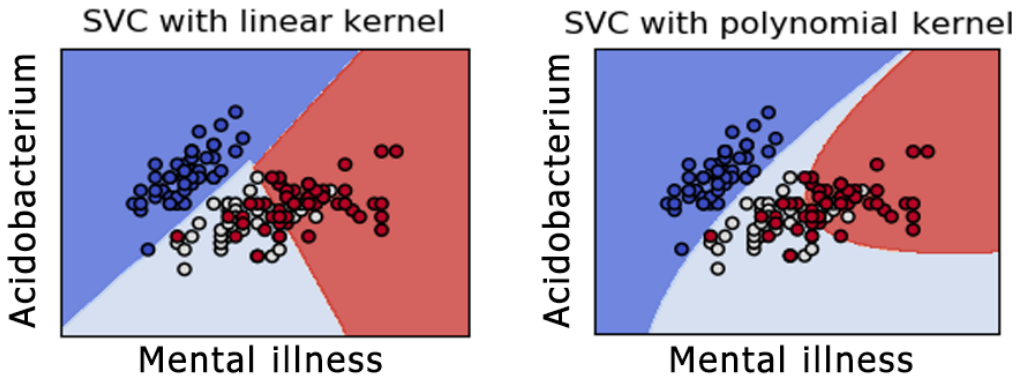

Figura 2.17: Representación gráfica de las funciones kernel lineal y polinomial

Sobresale en aplicaciones como:

- Clasificación de datos de expresión génica de microarrays, en el campo de Bioinformática, para identificar cáncer u otras enfermedades genéticas.
- Categorización de textos, para identificar idiomas utilizados en un documento o la clasificación de documentos por materias.
- Detección de eventos raros, pero importantes, como los fallos de motor de combustión, violaciones de seguridad o terremotos.

En la siguiente tabla se enumeran las fortalezas y debilidades generales de los modelos basados en este algoritmo.

Fortalezas	**Debilidades**
Puede ser utilizado para la clasificación o para problemas de predicción numérica.	Encontrar un buen modelo requiere probar varias combinaciones de kernel y parámetros del modelo.
No está demasiado influenciado por datos con ruido yno es propenso a la sobrealimentación.	Puede ser lento para la fase deentrenamiento, particularmente si elconjunto de datos de entrada tiene un gran número de características.
Puede ser más fácil de usar que la *Red Neuronal* debido al respaldo de los algoritmos *SVM*.	Los resultados se obtienen de una compleja caja negra casi imposible de interpretar.
Gana popularidad debido a su alta precisión y a lasvictorias obtenidas en competiciones de *Data Mining*.	

Cuadro 2.6: Fortalezas y debilidades del algoritmo Support Vector Machine

Este algoritmo permite utilizar diferente funciones kernel para el análisis de patrones. La tarea general del análisis de patrones es encontrar y estudiar tipos generales de relaciones entre grupos y/o clasificaciones en el conjuntos de datos. Las funciones utilizadas han sido:

* Función **kernel lineal** es la más simple de las funciones ya que se basa, principalmente, en la separación de una linea recta o un hiperplano de dimensión N.

* Función **kernel polinomial** es equivalente a una regresión polinómica. Es mucho más preciso que el kernel lineal y sobre todo es resistente al overfitting, a pesar de que penaliza con el elevado coste computacional. [16]

2.3.6. Support Vector Machine Resultados

A continuación se muestras los resultados obtenidos, para las enfermedades mentales *Depression* y *Schizophrenia*, por cada uno de los clasificadores, utilizando técnicas de *Re-sampling* y sin utilizarlas.

No sampling

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   736   86
##               True     42   18
##
##                Accuracy : 0.8549
##                  95% CI : (0.8299, 0.8775)
##     No Information Rate : 0.8821
##     P-Value [Acc > NIR] : 0.9936414
##
##                   Kappa : 0.1458
```

Figura 2.18: No sampling Depression kernel lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   879    1
##           True      1    1
##
##                Accuracy : 0.9977
##                  95% CI : (0.9918, 0.9997)
##     No Information Rate : 0.9977
##     P-Value [Acc > NIR] : 0.6767
##
##                   Kappa : 0.4989
```

Figura 2.19: No sampling Schizophrenia kernel lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   736   99
##               True     35   12
##
##                Accuracy : 0.8481
##                  95% CI : (0.8227, 0.8711)
##     No Information Rate : 0.8741
##     P-Value [Acc > NIR] : 0.9901
##
##                   Kappa : 0.0833
```

Figura 2.20: No sampling Depression kernel polinomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   877    4
##           True      0    1
##
##                Accuracy : 0.9955
##                  95% CI : (0.9884, 0.9988)
##     No Information Rate : 0.9943
##     P-Value [Acc > NIR] : 0.4400
##
##                   Kappa : 0.3321
```

Figura 2.21: No sampling Schizophrenia kernel polinomial

En comparación con el algoritmo BDT, SVM con la función lineal mejora sustancialmente los resultados a pesar de no haber aplicado *Re-sampling*. Como se observa en la figura 2.18, teniendo en cuenta un valor *kappa* bastante bajo, $0{,}1458$ se observa una

buena predisposición a detectar positivos, aunque si que hay que decir que la tasa de «Error de tipo I y II »también es elevada.

Pero por el contrario, en la figura 2.19, la mejora es significativa llegando a obtener un valor *kappa* de $0{,}4989$ y una tasa de «Error de tipo I y II »muy baja.

Under sampling

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   110   47
##               True     56   44
##
##                Accuracy : 0.5992
##                  95% CI : (0.5365, 0.6596)
##     No Information Rate : 0.6459
##     P-Value [Acc > NIR] : 0.9475
##
##                   Kappa : 0.143
```

Figura 2.22: Under sampling Depression kernel lineal.

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   178    1
##           True      0    0
##
##                Accuracy : 0.9944
##                  95% CI : (0.9693, 0.9999)
##     No Information Rate : 0.9944
##     P-Value [Acc > NIR] : 0.7358
##
##                   Kappa : 0
```

Figura 2.23: Under sampling Schizophrenia kernel lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   104   66
##               True     42   45
##
##                Accuracy : 0.5798
##                  95% CI : (0.5168, 0.6408)
##     No Information Rate : 0.5681
##     P-Value [Acc > NIR] : 0.37746
##
##                   Kappa : 0.1209
```

Figura 2.24: Under sampling Depression kernel polinomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   176    1
##           True      1    1
##
##                Accuracy : 0.9888
##                  95% CI : (0.9602, 0.9986)
##     No Information Rate : 0.9888
##     P-Value [Acc > NIR] : 0.6767
##
##                   Kappa : 0.4944
```

Figura 2.25: Under sampling Schizophrenia kernel polinomial

La aplicación del *Under sampling* ha empeorado los resultados en ambos casos, simplemente, porque el problema está en el escaso número de positivos. A pesar de ello,

Aplicando *Under sampling* tampoco se puede decir que los resultados obtenidos hayan mejorado. Los valores para *kappa* siguen siendo no significativos y para la *Depres-*

sion ha aumentado los «Errores de tipo I y II»como los aciertos. Asi pues, este ajuste tampoco se puede tomar en cuenta.

Over sampling

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   731   83
##               True     63   85
##
##                Accuracy : 0.8482
##                  95% CI : (0.824, 0.8703)
##     No Information Rate : 0.8254
##     P-Value [Acc > NIR] : 0.03223
##
##                   Kappa : 0.4476
```

Figura 2.26: Over sampling Depression kernel lineal.

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   877    2
##           True      0    5
##
##                Accuracy : 0.9977
##                  95% CI : (0.9919, 0.9997)
##     No Information Rate : 0.9921
##     P-Value [Acc > NIR] : 0.02919
##
##                   Kappa : 0.8322
```

Figura 2.27: Over sampling Schizophrenia kernel lineal.

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   705   79
##               True     76  102
##
##                Accuracy : 0.8389
##                  95% CI : (0.8141, 0.8616)
##     No Information Rate : 0.8119
##     P-Value [Acc > NIR] : 0.01637
##
##                   Kappa : 0.4692
```

Figura 2.28: Over sampling Depression kernel polinomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   877    3
##           True      0    4
##
##                Accuracy : 0.9966
##                  95% CI : (0.9901, 0.9993)
##     No Information Rate : 0.9921
##     P-Value [Acc > NIR] : 0.08094
##
##                   Kappa : 0.7257
```

Figura 2.29: Over sampling Schizophrenia kernel polinomial

De nuevo, se observa una mejora significativa en los resultados, sobre todo para la *Schizophrenia*, obteniedo un valor de *kappa* de $0{,}832$ con una precisión del $99{,}7\,\%$ y de $0{,}7257$ y una precisión del $99{,}6\,\%$ para las funciones lineal y polinomial respectivamente. Más dificultades ha tenido con la *Depression* con un valor de *kappa* de $0{,}447$ y una precisión del $84{,}8\,\%$ y de $0{,}469$ y una precisión del $83{,}8\,\%$ para las funciones lineal y polinomial respectivamente. Se concluye que ambas funciones son válidas para el ajuste del modelo y serán incluídas en la aplicación de software.

2.3.7. Random Forest

El método *Random Forest* está basado, principalmente, en los *Árboles de Decisión*. Se crea un gran número de árboles mediante el muestreo de individuos y variables en el conjunto de datos. Una diferencia clave con los Árboles de Decisión es que cada nodo está dividido por el mejor de un subconjunto aleatorio de variables, en lugar de la mejor de todas las variables. Cada individuo está clasificado por cada árbol y el resultado más común se utiliza como la clasificación final. Además se puede utilizar para realizar tareas de clasificación y de regresión. En muchos problemas el rendimiento del algoritmo *Random Forest* es muy similar a la del *boosting*, anteriormente probado, y es más simple de entrenar y ajustar. Esta característica lo convierte en popular y ampliamente utilizado.

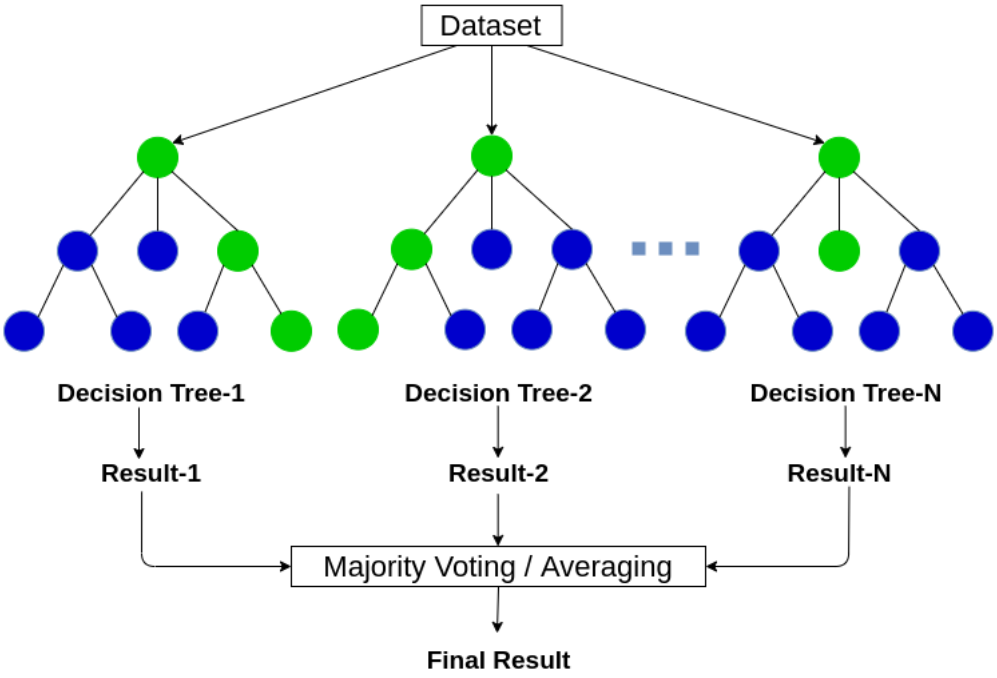

Figura 2.30: Representación gráfica del algoritmo Random Forest

La idea esencial es promediar muchos modelos ruidosos pero aproximadamente imparciales, y por tanto reducir la variación. Los árboles son los candidatos ideales, dado que ellos pueden registrar estructuras de interacción compleja en los datos, y si crecen con la suficiente profundidad, tienen relativamente baja parcialidad. [9]

En la siguiente tabla se enumeran las fortalezas y debilidades generales de los modelos basados en este algoritmo.

Fortalezas	Debilidades
Un modelo multipropósito que realiza bien la mayoría de los problemas.	A diferencia del *Árbol de Decisión*, el modelo no es fácil de interpretar.
Puede manejar datos ruidosos o faltantes así como variables categóricas o contínuas.	Es posible que se necesite algún trabajo extra para sintonizar el modelo con los datos.
Selecciona sólo las características más importantes.	
Se puede utilizar en datos con un gran número de características o ejemplos.	

Cuadro 2.7: Fortalezas y debilidades del algoritmo Random Forest

2.3.8. Random Forest Resultados

A continuación se muestras los resultados obtenidos, para las enfermedades mentales *Depression* y *Schizophrenia*, por cada uno de los clasificadores, utilizando técnicas de *Re-sampling* y sin utilizarlas.

No sampling

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   783   97
##               True      1    1
##
##                Accuracy : 0.8889
##                  95% CI : (0.8663, 0.9089)
##     No Information Rate : 0.8889
##     P-Value [Acc > NIR] : 0.5269
##
##                   Kappa : 0.0156
```

Figura 2.31: No sampling Depression.

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   878    4
##           True      0    0
##
##                Accuracy : 0.9955
##                  95% CI : (0.9884, 0.9988)
##     No Information Rate : 0.9955
##     P-Value [Acc > NIR] : 0.6288
##
##                   Kappa : 0
```

Figura 2.32: No sampling Schizophrenia.

Como era de esperar y tal como ha ocurrido con *BDT*, los resultados sin realizar *Resampling* han sido muy malos. Los valores para *kappa* para ambas enfermedades son prácticamente 0 y el «Error de tipo I »se ha disparado. Por lo tanto, este ajuste carece de importancia.

Under sampling

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   155   76
##               True     10   16
##
##                Accuracy : 0.6654
##                  95% CI : (0.6041, 0.7228)
##     No Information Rate : 0.642
##     P-Value [Acc > NIR] : 0.2381
##
##                   Kappa : 0.1347
```

Figura 2.33: Under sampling Depression.

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   176    3
##           True      0    0
##
##                Accuracy : 0.9832
##                  95% CI : (0.9518, 0.9965)
##     No Information Rate : 0.9832
##     P-Value [Acc > NIR] : 0.6472
##
##                   Kappa : 0
```

Figura 2.34: Under sampling Schizophrenia.

