Universitat Oberta de Catalunya

uoc.edu

Sistema de estratificación y predicción de salud mental en trabajadores tecnológicos

Mónica Arrúe Gabarain
Máster de Ciencia de Datos
Ciencia de Datos Aplicada a la Salud

Susana Pérez Álvarez
Àngels Rius Gavidia

9 de junio de 2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Sistema de estratificación y predicción de salud mental en trabajadores tecnológicos*
Nombre del autor:	*Mónica Arrúe Gabarain*
Nombre del consultor/a:	*Susana Pérez Álvarez*
Nombre del PRA:	*Àngels Rius Gavidia*
Fecha de entrega (mm/aaaa):	06/2019
Titulación::	*Máster de Ciencia de Datos*
Área del Trabajo Final:	*Ciencia de Datos Aplicada a la Salud*
Idioma del trabajo:	*Español*
Palabras clave	*data mining; machine learning; CART; mental health*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Las enfermedades mentales son uno de los grandes problemas de salud a nivel mundial. Según la OMS, aproximadamente 450 millones de personas en el mundo sufren algún trastorno mental o de conducta [1]. Aunque cada vez se conoce más sobre este tipo de enfermedades y su origen, aún queda mucho por descubrir sobre ellas. Mediante este trabajo se busca realizar una descripción de la situación actual de la salud mental de los trabajadores tecnológicos de diferentes países, con objetivo de descubrir la prevalencia de enfermedades mentales y los principales indicadores de salud mental en este grupo de trabajadores.

Abstract (in English, 250 words or less):

Mental illness is one of the world's major health problems. According to WHO, approximately 450 million people worldwide suffer from some form of mental or behavioural disorder [1]. Although more and more is known about these types of diseases and their origins, much remains to be discovered about them. This work seeks to make a description of the current mental health situation of tech workers in different

countries, with the main objective of discovering the prevalence of mental illnesses and the main mental health indicators in this group of workers.

Índice

Lista de figuras