Aplicando *Under sampling* tampoco se puede decir que los resultados obtenidos hayan mejorado. Los valores para *kappa* siguen siendo no significativos y para la *Depresion* ha aumentado los «Errores de tipo I y II»como los aciertos. Asi pues, este ajuste tampoco se puede tomar en cuenta.

Over sampling

```
## Confusion Matrix and Statistics
##
##
## dat.pred.depression False True
##               False   778   62
##               True      0  122
##
##                Accuracy : 0.9356
##                  95% CI : (0.9181, 0.9502)
##     No Information Rate : 0.8087
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7609
```

Figura 2.35: Over sampling Depression.

```
## Confusion Matrix and Statistics
##
##
## dat.pred.schizo False True
##           False   875    5
##           True      0    4
##
##                Accuracy : 0.9943
##                  95% CI : (0.9869, 0.9982)
##     No Information Rate : 0.9898
##     P-Value [Acc > NIR] : 0.11445
##
##                   Kappa : 0.613
```

Figura 2.36: Over sampling Schizophrenia.

Al aplicar el balanceo de los datos basado en *Over sampling* se nota una considerable mejoría en los resultados. Los valores de *kappa* son signifitcativos, siendo de $0{,}7699$ para la *Depression* y de $0{,}613$ para la *Schizophrenia*. A pesar de que el «Error de tipo I »se sigue manteniendo, los resultados positivos doblan a éstos en el primer caso y los iguala en el segundo. Por este motivo, se tendrá en cuenta esta configuración para el ajuste del modelo.

2.4. Software

Además del estudio sobre la relación entre el microbioma intestinal y ciertas enfermedades mentales, documentado en apartados anteriores, se ha querido ir más allá desarrollando una herramienta que permita al especialista obtener predicciones sobre pacientes, y además realizar la investigación sobre qué microorganismos tienen más relevancia cuando se detecta una enfermedad mental. Esta herramienta permite a usuarios con poco conocimiento de análisis estadístico y programación, el poder realizar predicciones e investigaciones sobre este tema.

2.4.1. Tecnologías

Para desarrollar la aplicación de predicción de enfermedades mentales, llamada **MI-Predictor**, se ha necesitado la combinación de dos tecnologías.

En primer lugar **R**, con un enfoque más matemático, responsable de todos los cálculos numéricos y capaz de manejar grandes cantidades de datos, orientado al análisis estadístico, que es utilizado ampliamente en el ámbito de *la ciencia de datos*. [19]

Y en segundo lugar y no por ellos menos importante, **Python** que en este caso por su versatilidad, ha servido como contenedor de todas las funciones de R y para que el usuario final pueda interactúar con la aplicación [20]. Se requiere la librería *RPY2* para poder comunicar una tecnología con otra. [21]

Pensando en el futuro y en la mejora de la aplicación, ambas tecnologías son de código abierto[8] y disponen de amplias comunidades de desarrolladores que las mantienen en constante evolución.

2.4.2. Interfaz gráfica de usuario

Para facilitar al usuario el manejo de la aplicación, cada una de las funcionalidades han sido separadas por pestañas o *tabs* y cada una de éstas tiene una finalidad específica que se explica a continuación.

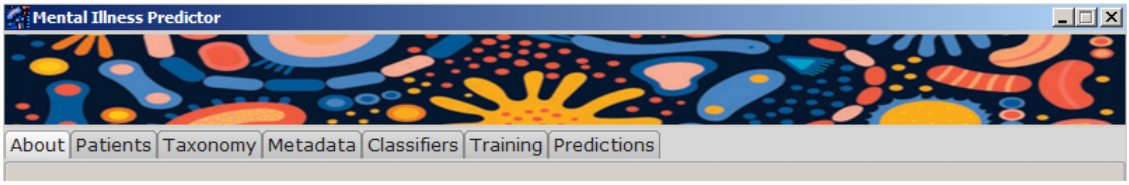

Figura 2.37: Tabs de cada una de las partes de la aplicación

Taxonomy

La pantalla *Taxonomy* se divide en dos apartados.

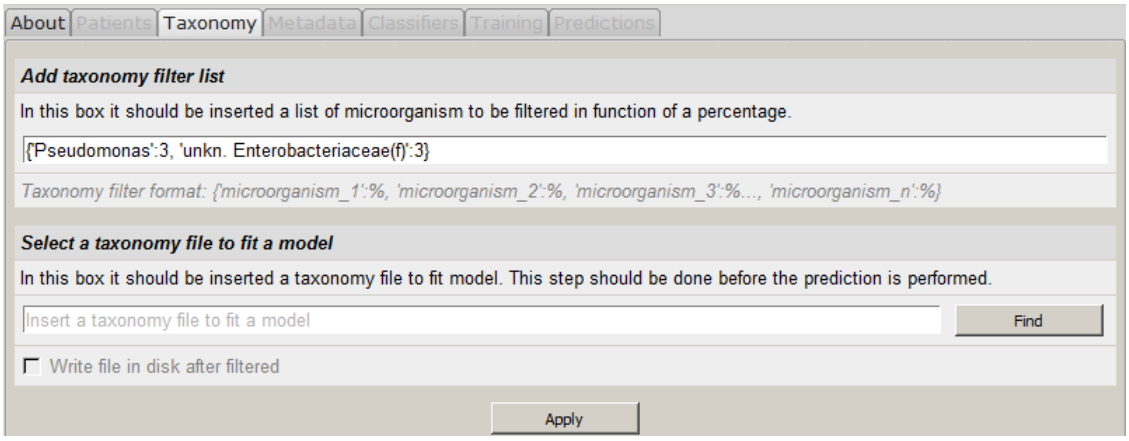

Figura 2.38: Tab par la carga y filtrado de la taxonomía

En el apartado **Add taxonomy filter list** se permite introducir una lista de duplas que deberán estar formadas por un nombre de taxón y un porcentaje, que será el que se tendrá en cuenta para el filtraje. Esta entrada permite introducir tantos elementos como sea necesario para el filtrado. Por defecto, están incluídos aquellos requeridos para este proyecto, pero como se ha comentado anteriormente, permite introducir muchos más.

En el apartado **Select a taxonomy file to fit a model** se debe seleccionar un fichero que contenga la taxonomía de una población, que servirá para el entrenamiento de los datos

Una vez seleccionado el filtro y el fichero de taxonomía se deberá pulsar el botón Apply para comenzar el proceso de filtrado.

Metadata

La pantalla *Metadata* se divide en dos apartados.

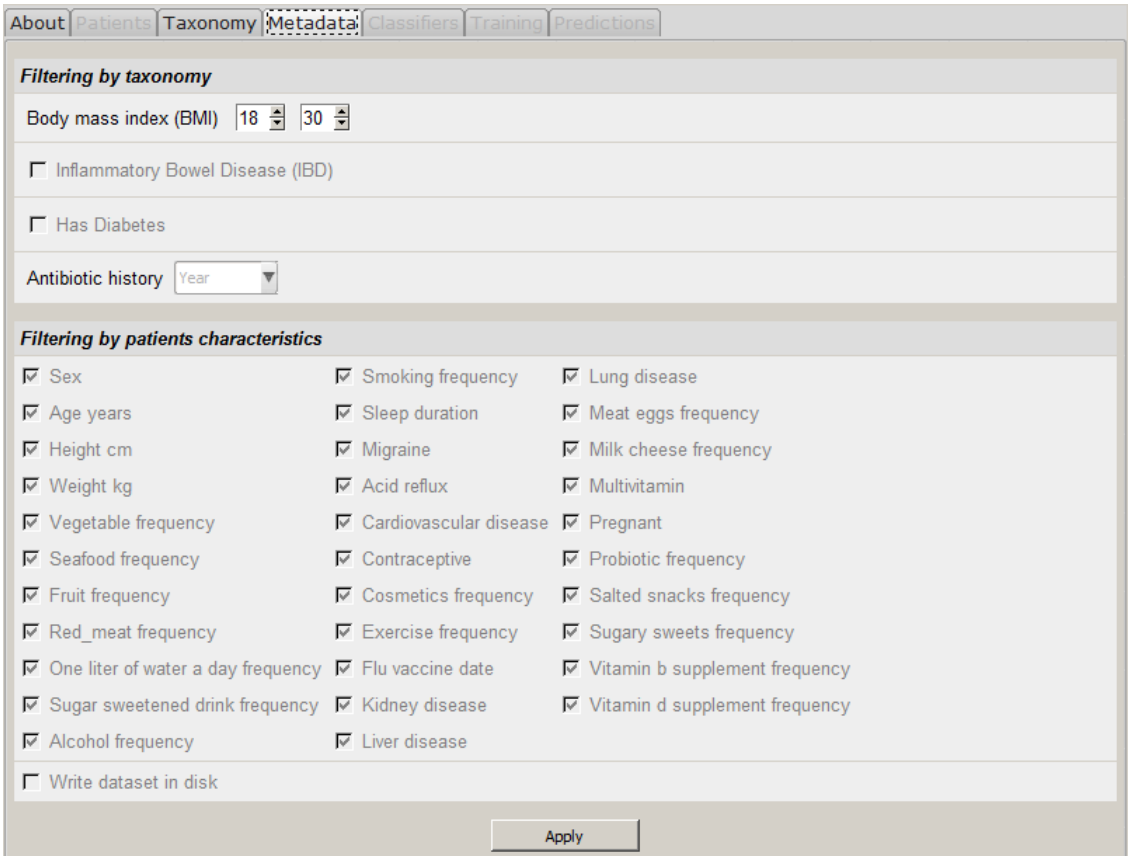

Figura 2.39: Tab para la carga y filtrado de los metadatos

En el apartado **Filtering by taxonomy** se permite configurar un segundo filtro basado en ciertas características del paciente. Se puede seleccionar el rango de *Body mass index* **BMI**, donde se debe indicar el valor mínimo y el máximo. El resto de configuraciones

están deshabilitadas y por defecto están aquellos valores requeridos para el proyecto.

En el apartado **Filtering by patients characteristics** aparecen otras características de los pacientes que se tendrán en cuenta para ajustar y entrenar el modelo. En este caso aparece deshabilitada la opción de marcarlas o no, y por defecto están todas seleccionadas.

Una vez elegidos los filtros, se deberá pulsar el botón Apply para comenzar el proceso de filtrado.

Classifiers

La pantalla *Classifiers* se divide en tres apartados.

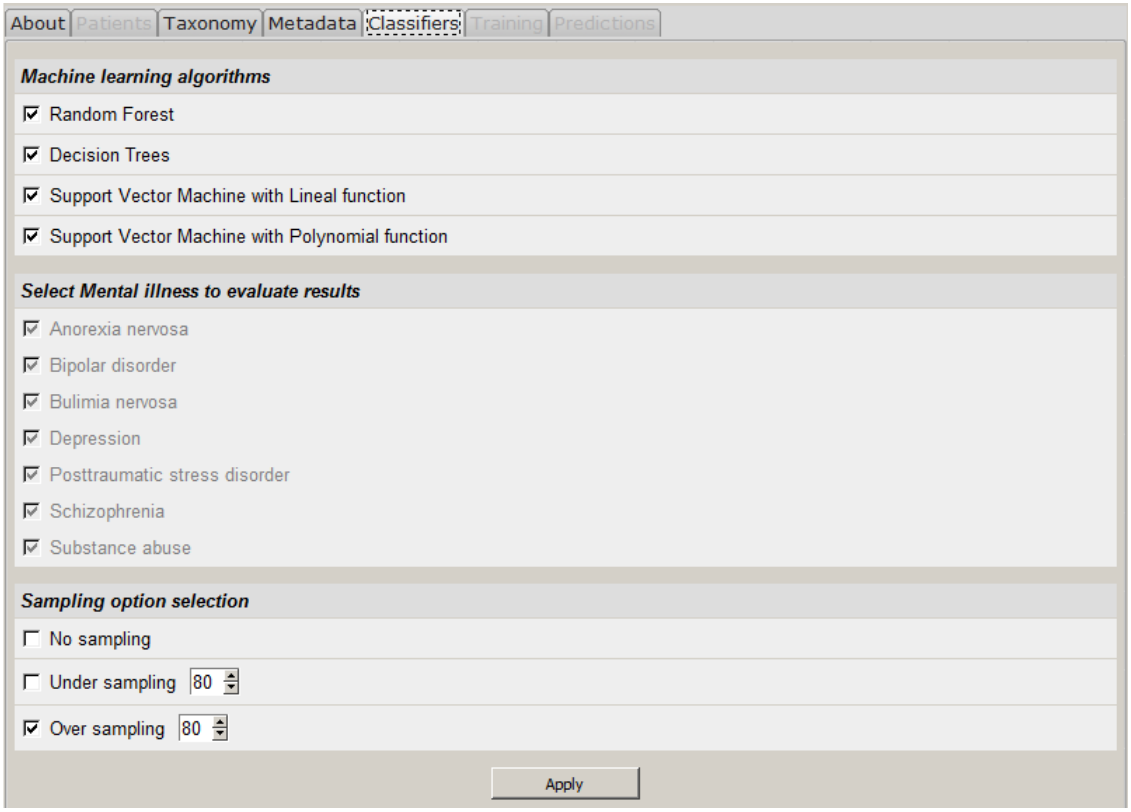

Figura 2.40: Tab para configurar los clasificadores

En el apartado **Machine learning algorithms** se puede seleccionar qué algoritmos se van a utilizar en el entrenamiento de los datos. Como mínimo se ha de seleccionar uno

de ellos. Todos los algoritmos seleccionados realizarán un entrenamiento de los datos y posteriormente se determinará cuál de ellos es el que tiene mejor rendimiento.

En el apartado **Select Mental illness to evaluate results** se permite seleccionar qué enfermedades mentales van a ser analizadas. Por defecto están todas seleccionadas.

En el apartado **Sampling option selection** se permite realizar ciertas transformaciones en los datos, antes de pasar por los algoritmos de clasificación. La opción *No sampling* no realiza ninguna transformación. La opción *Under sampling* permite eliminar un número de resultados negativos, en función de un porcentaje que también es ajustable. Por último, La opción *Over sampling* permite duplicar un número de resultados positivos, en función de un porcentaje que también es ajustable. Es importante indicar que solo puede haber una de estas tres opciones seleccionada.

Training

La pantalla *Training* se divide en dos apartados, y permite observar visualmente el resultado de los entrenamientos de cada uno de los algoritmos seleccionados para cada una de las enfermedades mentales, que están separadas por *tabs*.

En el primer apartado se muestran los resultados de cada uno de los algoritmos elegidos en el paso anterior. Se puede ver la tabla de contingencia y otros resultados como:

- F-F el sistema determinó falso cuando era falso.
- F-T el sistema determinó falso cuando era cierto.
- T-F el sistema determinó cierto cuando era falso.
- T-T el sistema determinó cierto cuando era cierto.
- Accuracy es el porcentaje de clasificaciones correctas de todas las instancias.
- CI (intervalos de confianza) es un posible rango de valores o intervalo (a; b), en el que, con una determinada probabilidad, sus límites contendrán el valor del parámetro poblacional que se está buscando.

Figura 2.41: Tab para el entrenamiento del modelo

- *p_value* corresponde al nivel de significación más pequeño posible que puede escogerse, para el cual todavía se aceptaría la hipótesis alternativa con las observaciones actuales. Cualquier nivel de significación escogido inferior a alpha = 0.05 comporta aceptar H0.

- kappa es similiar a la precisión de la clasificación, excepto porque se normaliza al inicio del azar en su conjunto de datos.

- Sensitivity también se conoce como Tasa de Verdaderos Positivos. Es la proporción de casos positivos que fueron correctamente identificados por el algoritmo.

- Specificity es la Tasa de Verdaderos Negativos. Se trata de los casos negativos que el algoritmo ha clasificado correctamente.

En el apartado **Chart results** se pueden visualizar los resultados de manera gráfica. Este apartado está divido a su vez en seis partes, una para cada tipo de resultado.

Indicar que, cuando un algoritmo no ha sido seleccionado para realizar el entrenamiento, la fila aparece con un color más oscuro que la fila de los algoritmos seleccionados para tal cometido.

Patients

La pantalla *Patients* se divide en tres apartados y permite realizar todo el proceso de la predicción.

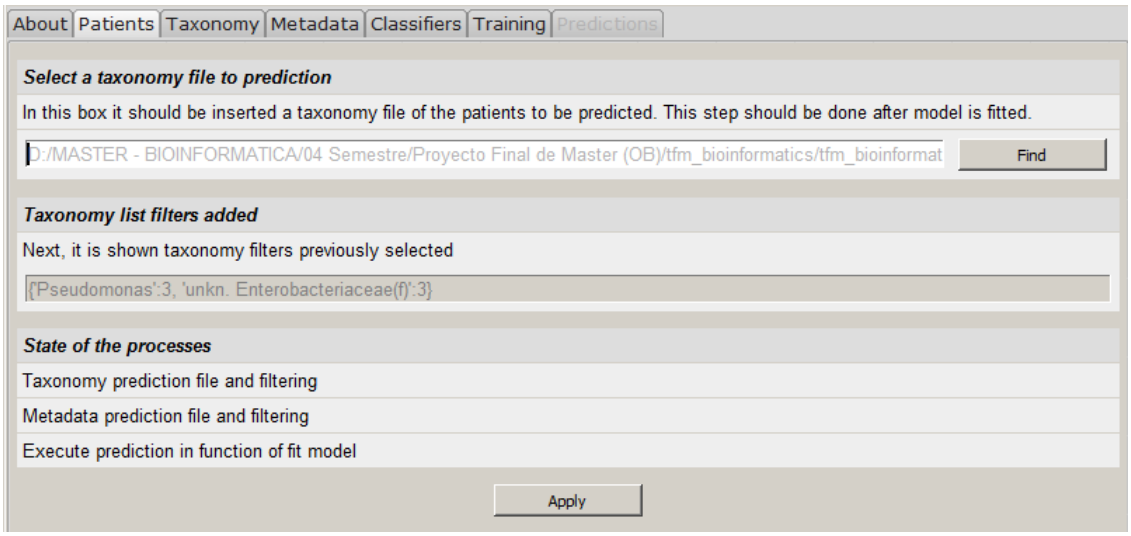

Figura 2.42: Tab para configurar los pacientes a predecir

En el apartado **Select a taxonomy file to prediction** se debe seleccionar un fichero que contenga la taxonomía de una población que servirá para realizar la predicción.

En el apartado **Taxonomy list filters added** se pueden observar la selección del filtro taxonómico tomada en la fase de entrenamiento con los taxones y sus respectivos porcentajes, pero ésta en este punto ya no es modificable.

En el apartado **State of processes** se puede observar los tres pasos que componen la predicción. Éstos son, el filtrado por taxonomía, el filtrado por metadatos y por último, la predicción en función del modelo ajustado anteriormente.

Cuando uno de estos pasos se inicia, se muestra un mensaje indicando *loading*. Una vez ha finalizado, el mensaje cambiará a *done*. De esta manera el usuario puede ir viendo el proceso de predicción. También hay que hacer referencia a que en la parte inferior de la pantalla irán apareciendo mensajes indicando cuál es el proceso actual que se está ejecutando.

Predictions

La pantalla *Predictions* se divide en tres apartados y permite visualizar los resultados de la predicción.

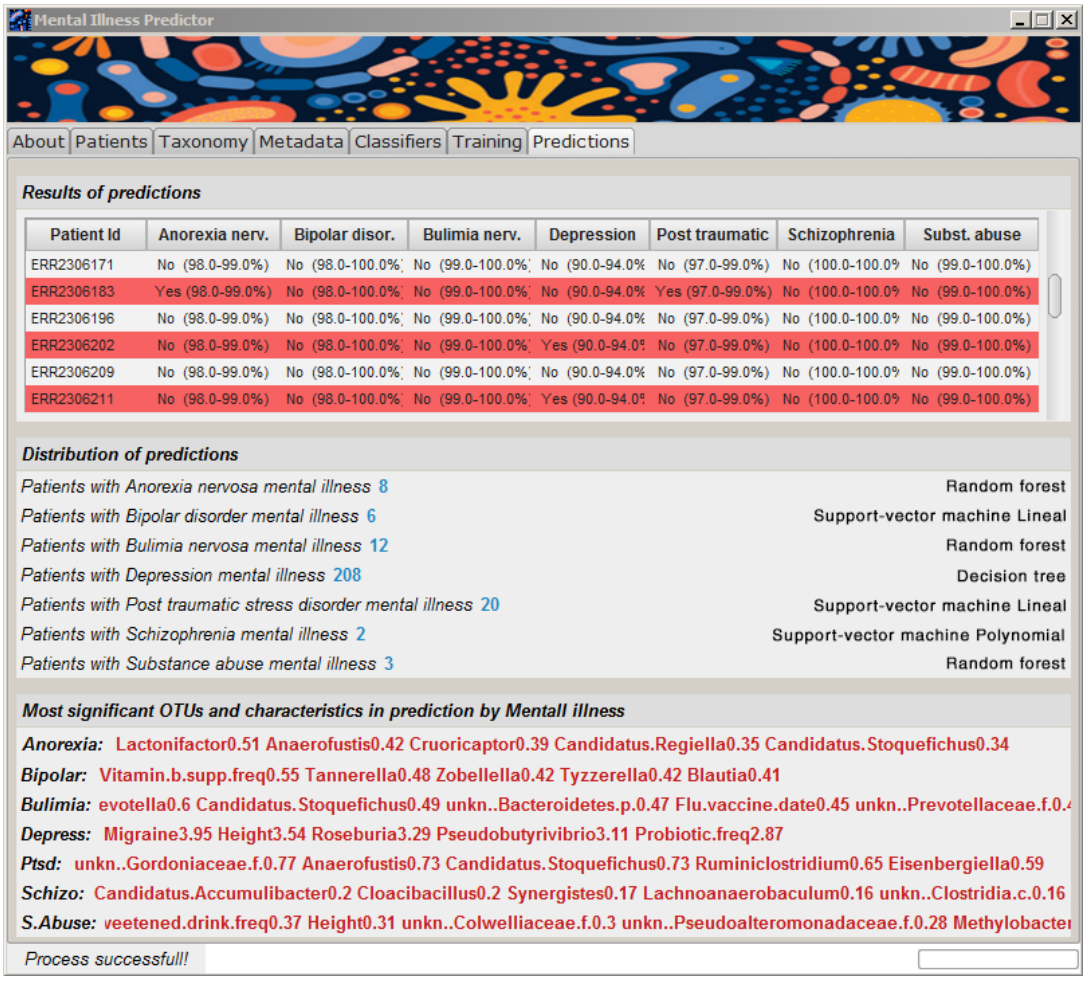

Figura 2.43: Tab con los resultados de las predicciones

En el apartado **Results of predictions** se muestran todos los resultados obtenidos sobre aquellos pacientes que han superado todos los filtros. Para cada paciente se muestra su identificador y si padece o no, con un cierto intervalo de confianza, cualquiera de las siete enfermedades mentales de las que se disponen datos.

En el apartado **Distribution of predictions** se muestra el número de pacientes que padecen una cierta enfermedad mental y con qué algoritmo se realizó dicha predicción. El algoritmo que realiza la predicción es el que mejores resultados obtuvo en la fase de entrenamiento y esta elección se hace automáticamente.

En el apartado **Most significant OTUs and characteristics in prediction by Mental illness** se muestran aquellas unidades taxonómicas y/o características más significativas para determinar si un paciente padece cierta enfermedad mental.

2.4.3. Resultados

En la gran mayoría de las pruebas que han sido realizadas, los resultados han sido óptimos con intervalos de confianza en el acierto de la predicción muy elevados.

Resultados en la predicción

Results of predictions

Patient Id	Anorexia nerv.	Bipolar disor.	Bulimia nerv.	Depression	Post traumatic	Schizophrenia	Subst. abuse
ERR2032863	Yes (98.0-100.0%)	No (97.0-99.0%)	No (99.0-100.0%	Yes (91.0-94.0%	Yes (98.0-99.0%)	No (99.0-100.0%	No (99.0-100.0%)
ERR2032867	No (98.0-100.0%)	Yes (97.0-99.0%)	No (99.0-100.0%	No (91.0-94.0%	No (98.0-99.0%)	No (99.0-100.0%	No (99.0-100.0%)
ERR2032869	No (98.0-100.0%)	No (97.0-99.0%)	No (99.0-100.0%	No (91.0-94.0%	No (98.0-99.0%)	No (99.0-100.0%	No (99.0-100.0%)
ERR2032893	No (98.0-100.0%)	No (97.0-99.0%)	No (99.0-100.0%	No (91.0-94.0%	No (98.0-99.0%)	No (99.0-100.0%	No (99.0-100.0%)
ERR2032896	No (98.0-100.0%)	No (97.0-99.0%)	No (99.0-100.0%	No (91.0-94.0%	No (98.0-99.0%)	No (99.0-100.0%	No (99.0-100.0%)
ERR2032897	No (98.0-100.0%)	No (97.0-99.0%)	No (99.0-100.0%	No (91.0-94.0%	No (98.0-99.0%)	No (99.0-100.0%	No (99.0-100.0%)

Figura 2.44: Resultados en la detección de Anorexia nervosa, Bipolar disorder, Depression y Post traumatic stress disorder

Como se observa en la figura 2.44, el paciente **ERR2032863** padece *Anorexia nervosa* con una certeza del $98 - 100\,\%$, además de *Depression* entre un $91 - 94\,\%$ y *Post traumatic stress disorder* entre un $98 - 99\,\%$. En el caso del paciente **ERR2032867**, padece *Bipolar disorder* entre un $97 - 99\,\%$.

Results of predictions

Patient Id	Anorexia nerv.	Bipolar disor.	Bulimia nerv.	Depression	Post traumatic	Schizophrenia	Subst. abuse
ERR2306171	No (98.0-99.0%)	No (98.0-100.0%	No (99.0-100.0%	No (90.0-94.0%	No (97.0-99.0%)	No (100.0-100.0%	No (99.0-100.0%)
ERR2306183	Yes (98.0-99.0%)	No (98.0-100.0%	No (99.0-100.0%	No (90.0-94.0%	Yes (97.0-99.0%)	No (100.0-100.0%	No (99.0-100.0%)
ERR2306196	No (98.0-99.0%)	No (98.0-100.0%	No (99.0-100.0%	No (90.0-94.0%	No (97.0-99.0%)	No (100.0-100.0%	No (99.0-100.0%)
ERR2306202	No (98.0-99.0%)	No (98.0-100.0%	No (99.0-100.0%	Yes (90.0-94.0%	No (97.0-99.0%)	No (100.0-100.0%	No (99.0-100.0%)
ERR2306209	No (98.0-99.0%)	No (98.0-100.0%	No (99.0-100.0%	No (90.0-94.0%	No (97.0-99.0%)	No (100.0-100.0%	No (99.0-100.0%)
ERR2306211	No (98.0-99.0%)	No (98.0-100.0%	No (99.0-100.0%	Yes (90.0-94.0%	No (97.0-99.0%)	No (100.0-100.0%	No (99.0-100.0%)

Figura 2.45: Resultados en la detección de Anorexia nervosa, Depression y Post traumatic stress disorder.

Otros resultados obtenidos se observan en la figura 2.45, donde el paciente **ERR2306183** padece *Anorexia nervosa* con una certeza del $98 - 99\,\%$ y *Post traumatic stress disorder* con un $97 - 99\,\%$ de acierto. Los pacientes **ERR2306202** y **ERR2306211** padecen *Depression* con una veracidad del $90 - 94\,\%$.

Results of predictions

Patient Id	Anorexia nerv.	Bipolar disor.	Bulimia nerv.	Depression	Post traumatic	Schizophrenia	Subst. abuse
ERR2092109	No (98.0-100.0%)	No (97.0-99.0%)	No (99.0-100.0%	Yes (91.0-94.0%	No (98.0-99.0%)	No (99.0-100.0%	Yes (99.0-100.0%)
ERR2092112	No (98.0-100.0%)	No (97.0-99.0%)	No (99.0-100.0%	No (91.0-94.0%	No (98.0-99.0%)	No (99.0-100.0%	No (99.0-100.0%)
ERR2092143	No (98.0-100.0%)	No (97.0-99.0%)	No (99.0-100.0%	No (91.0-94.0%	No (98.0-99.0%)	No (99.0-100.0%	No (99.0-100.0%)
ERR2092146	No (98.0-100.0%)	No (97.0-99.0%)	No (99.0-100.0%	No (91.0-94.0%	No (98.0-99.0%)	No (99.0-100.0%	No (99.0-100.0%)
ERR2092151	No (98.0-100.0%)	No (97.0-99.0%)	No (99.0-100.0%	No (91.0-94.0%	No (98.0-99.0%)	No (99.0-100.0%	No (99.0-100.0%)
ERR2092152	No (98.0-100.0%)	No (97.0-99.0%)	No (99.0-100.0%	No (91.0-94.0%	No (98.0-99.0%)	No (99.0-100.0%	No (99.0-100.0%)

Figura 2.46: Resultados en la detección de Depression y Schizophrenia

En la figura 2.46, se observa como el paciente **ERR2092109** padece *Depression* con un acierto entre el $91-94\,\%$ de las veces y *Substance abuse* con un intervalo de confianza en que será cierto de entre el $99 - 100\,\%$ de las veces.

Results of predictions

Patient Id	Anorexia nerv.	Bipolar disor.	Bulimia nerv.	Depression	Post traumatic	Schizophrenia	Subst. abuse
ERR2302863	No (98.0-99.0%)	No (98.0-100.0%	Yes (99.0-100.0%	No (90.0-94.0%	No (97.0-99.0%)	No (100.0-100.0%	No (99.0-100.0%)
ERR2308941	No (98.0-99.0%)	No (98.0-100.0%	No (99.0-100.0%	No (90.0-94.0%	No (97.0-99.0%)	No (100.0-100.0%	No (99.0-100.0%)
ERR2308943	No (98.0-99.0%)	No (98.0-100.0%	No (99.0-100.0%	No (90.0-94.0%	No (97.0-99.0%)	No (100.0-100.0%	No (99.0-100.0%)
ERR2308944	No (98.0-99.0%)	No (98.0-100.0%	No (99.0-100.0%	No (90.0-94.0%	No (97.0-99.0%)	No (100.0-100.0%	No (99.0-100.0%)
ERR2308945	No (98.0-99.0%)	No (98.0-100.0%	No (99.0-100.0%	No (90.0-94.0%	No (97.0-99.0%)	Yes (100.0-100.0'	No (99.0-100.0%)
ERR2308946	No (98.0-99.0%)	No (98.0-100.0%	No (99.0-100.0%	Yes (90.0-94.0%	No (97.0-99.0%)	Yes (100.0-100.0'	No (99.0-100.0%)

Figura 2.47: Resultados en la detección de Bulimia nervosa, Depression y Schizophrenia

En la figura 2.47, se observa como el paciente **ERR2302863** padece *Bulimia nervosa* con un acierto entre el $99 - 100\,\%$ de las veces, el paciente **ERR2308945** padece *Schizoprenia* con una veracidad del $100\,\%$ y por último, el paciente **ERR2308946** padece *Depression* entre el $90 - 94\,\%$ y también padece *Schizoprenia* con una veracidad del $100\,\%$.

Distribución de las predicciones

Como se ha comentado en el apartado que trata sobre los algoritmos, aquellos que presenten mejores resultados a la hora de ajustar el modelo, serán los utilizados en la predicción de las enfermedades mentales.

Distribution of predictions	
Patients with Anorexia nervosa mental illness 8	Random forest
Patients with Bipolar disorder mental illness 6	Support-vector machine Lineal
Patients with Bulimia nervosa mental illness 12	Random forest
Patients with Depression mental illness 208	Decision tree
Patients with Post traumatic stress disorder mental illness 20	Support-vector machine Lineal
Patients with Schizophrenia mental illness 2	Support-vector machine Polynomial
Patients with Substance abuse mental illness 3	Random forest

Figura 2.48: Distribución de los resultados obtenidos

Random Forest fue el que tuvo mejor ajuste para la detección de las enfermedades mentales de *Anorexia nervosa*, *Bulimia nervosa* y *Substance abuse*, detectando a 8, 12 y 3 pacientes con estos trastornos, respectivamente.

Support Vector Machine Lineal fue el que tuvo mejor ajuste para la detección de las enfermedades mentales de *Bipolar disorder* y *Post traumatic stress disorder*, detectando a 6 y 20 pacientes con estos trastornos, respectivamente.

Support Vector Machine Polynomial fue el que tuvo mejor ajuste para la detección de la enfermedad mental *Schizophrenia*, detectando a 2 pacientes con este trastorno.

Boost Decision Tree fue el que tuvo mejor ajuste para la detección de la enfermedad mental *Depression*, detectando a 208 pacientes con este trastorno.

Variables más significativas en la predicción

Este es el apartado más interesante, ya que muestra aquellas unidades traxonómicas operativas y aquellos hábitos de los pacientes más significativos a la hora de realizar la predicción. *MeanDecreaseGini* proporciona una medida de importancia que tiene en cuenta la contribución que la variable hace a la precisión como el grado de clasificación errónea.

En **Anorexia nervosa** la variable más importante e influyente es OTU *Bilophila*.

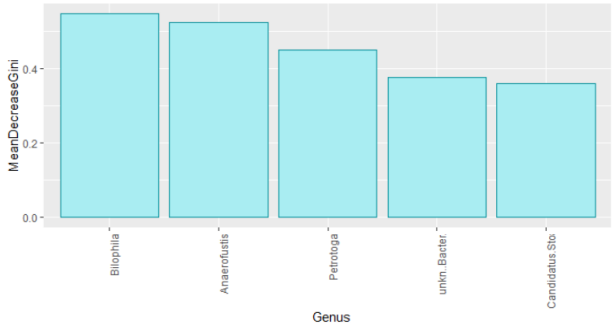

Figura 2.49: Variables más importantes en la predicción de la Anorexia nervosa

Unidad Taxonómica	MeanDecreaseGini
Bilophila	0.5461412
Anaerofustis	0.5227212
Petrotoga	0.4502741
unkn..Bacteroidales.o.	0.3752048
Candidatus.Stoquefichus	0.3584045

Cuadro 2.8: Valores de las variables más importantes en la predicción de la Anorexia nervosa

En **Bipolar disorder** la variable más importante e influyente es OTU *unkn..Chitinophagaceae.f.*.

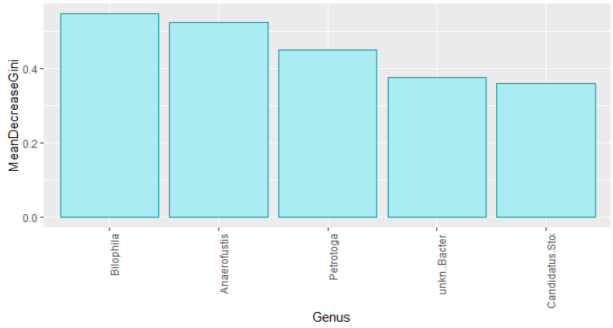

Figura 2.50: Variables más importantes en la predicción de la Bipolar disorder

Unidad Taxonómica	MeanDecreaseGini
unkn..Chitinophagaceae.f.	0.5542027
Odoribacter	0.5393127
unkn..Enterobacterales.o.	0.4937905
Lachnoanaerobaculum	0.4677285
Christensenella	0.4268557

Cuadro 2.9: Valores de las variables más importantes en la predicción de Bipolar disorder

En **Bulimia nervosa** la variable más importante e influyente es OTU *Lacihabitans*.

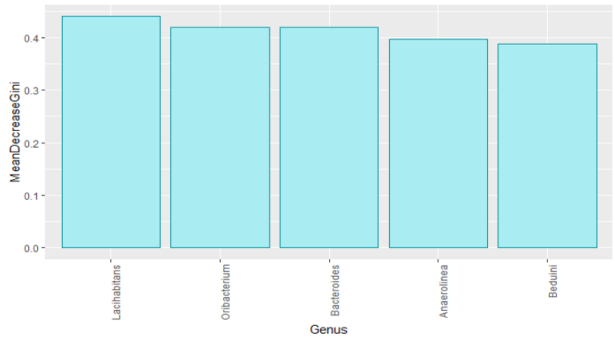

Figura 2.51: Variables más importantes en la predicción de la Bulimia nervosa

Unidad Taxonómica	MeanDecreaseGini
Lacihabitans	0.4396620
Oribacterium	0.4194204
Bacteroides	0.4191650
Anaerolinea	0.3972196
Beduini	0.3872044

Cuadro 2.10: Valores de las variables más importantes en la predicción de Bulimia nervosa

En **Depression** la variable más importante e influyente es OTU *Lachnoclostridium*.

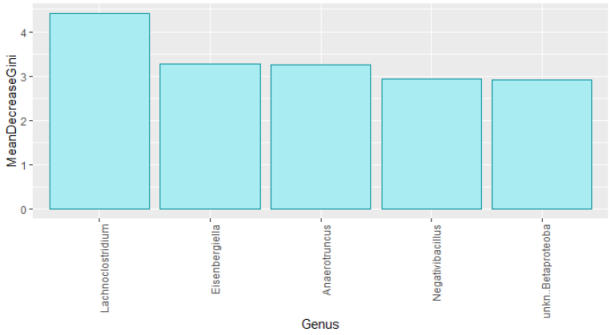

Figura 2.52: Variables más importantes en la predicción de la Depression

Unidad Taxonómica	MeanDecreaseGini
Lachnoclostridium	4.412455
Eisenbergiella	3.263250
Anaerotruncus	3.257128
Negativibacillus	2.923980
unkn..Betaproteobacteria.c.	2.903739

Cuadro 2.11: Valores de las variables más importantes en la predicción de la Depression

En **Post traumatic stress disorder** la variable más importante e influyente es OTU *Gabonibacter*.

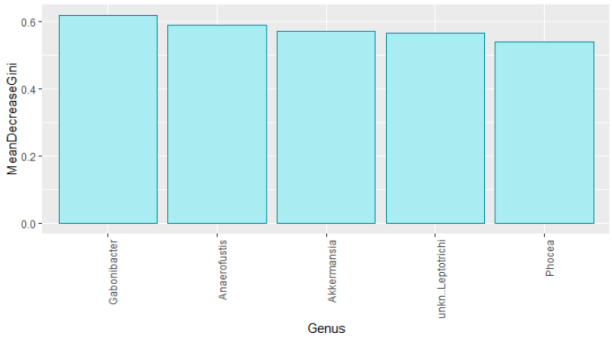

Figura 2.53: Variables más importantes en la predicción de la Post traumatic stress disorder

Unidad Taxonómica	MeanDecreaseGini
Gabonibacter	0.6202528
Anaerofustis	0.5897360
unkn..Leptotrichiaceae.f.	0.570665
Negativibacillus	0.5654520
Phocea	0.5392301

Cuadro 2.12: Valores de las variables más importantes en la predicción de la Post traumatic stress disorder

En **Schizophrenia** la variable más importante e influyente es OTU *Myxococcus*.

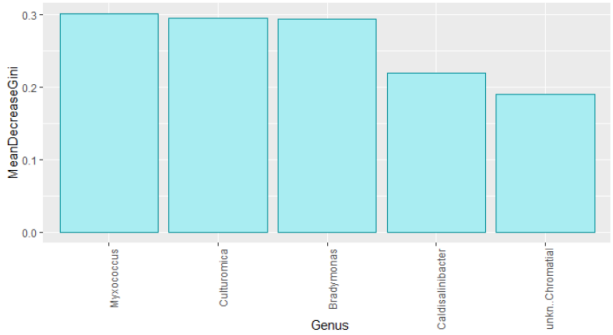

Figura 2.54: Variables más importantes en la predicción de la Schizophrenia

Unidad Taxonómica	MeanDecreaseGini
Myxococcus	0.3012020
Culturomica	0.2948985
Bradymonas	0.2934939
Caldisalinibacter	0.2200814
unkn..Chromatiales.o.	0.2200814

Cuadro 2.13: Valores de las variables más importantes en la predicción de la Schizophrenia

En **Susbstance abuse** la variable más importante e influyente es OTU *Dokdonia*.

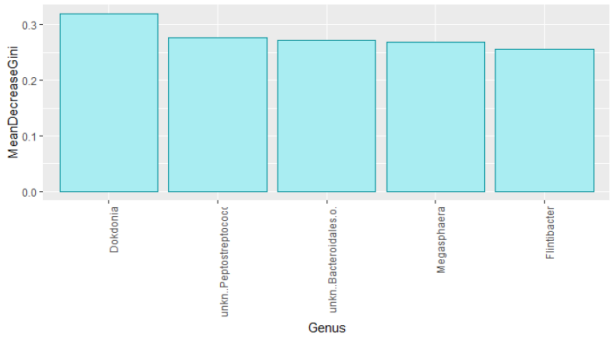

Figura 2.55: Variables más importantes en la predicción de Susbstance abuse

Unidad Taxonómica	MeanDecreaseGini
Dokdonia	0.3198506
unkn..Peptostreptococcaceae.f.	0.2758769
unkn..Bacteroidales.o.	0.2714078
Megasphaera	0.2684625
Flintibacter	0.2565395

Cuadro 2.14: Valores de las variables más importantes en la predicción de Susbstance abuse

Capítulo 3

Conclusiones

A medida que se ha ido profundizando más en el tema, han ido surgiendo nuevos retos y se han ido aplicando técnicas de análisis ya existentes, pero sobre todo, se han ido adaptando otras a las propias necesidades del estudio. Está claro, que no existe un patrón universal donde se presente un problema y éste quede resuelto sin necesidad de adaptarlo. El gran volumen de datos que se ha manejado en este proyecto y el bajo poder de cómputo de las herramientas utilizadas, ha obligado a agudizar aun más el ingenio.

Lo primero que hay que destacar es que todos los objetivos han sido cumplidos e incluso se ha querido ir más allá y se han conseguido otros que no estaban previstos inicialmente, pero que a medida que ha ido evolucionando en el proyecto ha sido oportuno añadir.

Este trabajo se compone de dos partes bien diferenciadas. La primera, basada en el estudio y análisis de los datos suministrados y una segunda, centrada en el desarrollo de una aplicación de software donde se pone en práctica el análisis anterior.

El objetivo principal en la primera parte y el fin en si mismo, era que a partir de una población de unidades taxonómicas, poder llegar a detectar cuáles son las más influyentes en cada una de las enfermedades mentales a analizar. El primer reto ha sido manejar toda esta información y prepararla para su posterior análisis y ajuste de un modelo capaz de realizar predicciones.

La dificultad ha estado en este punto en concreto, porque a pesar de disponer de más de 16000 pacientes con sus respectivos microbiomas, una gran parte de ellos no disponían de toda la información necesaria, o quedaron descartados una vez pasada la fase de filtrado, reduciéndose así la población a unos 2500 pacientes en total para el análisis.

Además, salvo en el caso de la Depression, el resto de enfermedades mentales tenían muy pocos casos positivos. Esto mismo ha dificultado la tarea a los algoritmos de clasificación y ha obligado a utilizar técnicas de *Re-sampling*. A pesar de que el poner en práctica estas técnicas podría dar la sensación de que se están manipulando los datos de entrada, lo que nos está diciendo esto realmente, es que con más datos positivos el modelo estará mejor ajustado y por lo tanto, la predicción será mejor. Este desbalanceo en los datos se ha solucionado de esta manera y a partir de ese momento, se empezaron a tener resultados significativos y realmente buenos.

Una vez realizada la preparación de los datos y la evaluación de los algoritmos, se ha desarrollado una aplicación de software que engloba todo este análisis, conviertiéndose en una herramienta de predicción de enfermedades mentales, ajustable a las necesidades del usuario.

Además de realizar dicha predicción, esta herramienta nos informa de aquellas variables o factores más importantes que son las que han influenciado en la decisión final. Por lo tanto, esta aplicación facilita la tarea del usuario, permitiendo realizar predicciones sobre enfermedades mentales de pacientes a partir del microbioma intestinal. Esto es extrapolable a cualquier otro tipo de microbioma, realizando los cambios pertinentes en el software. Cabe destacar que esta herramienta permite estudiar aquellos OTUs significativos en el resultado, además de permitir ajustar el balanceo de los datos en función de las necesidades del momento.

La planificación se ha seguido estrictamente, a pesar de la situación actual de pandemia que estamos viviendo actualmente. Se ha tenido que realizar un esfuerzo sobrehumano para poder cumplir los plazos, teniendo en cuenta que se debia compaginar el teletrabajo con el cuidado de mis dos hijos pequeños y la realización de este trabajo final de máster.

En este sentido, he de agradecer la extensión de tiempo de dos semanas que me ha permitido no solo cumplir con lo planificado, sino realizar aquellas mejoras no previstas, pero bajo mi punto de vista, importantes para el proyecto.

El trabajar en este proyecto me ha permitido pensar sobre temas todavía desconocidos para la investigación. Temas que me han rondado durante mucho tiempo en la cabeza y que, de una forma u otra, he podido dar a luz, como por ejemplo, un analizador en tiempo real del estado de nuestro microbioma, capaz de detectar alteraciones en la variabilidad de éste. Además, esta herramienta debería ser portátil y de fácil uso y adquisición, ya que el usuario final es el propio analizado. Y yendo mucho más allá, sería aun más efectivo que este analizador sea un dispositivo electrónico insertado en el propio intestino, suministrando información relevante de todo lo que está pasando en él.

La mente humana es el gran desconocido para la propia humanidad, donde diferentes ciencias intentan descrifrar y desvelar sus interioridades. ¿Somos conscientes, porque pensamos? ¿Quién o qué alimenta ese pensamiento? ¿Se mapeará la mente y se correlacionará cada parte de ella con nuestro microbioma?

Solo hay que ponerse manos a la obra, porque:
” *El hombre que mueve montañas empieza apartando piedrecitas.*” **CONFUCIO**

Glosario

AGP Plataforma abierta para la investigación de microbiomas de ciencia ciudadana. 13

BDT Boost decision tree son modelos de predicción que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva para la resolución de un problema. 4

Crohn Alteración de la microbiota intestinal. 12

disbiosis Alteración de la microbiota. 1, 12

Distancia de Manhattan Métrica en la que la distancia entre dos puntos es la suma de las diferencias (absolutas) de sus coordenadas. 18

JSON Formato de texto sencillo para el intercambio de datos. 4, 14

K-means Método de agrupamiento. 18

kappa Medida estadística que ajusta el efecto del azar. 26, 27, 30–32, 35, 36

Machine learning Conjunto de algoritmos que transforman los datos en conocimiento procesable. 2

Metadatos Datos que describen otros datos. 4

microbiómico Microbioma es el conjunto total de genes de la microbiota. 12

microbiota Conjunto de microorganismos que conviven simbióticamente con el nuestro. 11

OTU Operational Taxonomic Unit, es una unidad muy utilizada en la investigación de la diversidad microbiana. 4, 5, 13, 16, 54

Over sampling Sobremuestreo de los resultados positivos. 22, 27, 36, 41

Python Lenguaje de programación orientado a objetos interpretado. 3

R Entorno y lenguaje de programación con un enfoque al análisis estadístico. 3

Re-sampling Conjunto de métodos para garantizar que el modelo es lo suficientemente bueno para manejar variaciones en los datos. 22, 30, 34, 35, 54

RF Random Forest está basado, principalmente, en los Árboles de Decisión con la diferencia de que cada nodo está dividido por el mejor de un subconjunto aleatorio de variables, en lugar de la mejor de todas las variables. 4

RPY2 Proporciona acceso desde Python para utilizar funciones R. 37

SVM Support Vector Machine son un conjunto de algoritmos de aprendizaje supervisado relacionados con problemas de clasificación y regresión. 4

taxonomía Ciencia que trata de los principios, métodos y fines de la clasificación, generalmente científica; se aplica, en especial, dentro de la biología para la ordenación jerarquizada y sistemática de los grupos de animales y de vegetales. 2

Under sampling Submuestreo de los resultados negativos. 22, 31

Bibliografía

[1] Lantz, B. (2015). ″Machine Learning with R: discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R″.

[2] Everitt, B., Hothorn, T. (2011). ″An introduction to applied multivariate analysis with R. New York: Springer″.

[3] Michel Aceves, R., Izeta Gutiérrez, AC., Torres Alarcón, G., Margarita Michel Izeta, AC. (2017). ″ La microbiota y el microbioma intestinal humano. (Entre las llaves del reino y una nueva caja de Pandora)″.

[4] Salaverry, O. (2012). ″The stone of madness: Starting points of the history of mental health″.

[5] Alles-Colomer, M., Falony, G., Darzi, Y. et al. (2019). ″The neuroactive potential of the human gut microbiota in quality of life and depression″.

[6] Svoboda, E. (2020). ″Could the gut microbiome be linked to autism?″

[7] Salagre, E., Vieta, E., Grande, I. (2017). ″The visceral brain: Bipolar disorder and microbiota″.

[8] Levine, S. S., & Prietula, M. J. (2014). ″Open collaboration for innovation: Principles and performance.″. Organization Science, 25(5), 1414-1433. doi: 10.1287orsc.2013.0872

[9] Ho, Tin Kam (1995). ″Random Decision Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition″, Montreal, QC,

14–16 August 1995. pp. 278-282. Archivado desde el original el 4 de julio de 2008.

[10] Handling Imbalanced Datasets in Deep Learning. https://towardsdatascience.com/handling-imbalanced-datasets-in-deep.

[11] Hannah Ritchie and Max Roser (2019). ″Mental Health″. Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/mental-health'

[12] OMS, ACNUR (2015). ″Evaluación de necesidades y recursos psicosociales y de salud mental″. Número de páginas: 84, Fecha de publicación: 2015, Idiomas: Árabe, español, francés, inglés, ruso; ISBN: 978 92 4 354853 1

[13] Daniel McDonald, Embriette Hyde, Justine W. Debelius, James T. Morton, Antonio Gonzalez, Gail Ackermann, Alexander A. Aksenov, Bahar Behsaz, Caitriona Brennan, Yingfeng Chen, Lindsay DeRight Goldasich, Pieter C. Dorrestein, Robert R. Dunn, Ashkaan K. Fahimipour, James Gaffney, Jack A. Gilbert, Grant Gogul, Jessica L. Green, Philip Hugenholtz, Greg Humphrey, Curtis Huttenhower, Matthew A. Jackson, Stefan Janssen, Dilip V. Jeste, Lingjing Jiang, Scott T. Kelley, Dan Knights, Tomasz Kosciolek, Joshua Ladau, Jeff Leach, Clarisse Marotz, Dmitry Meleshko, Alexey V. Melnik, Jessica L. Metcalf, Hosein Mohimani, Emmanuel Montassier, Jose Navas-Molina, Tanya T. Nguyen, Shyamal Peddada, Pavel Pevzner, Katherine S. Pollard, Gholamali Rahnavard, Adam Robbins-Pianka, Naseer Sangwan, Joshua Shorenstein, Larry Smarr, Se Jin Song, Timothy Spector, Austin D. Swafford, Varykina G. Thackray, Luke R. Thompson, Anupriya Tripathi, Yoshiki Vázquez-Baeza, Alison Vrbanac, Paul Wischmeyer, Elaine Wolfe, Qiyun Zhu, The American Gut Consortium, Rob Knight, Casey S. Greene (2018). ″American Gut: an Open Platform for Citizen Science Microbiome Research″. http://americangut.org; Editor DOI: 10.1128/mSystems.00031-18

[14] NCBI (2015). ″American Gut Project″, BioProject: https://www.ncbi.nlm.nih.gov/bioproject?term=PRJEB11419&cmd=DetailsSearch

[15] Wing Sun Faith Chung, Alan W. Walker, Petra Louis, Julian Parkhill, Joan Vermeiren, Douwina Bosscher, Sylvia H. Duncan, and Harry J. Flint (2016). ″Modulation of the human gut microbiota by dietary fibres occurs at the species level″ Published online 2016 Jan 11. doi: 10.1186/s12915-015-0224-3

[16] Lin, Chih-Jen (2012). ″Machine learning software: design and practical use ″. Machine Learning Summer School. Kyoto.

[17] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, Daniel R Mende, Gabriel R Fernandes, Julien Tap, Thomas Bruls, Jean-Michel Batto, Marcelo Bertalan, Natalia Borruel, Francesc Casellas, Leyden Fernandez, Laurent Gautier, Torben Hansen, Masahira Hattori, Tetsuya Hayashi, Michiel Kleerebezem, Ken Kurokawa, Marion Leclerc, Florence Levenez, Chaysavanh Manichanh, H Bjørn Nielsen, Trine Nielsen, Nicolas Pons, Julie Poulain, Junjie Qin, Thomas Sicheritz-Ponten, Sebastian Tims, David Torrents, Edgardo Ugarte, Erwin G Zoetendal, Jun Wang, Francisco Guarner, Oluf Pedersen, Willem M de Vos, Søren Brunak, Joel Doré, MetaHIT Consortium; María Antolín, François Artiguenave, Hervé M Blottiere, Mathieu Almeida, Christian Brechot, Carlos Cara, Christian Chervaux, Antonella Cultrone, Christine Delorme, Gérard Denariaz, Rozenn Dervyn, Konrad U Foerstner, Carsten Friss, Maarten van de Guchte, Eric Guedon, Florence Haimet, Wolfgang Huber, Johan van Hylckama-Vlieg, Alexandre Jamet, Catherine Juste, Ghalia Kaci, Jan Knol, Omar Lakhdari, Severine Layec, Karine Le Roux, Emmanuelle Maguin, Alexandre Mérieux, Raquel Melo Minardi, Christine M'rini, Jean Muller, Raish Oozeer, Julian Parkhill, Pierre Renault, Maria Rescigno, Nicolas Sanchez, Shinichi Sunagawa, Antonio Torrejon, Keith Turner, Gaetana Vandemeulebrouck, Encarna Varela, Yohanan Winogradsky, Georg Zeller, Jean Weissenbach, S Dusko Ehrlich, Peer Bork (2011). ″Enterotypes of the Human Gut Microbiome″. Nature. 2011 May 12;473(7346):174-80. doi: 10.1038/nature09944. Epub 2011 Apr 20.

[18] Colín, Carlos A. Núñez, & Alejandro F. Barrientos Priego (2006). ″Estimación de la variabilidad interna de muestras poblacionales, mediante análisis de

componentes principales″. Interciencia: Revista de ciencia y tecnología de América, 31(11), 802-806.

[19] «rstudio/rstudio». GitHub. ″RStudio″. Consultado el 18 de diciembre de 2016.

[20] Guttag, John V. (12 August 2016). ″Introduction to Computation and Programming Using Python: With Application to Understanding Data″. MIT Press. ISBN 978-0-262-52962-4.

[21] Laurent Gautier (embedded R) (2020). ″Python interface to the R language″. https://pypi.org/project/rpy2/

Anexos

Anexo A

Métodos y funciones de la aplicación

El código fuente de la aplicación y del análisis de datos se encuentran disponibles en *https://bitbucket.org/derdritteraum/workspace/projects/BIOIN*

A.1. Front-end

Para poder permitir al usuario acceder a todas las funcionalidades de la aplicación, ha sido necesario implementar una GUI. Para ello, se ha creado una clase llamada MIP-Window desde la que el usuario realizará, de forma directa o indirecta, todas las acciones necesarias para completar su predicción.

- En el constructor se declaran principalmente aquellas variables que se van a ir actualizando durante todo el proceso, el estilo de la GUI y la fuente de los textos, y sobre todo las dimensiones de la ventana, título, icono y colores.
- Método **layout()** permite habilitar todos los tabs de la aplicación y llamar a aquellos métodos responsables de su implementación.
- Método **patients_tab()** permite mostrar al usuario todo lo relacionado con la predicción.
- Método **taxonomy_tab()** permite mostrar al usuario todo lo relacionado con la taxonomía.

- Método **metadata_tab()** permite mostrar al usuario todo lo relacionado con los metadatos.
- Método **classifiers_tab()** permite mostrar al usuario todo lo relacionado con los algoritmos de clasificación.
- Método **results_tab()** permite mostrar al usuario todo lo relacionado con los resultados obtenidos después del entrenamiento.
- Método **prediction_tab()** permite mostrar al usuario todos los resultados obtenidos después de la predicción.
- Método **about_tab()** permite mostrar al usuario información sobre el proyecto y partes implicadas en él.
- Método **generic_result()** permite mostrar los resultados obtenidos después del entrenamiento tanto a nivel numérico como a nivel gráfico. Esta pantalla al ser genérica se reutilizará para todas las diferentes enfermedades mentales.
- Método **simple_chart()** permite realizar una representación gráfica de un resultado por algoritmo.
- Método **multi_chart()** ermite realizar una representación gráfica de varios resultados por algoritmo.
- Método **start_taxonomy_filtered()** permite cargar los ficheros de taxonomía y aplicar los filtros, tanto para la fase de entrenamiento como para la fase de predicción.
- Método **start_metadata_filtered()** permite cargar los ficheros de metadatos y aplicar los filtros, tanto para la fase de entrenamiento como para la fase de predicción.
- Método **start_generate_model()** permite generar el modelo para cada una de las enfermedades mentales.
- Método **execute_prediction()** permite realizar la predicción basándose en el modelo ajustado para cada una de las enfermedades mentales.

- Método **xxx_RFunction_calculations()** permiten calcular todos los resultados para cada algoritmo para cada enfermedad mental.
- Método **decide_best_algorithm()** permite decidir cuál de los algoritmos seleccionados en la fase de entrenamiento ha obtenido mejor resultado para cada enfermedad mental, y por lo tanto, será el utilizado en la fase de predicción. Este criterio está basado en obtener el mejor resultado después de sumar la precisión accuracy y el estadístico kappa.

A.2. Back-end

Para realizar la extracción de los datos taxonómicos y su posterior filtrado, se ha creado una clase llamada **Taxonomy** preparada para gestionar toda la información sobre este tipo de datos.

- **load_file()** se encarga de realizar la lectura de los datos desde el fichero de taxonomía para poder crear una matriz con ellos.
- **create_matrix()** es responsable directo de la creación de la matriz de datos. Recoge línea a línea el contenido del fichero y las convierte en listas.
- **insert_list_in_matrix()** toma cada una de estas líneas leídas desde el fichero y separa cada uno de los elementos que completará finalmente la lista.

Con estos tres métodos se pasa toda la información del fichero a una matriz o lista de listas. El siguiente paso es el filtrado o supresión de aquella información no necesaria. En este caso lo que se pretende es suprimir a aquellos individuos de la lista con un porcentaje de Pseudomonas superior al 3 % y un porcentaje de (unkn.) Enterobacteiaceae(f) también superior al 3 %.

- **microorganism_filter()** guarda una lista de índices de pacientes que no cumplen los requisitos del filtro, para ser eliminados de la matriz de taxonomía. Este método recibe como parámetro un diccionario con el microorganismo y el porcentaje. De esta manera, sin en un futuro se quisieran añadir más microorganismos u otros valores en los porcentajes, no habría que modificarlo.

- **filter_taxonomy_matrix()** es el responsable directo de eliminar aquellos pacientes que no cumplen con las condiciones deseadas. Recibe una lista de índices a suprimir, generada por el método *microorganism_filter()*, y de mayor a menor procede a la supresión.
- **get_taxonomy_matrix()** retorna la matriz de taxonomía una vez realizados todos los filtrados.
- **get_patientsId_list()** retorna la lista con todos los identificadores de los pacientes que han pasado el filtrado.
- **get_microorganism_list()** retorna la lista de los nombres de todos los microorganismos.
- **change_bacteria_name()** reduce el nombre de la bacteria de una manera drástica para un mejor manejo de los datos y evitar nombre muy largos.
- **get_bacteria_name()** retorna el nombre original de la bacteria, es decir, su versión larga del nombre.
- **write_file()** permite generar un fichero, con los datos obtenidos después del filtrado, de la matriz de taxonomía.

Para realizar la extracción de los metadatos de cada uno de los pacientes, se ha utilizado la tecnología Python. Para este caso en concreto, se ha creado una clase llamada **Metadata** preparada para gestionar toda la información sobre este tipo de datos.

- **filter_by_taxonomy()** se encarga de la obtención de la matriz de taxonomía generada a través de la clase Taxonomy.
- **filter_by_metadata()** es el más importante de la clase, ya que es responsable del filtrado total de todos los datos. Toma la lista de pacientes válidos en el filtrado taxonómico y uno a uno comprobará si cumplen con los requisitos de filtrado.
- Índice de masa corporal (BMI) si este valor no está entre los valores **18** y **30** o no se tuviera información al respecto, este paciente quedará descartado del juego de datos a analizar.

- Padecer enfermedad intestinal **(IBD)** si el resultado es un diagnóstico positivo, ya sea por un profesional médico o un practicante de medicina alternativa, por el mismo paciente o no se tuviera información al respecto, este paciente quedará descartado del juego de datos a analizar.

- Padecer**Diabetes** si el resultado es un diagnóstico positivo, ya sea por un profesional médico o un practicante de medicina alternativa, por el mismo paciente o no se tuviera información al respecto, este paciente quedará descartado del juego de datos a analizar.

- Haber **tomado** algún tipo de **antibiótico** en los últimos seis meses también descartará a este paciente del juego final de datos.

- Es importante tener información sobre si los pacientes tienen o no alguna enfermedad mental. En el caso en que el valor obtenido en todas las enfermedades mentales no este especificado, *Unspecified*, también quedará descartado el paciente.

Cuando un paciente supera todos estos filtros será insertada toda esta información más otras características al conjunto de datos. Se han creado otros métodos que ayudan a la obtención de los resultados, preparándolos para el análisis.

- **check_characteristic_if_has_condition()** recibe un resultado obtenido de una característica del paciente y se retorna un resultado más significativo o con una longitud del literal mucho más corto. Como por ejemplo, si se recibe *'Diagnosed by a medical professional (doctor, physician assistant)'* se retornará *'Medical diagnose'*.

- **check_mental_illness_character_result()** retorna *'True'* o *'False'* dependiendo si se padece o no cierta enfermedad mental y en cualquier otro caso se retornará sin especificar *'Unspecified'*.

- **getCharacteristics()** dependiendo de la característica a buscar retorna su valor.

Una vez obtenido todos los datos, el siguiente paso es unificarlos con la matriz de taxonomía.

- **merge_taxonomy_and_metadata_matrix()** añade la final de la matriz de taxonomía todos los resultados obtenidos a partir de los metadatos de cada paciente.

Puede suceder que una bacteria sea irrelevante debido a la falta de resultados o incluso por carecer de ellos. Por lo tanto, ya que no va a realizar ninguna aportación al estudio y además va a incrementar el consumo de los recursos del sistema en los cálculos, se procederá a eliminar aquellas poco relevantes.

- **clean_merge_matrix()** permite eliminar bacterias irrelevantes. Para determinar cuáles son, se comprobará el número de valores diferente a '0' encontrados.
- **write_file_dataset()** tomará lista a lista de la matriz resultante y las insertará en un fichero a modo de juego de datos. Éste está preparado para su uso con la tecnología R.

A.3. Llamadas a las funciones R

Para realizar llamadas a las funciones de R se ha necesitado crear una nueva clase llamada RFunctions que se encarga de todas estas tareas.

- En el constructor se cargan todas las librerías necesarias, como por ejemplo las librerías que procesan los algoritmos de entrenamiento.
- Método **load_data()** y **load_data_prediction()** tienen el mismo cometido, que es la de cargar los ficheros *.csv* generados en los pasos previos al entrenamiento. La diferencia es que el primer método carga el fichero *.csv* para el entrenamiento, y el segundo método carga el fichero *.csv* para la predicción.
- Método **distribute_data()** y **distribute_data_prediction()** se encargan de distribuir los datos leídos del juego de datos en función de si son taxonómicos, o metadatos para el entrenamiento y la predicción, y si son enfermedades mentales que en este caso solo afecta al entrenamiento.
- Método **normalize_data()** y **normalize_data_prediction()** realizan la normalización de los datos para el entrenamiento y para la predicción.

- Método **add_mental_illness_data()** permite añadir cada una de las enfermedades mentales al dataframe final.
- Método **f_under_sampling()** permite recortar resultados negativos del juego de datos.
- Método **f_over_sampling()** permite añadir resultados positivos al juego de datos.
- Método **f_confusion_matrix_xxx()** permite crear la matriz de confusión para cada una de las enfermedades mentales tomadas como variables de salida.
- Método **calculate_result()** permite obtener el resultado después del entrenamiento y después de la predicción.
- Método **get_patientsId_prediction_list()** retorna una lista con los identificadores de cada uno de los pacientes a los que se les ha hecho la predicción.

Anexo B

Distribuciones

B.1. Estadística descriptiva

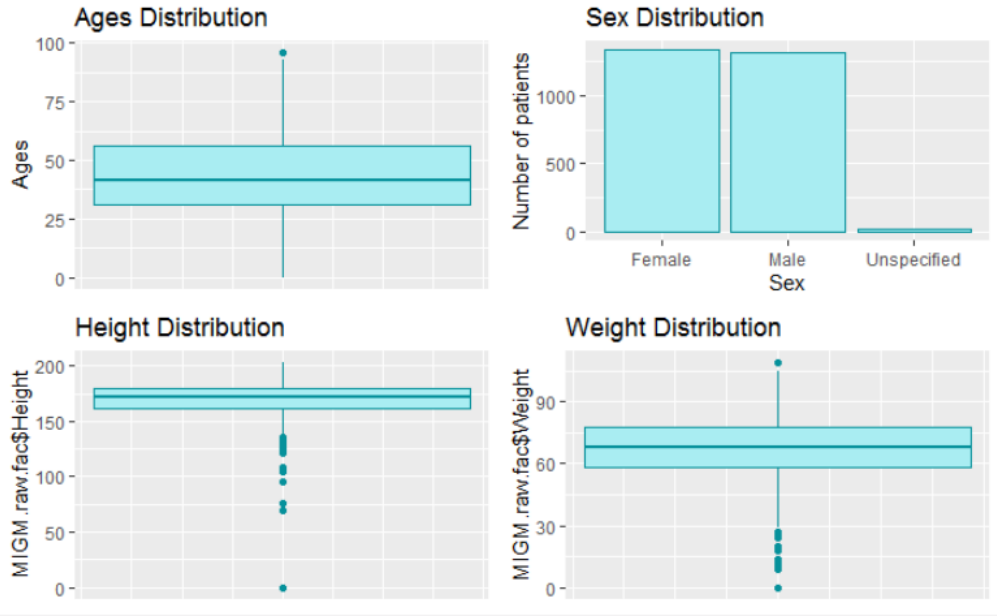

Figura B.1: Distribución de los datos referente a la edad, el sexo, la altura y el peso

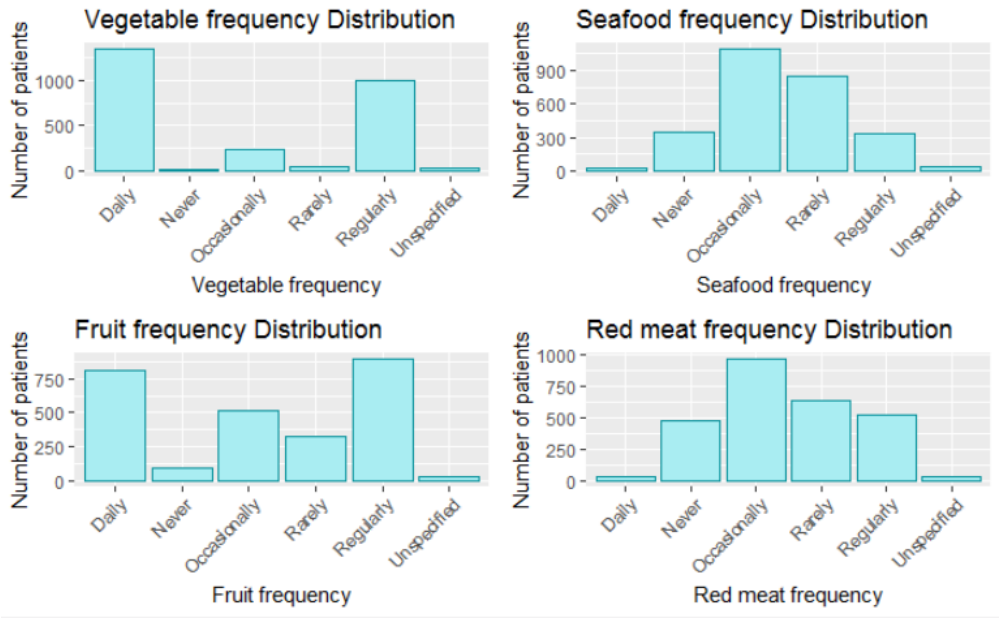

Figura B.2: Representación gráfica de frecuencia en la ingesta de vegetales, marisco, fruta y carne roja

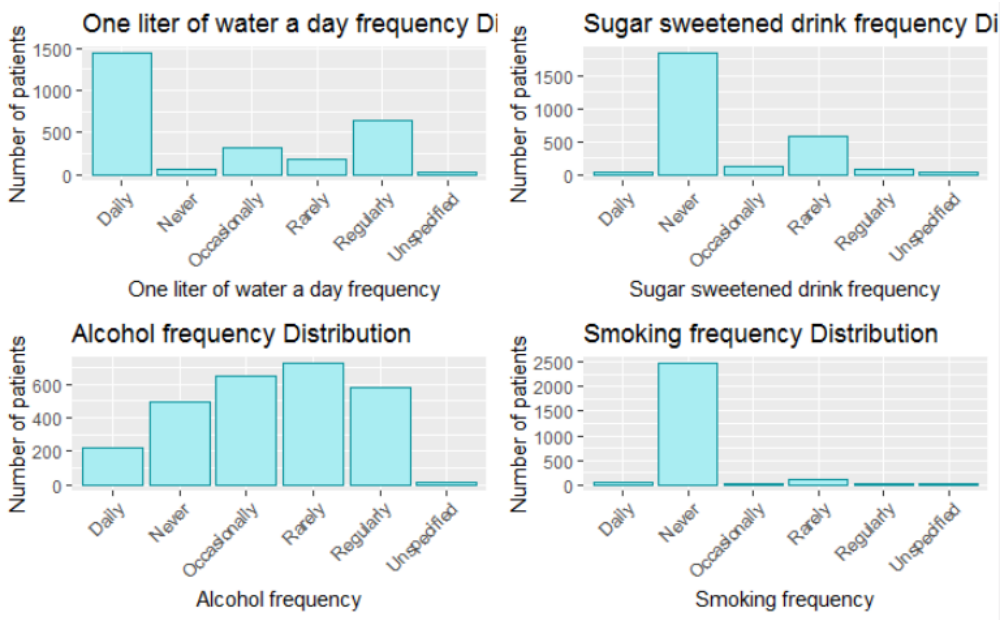

Figura B.3: Representación gráfica de frecuencia en la toma de un litro de agua, bebidas azucaradas, alcohol y tabaco

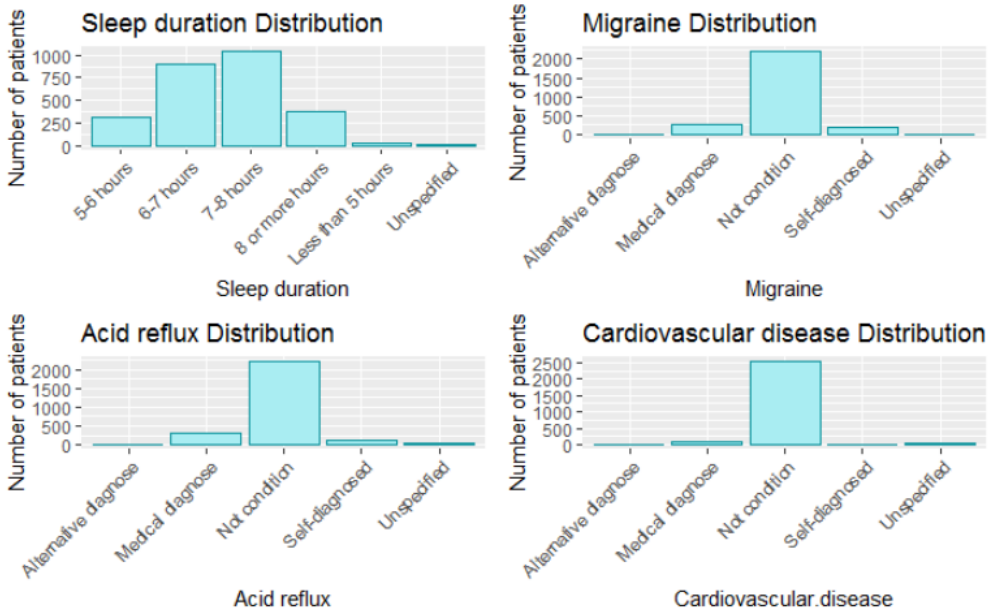

Figura B.4: Representación gráfica de los datos referente a la duración del sueño, diagnóstico de tener migraña, reflujo de ácidos y enfermedades cardiovasculares

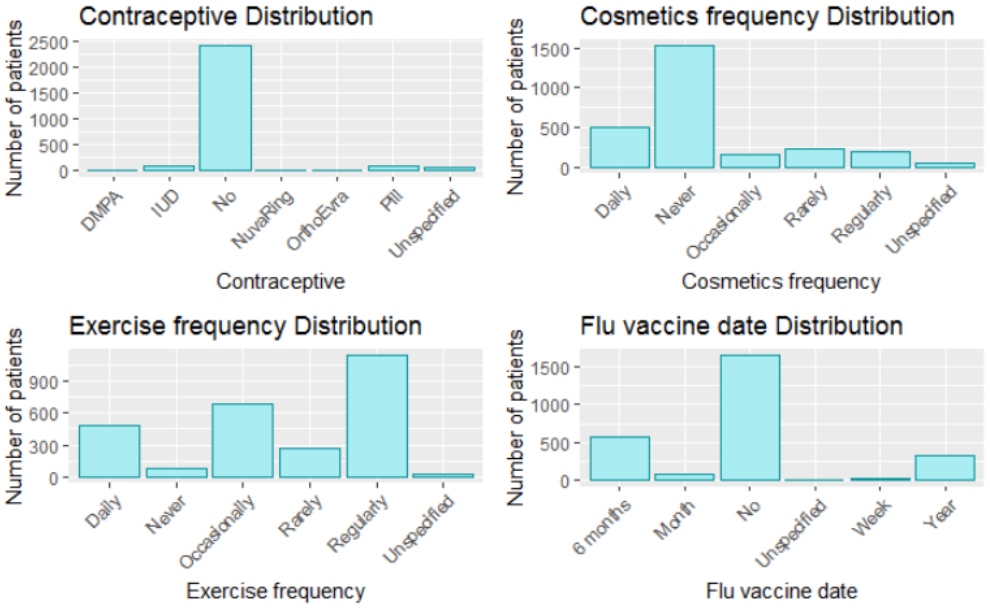

Figura B.5: Representación gráfica de frecuencia del uso de cosméticos, ejercicio y vacuna para la gripe, además del tipo de anticonceptivo usado

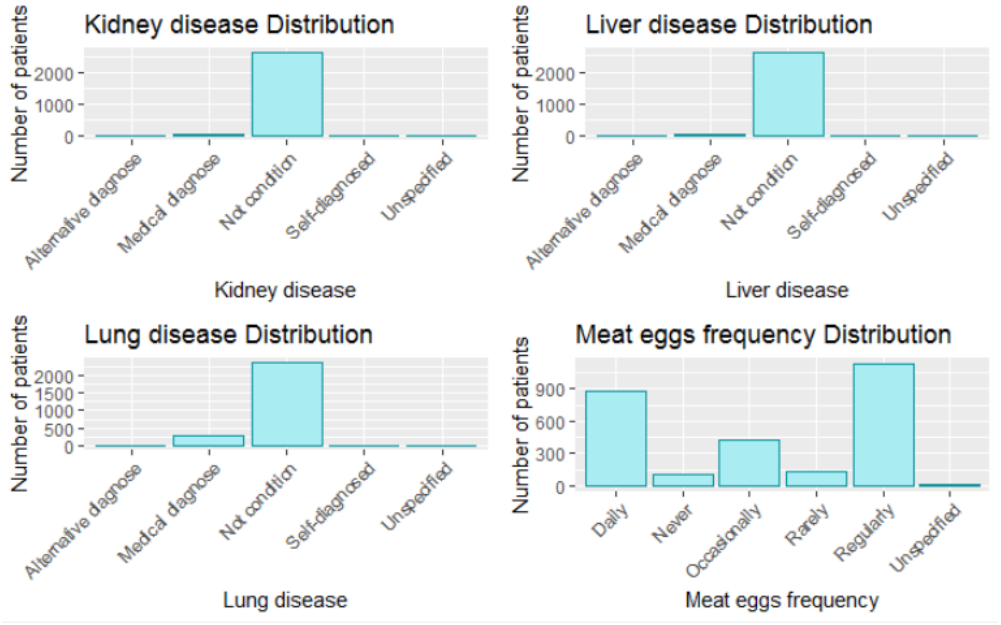

Figura B.6: Representación gráfica de los datos referente enfermedades de riñón, hígado y pulmón, y la frecuencia de consumo de huevo de carne.

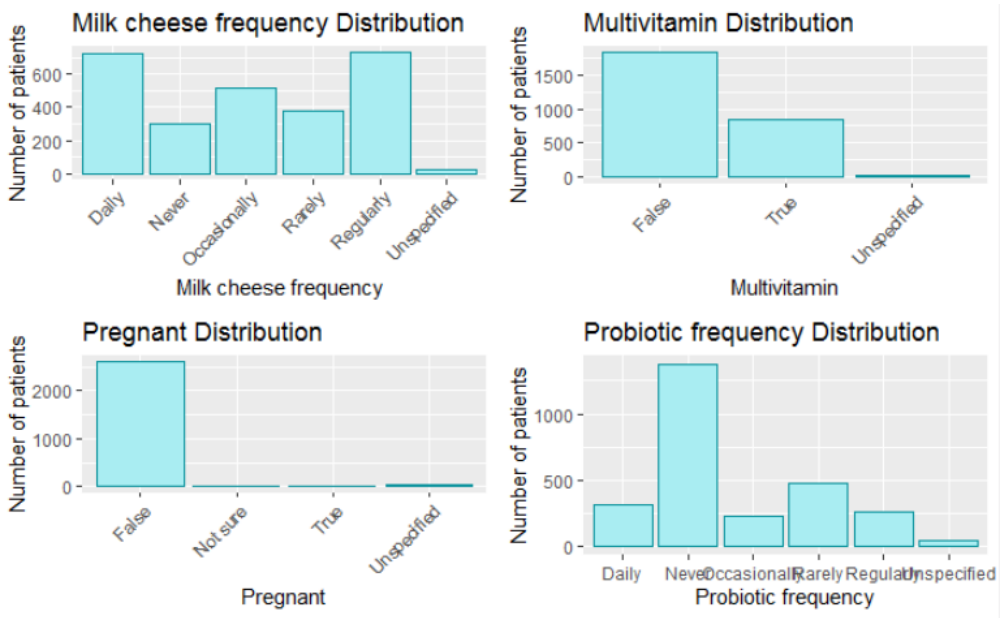

Figura B.7: Representación gráfica de frecuencia del consumo de queso, multivitaminas y probióticos, además de la condición de estar embarazada

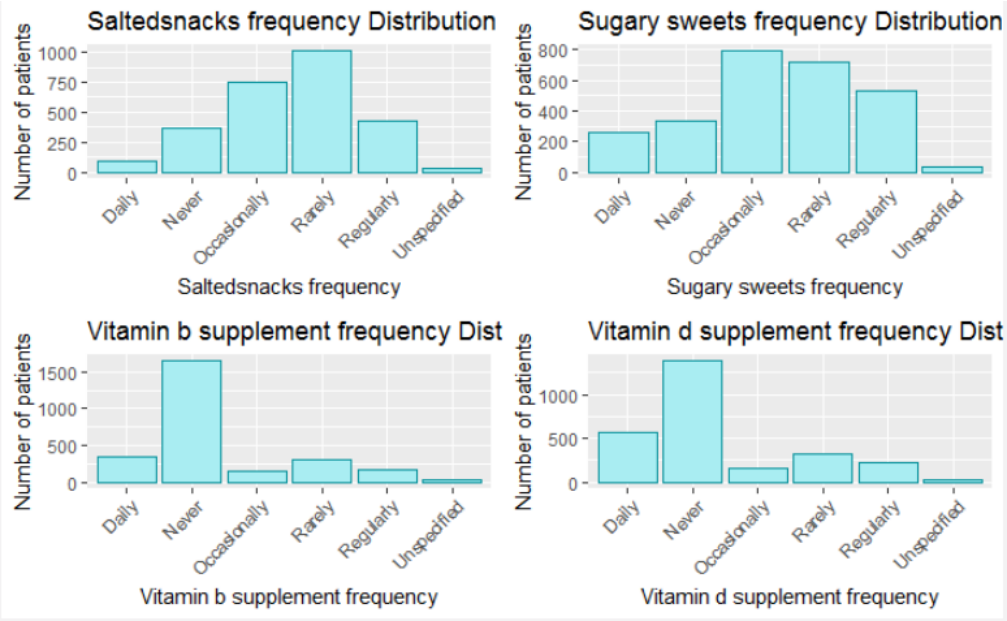

Figura B.8: Representación gráfica de frecuencia del consumo de snaks, azucar, vitamina b y vitamina d

Anexo C

Clasificadores

A continuación, se muestran las tablas de confusión para cada uno de los clasificadores en cada una de la enfermedades mentales, después de ajustar el modelo.

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   876    6
##             True      0    0
##
##                Accuracy : 0.9932
##                  95% CI : (0.9853, 0.9975)
##     No Information Rate : 0.9932
##     P-Value [Acc > NIR] : 0.60630
##
##                   Kappa : 0
```

Figura C.1: No sampling Anorexia - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   881    1
##             True      0    0
##
##                Accuracy : 0.9989
##                  95% CI : (0.9937, 1)
##     No Information Rate : 0.9989
##     P-Value [Acc > NIR] : 0.7358
##
##                   Kappa : 0
```

Figura C.2: No sampling Anorexia - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   876    5
##             True      1    0
##
##                Accuracy : 0.9932
##                  95% CI : (0.9853, 0.9975)
##     No Information Rate : 0.9943
##     P-Value [Acc > NIR] : 0.7626
##
##                   Kappa : -0.0019
```

Figura C.3: No sampling Anorexia - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   873    5
##             True      2    2
##
##                Accuracy : 0.9921
##                  95% CI : (0.9837, 0.9968)
##     No Information Rate : 0.9921
##     P-Value [Acc > NIR] : 0.5987
##
##                   Kappa : 0.3599
```

Figura C.4: No sampling Anorexia - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   176    5
##             True      0    0
##
##                Accuracy : 0.9724
##                  95% CI : (0.9367, 0.9915)
##     No Information Rate : 0.9724
##     P-Value [Acc > NIR] : 0.61597
##
##                   Kappa : 0
```

Figura C.5: Under sampling Anorexia - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   177    4
##             True      0    0
##
##                Accuracy : 0.9779
##                  95% CI : (0.9444, 0.9939)
##     No Information Rate : 0.9779
##     P-Value [Acc > NIR] : 0.6288
##
##                   Kappa : 0
```

Figura C.6: Under sampling Anorexia - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   176    1
##             True      3    1
##
##                Accuracy : 0.9779
##                  95% CI : (0.9444, 0.9939)
##     No Information Rate : 0.989
##     P-Value [Acc > NIR] : 0.9483
##
##                   Kappa : 0.3234
```

Figura C.7: Under sampling Anorexia - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   173    4
##             True      1    3
##
##                Accuracy : 0.9724
##                  95% CI : (0.9367, 0.991)
##     No Information Rate : 0.9613
##     P-Value [Acc > NIR] : 0.2957
##
##                   Kappa : 0.5323
```

Figura C.8: Under sampling Anorexia - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   877    4
##             True      0    5
##
##                Accuracy : 0.9955
##                  95% CI : (0.9885, 0.9988)
##     No Information Rate : 0.9898
##     P-Value [Acc > NIR] : 0.05411
##
##                   Kappa : 0.7122
```

Figura C.9: Over sampling Anorexia - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   878    2
##             True      0    6
##
##                Accuracy : 0.9977
##                  95% CI : (0.9919, 0.9997)
##     No Information Rate : 0.991
##     P-Value [Acc > NIR] : 0.01346
##
##                   Kappa : 0.856
```

Figura C.10: Over sampling Anorexia - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   877    6
##             True      0    3
##
##                Accuracy : 0.9932
##                  95% CI : (0.9853, 0.9975)
##     No Information Rate : 0.9898
##     P-Value [Acc > NIR] : 0.20539
##
##                   Kappa : 0.4974
```

Figura C.11: Over sampling Anorexia - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.anorexia False True
##             False   879    1
##             True      1    5
##
##                Accuracy : 0.9977
##                  95% CI : (0.9919, 0.9997)
##     No Information Rate : 0.9932
##     P-Value [Acc > NIR] : 0.06136
##
##                   Kappa : 0.8322
```

Figura C.12: Over sampling Anorexia - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   874    8
##            True      0    0
##
##                Accuracy : 0.9909
##                  95% CI : (0.9822, 0.9961)
##     No Information Rate : 0.9909
##     P-Value [Acc > NIR] : 0.59255
##
##                   Kappa : 0
```

Figura C.13: No sampling Bipolar - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   875    7
##            True      0    0
##
##                Accuracy : 0.9921
##                  95% CI : (0.9837, 0.9968)
##     No Information Rate : 0.9921
##     P-Value [Acc > NIR] : 0.59871
##
##                   Kappa : 0
```

Figura C.14: No sampling Bipolar - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   863   13
##            True      6    0
##
##                Accuracy : 0.9785
##                  95% CI : (0.9666, 0.987)
##     No Information Rate : 0.9853
##     P-Value [Acc > NIR] : 0.9585
##
##                   Kappa : -0.0094
```

Figura C.15: No sampling Bipolar - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   873    6
##            True      3    0
##
##                Accuracy : 0.9898
##                  95% CI : (0.9807, 0.9953)
##     No Information Rate : 0.9932
##     P-Value [Acc > NIR] : 0.9168
##
##                   Kappa : -0.0046
```

Figura C.16: No sampling Bipolar - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   174    9
##            True      0    0
##
##                Accuracy : 0.9508
##                  95% CI : (0.9087, 0.9773)
##     No Information Rate : 0.9508
##     P-Value [Acc > NIR] : 0.587436
##
##                   Kappa : 0
```

Figura C.17: Under sampling Bipolar - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   172   11
##            True      0    0
##
##                Accuracy : 0.9399
##                  95% CI : (0.895, 0.9696)
##     No Information Rate : 0.9399
##     P-Value [Acc > NIR] : 0.579305
##
##                   Kappa : 0
```

Figura C.18: Under sampling Bipolar - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   169   11
##            True      2    1
##
##                Accuracy : 0.929
##                  95% CI : (0.8816, 0.9616)
##     No Information Rate : 0.9344
##     P-Value [Acc > NIR] : 0.6852
##
##                   Kappa : 0.11
```

Figura C.19: Under sampling Bipolar - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   171    6
##            True      5    1
##
##                Accuracy : 0.9399
##                  95% CI : (0.895, 0.9696)
##     No Information Rate : 0.9617
##     P-Value [Acc > NIR] : 0.9501
##
##                   Kappa : 0.1229
```

Figura C.20: Under sampling Bipolar - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   874    8
##            True      0    7
##
##                Accuracy : 0.991
##                  95% CI : (0.9823, 0.9961)
##     No Information Rate : 0.9831
##     P-Value [Acc > NIR] : 0.03630
##
##                   Kappa : 0.6324
```

Figura C.21: Over sampling Bipolar - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   873   12
##            True      0    4
##
##                Accuracy : 0.9865
##                  95% CI : (0.9765, 0.993)
##     No Information Rate : 0.982
##     P-Value [Acc > NIR] : 0.190722
##
##                   Kappa : 0.3956
```

Figura C.22: Over sampling Bipolar - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   867    9
##            True      4    9
##
##                Accuracy : 0.9854
##                  95% CI : (0.9751, 0.9922)
##     No Information Rate : 0.9798
##     P-Value [Acc > NIR] : 0.1400
##
##                   Kappa : 0.5734
```

Figura C.23: Over sampling Bipolar - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bipolar False True
##            False   873    5
##            True      2    9
##
##                Accuracy : 0.9921
##                  95% CI : (0.9838, 0.9968)
##     No Information Rate : 0.9843
##     P-Value [Acc > NIR] : 0.03066
##
##                   Kappa : 0.7161
```

Figura C.24: Over sampling Bipolar - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   872   10
##            True      0    0
##
##                Accuracy : 0.9887
##                  95% CI : (0.9792, 0.9946)
##     No Information Rate : 0.9887
##     P-Value [Acc > NIR] : 0.583041
##
##                   Kappa : 0
```

Figura C.25: No sampling Bulimia - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   872   10
##            True      0    0
##
##                Accuracy : 0.9887
##                  95% CI : (0.9792, 0.9946)
##     No Information Rate : 0.9887
##     P-Value [Acc > NIR] : 0.583041
##
##                   Kappa : 0
```

Figura C.26: No sampling Bulimia - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   871    8
##            True      0    3
##
##                Accuracy : 0.9909
##                  95% CI : (0.9822, 0.9961)
##     No Information Rate : 0.9875
##     P-Value [Acc > NIR] : 0.23031
##
##                   Kappa : 0.4255
```

Figura C.27: No sampling Bulimia - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   870    8
##            True      1    3
##
##                Accuracy : 0.9898
##                  95% CI : (0.9807, 0.9953)
##     No Information Rate : 0.9875
##     P-Value [Acc > NIR] : 0.3391
##
##                   Kappa : 0.396
```

Figura C.28: No sampling Bulimia - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   175    9
##            True      0    0
##
##                Accuracy : 0.9511
##                  95% CI : (0.9092, 0.9774)
##     No Information Rate : 0.9511
##     P-Value [Acc > NIR] : 0.587436
##
##                   Kappa : 0
```

Figura C.29: Under sampling Bulimia - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   175    8
##            True      0    1
##
##                Accuracy : 0.9565
##                  95% CI : (0.9161, 0.981)
##     No Information Rate : 0.9511
##     P-Value [Acc > NIR] : 0.45234
##
##                   Kappa : 0.1921
```

Figura C.30: Under sampling Bulimia - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   175    3
##            True      4    2
##
##                Accuracy : 0.962
##                  95% CI : (0.9232, 0.9846)
##     No Information Rate : 0.9728
##     P-Value [Acc > NIR] : 0.8695
##
##                   Kappa : 0.3442
```

Figura C.31: Under sampling Bulimia - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   168    5
##            True      5    6
##
##                Accuracy : 0.9457
##                  95% CI : (0.9023, 0.9736)
##     No Information Rate : 0.9402
##     P-Value [Acc > NIR] : 0.4562
##
##                   Kappa : 0.5166
```

Figura C.32: Under sampling Bulimia - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   873    4
##            True      0   12
##
##                Accuracy : 0.9955
##                  95% CI : (0.9885, 0.9988)
##     No Information Rate : 0.982
##     P-Value [Acc > NIR] : 0.0003682
##
##                   Kappa : 0.8549
```

Figura C.33: Over sampling Bulimia - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   872   13
##            True      0    4
##
##                Accuracy : 0.9854
##                  95% CI : (0.9751, 0.9922)
##     No Information Rate : 0.9809
##     P-Value [Acc > NIR] : 0.1983323
##
##                   Kappa : 0.3764
```

Figura C.34: Over sampling Bulimia - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   864    5
##            True      3   17
##
##                Accuracy : 0.991
##                  95% CI : (0.9823, 0.9961)
##     No Information Rate : 0.9753
##     P-Value [Acc > NIR] : 0.0005136
##
##                   Kappa : 0.8049
```

Figura C.35: Over sampling Bulimia - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   865    5
##            True      6   13
##
##                Accuracy : 0.9876
##                  95% CI : (0.978, 0.9938)
##     No Information Rate : 0.9798
##     P-Value [Acc > NIR] : 0.05315
##
##                   Kappa : 0.6964
```

Figura C.36: Over sampling Bulimia - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.ptsd False True
##         False   864   18
##         True      0    0
##
##                Accuracy : 0.9796
##                  95% CI : (0.9679, 0.9879)
##     No Information Rate : 0.9796
##     P-Value [Acc > NIR] : 0.5622
##
##                   Kappa : 0
```

Figura C.37: No sampling Post traumatic stress disorder - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   872   10
##            True      0    0
##
##                Accuracy : 0.9887
##                  95% CI : (0.9792, 0.9946)
##     No Information Rate : 0.9887
##     P-Value [Acc > NIR] : 0.583041
##
##                   Kappa : 0
```

Figura C.38: No sampling PTSD - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.ptsd False True
##         False   869   12
##         True      0    1
##
##                Accuracy : 0.9864
##                  95% CI : (0.9764, 0.993)
##     No Information Rate : 0.9853
##     P-Value [Acc > NIR] : 0.462288
##
##                   Kappa : 0.141
```

Figura C.39: No sampling PTSD - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.ptsd False True
##         False   856   14
##         True      9    3
##
##                Accuracy : 0.9739
##                  95% CI : (0.9611, 0.9834)
##     No Information Rate : 0.9807
##     P-Value [Acc > NIR] : 0.9386 8
##
##                   Kappa : 0.194
```

Figura C.40: No sampling PTSD - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.ptsd False True
##         False   174   13
##         True      0    0
##
##                Accuracy : 0.9305
##                  95% CI : (0.8841, 0.9625)
##     No Information Rate : 0.9305
##     P-Value [Acc > NIR] : 0.5730921
##
##                   Kappa : 0
```

Figura C.41: Under sampling PTSD - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.ptsd False True
##         False   176   10
##         True      1    0
##
##                Accuracy : 0.9412
##                  95% CI : (0.8972, 0.9703)
##     No Information Rate : 0.9465
##     P-Value [Acc > NIR] : 0.69997
##
##                   Kappa : -0.0098
```

Figura C.42: Under sampling PTSD - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.ptsd False True
##         False   162    7
##         True     11    7
##
##                Accuracy : 0.9037
##                  95% CI : (0.8521, 0.9419)
##     No Information Rate : 0.9251
##     P-Value [Acc > NIR] : 0.8913
##
##                   Kappa : 0.3858
```

Figura C.43: Under sampling PTSD - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.ptsd False True
##         False   166    9
##         True      7    5
##
##                Accuracy : 0.9144
##                  95% CI : (0.8648, 0.9503)
##     No Information Rate : 0.9251
##     P-Value [Acc > NIR] : 0.7628
##
##                   Kappa : 0.3389
```

Figura C.44: Under sampling PTSD - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.ptsd False True
##         False   881    3
##         True      0    8
##
##                Accuracy : 0.9966
##                  95% CI : (0.9902, 0.9993)
##     No Information Rate : 0.9877
##     P-Value [Acc > NIR] : 0.004735
##
##                   Kappa : 0.8404
```

Figura C.45: Over sampling PTSD - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.ptsd False True
##         False   851   32
##         True      0    9
##
##                Accuracy : 0.9641
##                  95% CI : (0.9497, 0.9753)
##     No Information Rate : 0.954
##     P-Value [Acc > NIR] : 0.08335
##
##                   Kappa : 0.3492
```

Figura C.46: Over sampling PTSD - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.ptsd False True
##         False   858    9
##         True     10   15
##
##                Accuracy : 0.9787
##                  95% CI : (0.9669, 0.9871)
##     No Information Rate : 0.9731
##     P-Value [Acc > NIR] : 0.1767
##
##                   Kappa : 0.6013
```

Figura C.47: Over sampling PTSD - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.ptsd False True
##         False   854    5
##         True     11   22
##
##                Accuracy : 0.9821
##                  95% CI : (0.971, 0.9897))
##     No Information Rate : 0.9697
##     P-Value [Acc > NIR] : 0.01483
##
##                   Kappa : 0.7241
```

Figura C.48: Over sampling PTSD - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.sust_abuse False True
##               False   876    6
##               True      0    0
##
##                Accuracy : 0.9932
##                  95% CI : (0.9853, 0.9975)
##     No Information Rate : 0.9932
##     P-Value [Acc > NIR] : 0.60630
##
##                   Kappa : 0
```

Figura C.49: No sampling Substance Abuse - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.bulimia False True
##            False   872   10
##            True      0    0
##
##                Accuracy : 0.9887
##                  95% CI : (0.9792, 0.9946)
##     No Information Rate : 0.9887
##     P-Value [Acc > NIR] : 0.583041
##
##                   Kappa : 0
```

Figura C.50: No sampling Substance Abuse - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.sust_abuse False True
##               False   879    2
##               True      0    1
##
##                Accuracy : 0.9977
##                  95% CI : (0.9918, 0.9997)
##     No Information Rate : 0.9966
##     P-Value [Acc > NIR] : 0.4228
##
##                   Kappa : 0.4991
```

Figura C.51: No sampling Substance Abuse - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.sust_abuse False True
##               False   874    6
##               True      0    2
##
##                Accuracy : 0.9932
##                  95% CI : (0.9853, 0.9975)
##     No Information Rate : 0.9909
##     P-Value [Acc > NIR] : 0.31226
##
##                   Kappa : 0.3978
```

Figura C.52: No sampling Substance Abuse - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.sust_abuse False True
##               False   177    4
##               True      0    0
##
##                Accuracy : 0.9779
##                  95% CI : (0.9444, 0.9939)
##     No Information Rate : 0.9779
##     P-Value [Acc > NIR] : 0.6288
##
##                   Kappa : 0
```

Figura C.53: Under sampling Substance Abuse - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.sust_abuse False True
##               False   176    5
##               True      0    0
##
##                Accuracy : 0.9724
##                  95% CI : (0.9367, 0.991)
##     No Information Rate : 0.9724
##     P-Value [Acc > NIR] : 0.61597
##
##                   Kappa : 0
```

Figura C.54: Under sampling Substance Abuse - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.sust_abuse False True
##               False   175    3
##               True      1    2
##
##                Accuracy : 0.9779
##                  95% CI : (0.9444, 0.9939)
##     No Information Rate : 0.9724
##     P-Value [Acc > NIR] : 0.4380
##
##                   Kappa : 0.4894
```

Figura C.55: Under sampling Substance Abuse - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.sust_abuse False True
##               False   175    3
##               True      0    3
##
##                Accuracy : 0.9834
##                  95% CI : (0.9523, 0.9966)
##     No Information Rate : 0.9669
##     P-Value [Acc > NIR] : 0.1467
##
##                   Kappa : 0.6591
```

Figura C.56: Under sampling Substance Abuse - SVM polynomial

```
## Confusion Matrix and Statistics
##
##
## dat.pred.sust_abuse False True
##               False   882    0
##               True      0    4
##
##                Accuracy : 1
##                  95% CI : (0.9958, 1)    )
##     No Information Rate : 0.9955
##     P-Value [Acc > NIR] : 0.01815
##
##                   Kappa : 1
```

Figura C.57: Over sampling Substance Abuse - RF

```
## Confusion Matrix and Statistics
##
##
## dat.pred.sust_abuse False True
##               False   877    5
##               True      0    4
##
##                Accuracy : 0.9944
##                  95% CI : (0.9869, 0.9982)
##     No Information Rate : 0.9898
##     P-Value [Acc > NIR] : 0.11445
##
##                   Kappa : 0.613
```

Figura C.58: Over sampling Substance Abuse - BDT

```
## Confusion Matrix and Statistics
##
##
## dat.pred.sust_abuse False True
##               False   877    1
##               True      1    7
##
##                Accuracy : 0.9977
##                  95% CI : (0.9919, 0.9997)
##     No Information Rate : 0.991
##     P-Value [Acc > NIR] : 0.01346
##
##                   Kappa : 0.8739
```

Figura C.59: Over sampling Substance Abuse - SVM lineal

```
## Confusion Matrix and Statistics
##
##
## dat.pred.sust_abuse False True
##               False   875    2
##               True      3    6
##
##                Accuracy : 0.9944
##                  95% CI : (0.9869, 0.9982)
##     No Information Rate : 0.991
##     P-Value [Acc > NIR] : 0.19
##
##                   Kappa : 0.703
```

Figura C.60: Over sampling Substance Abuse - SVM polynomial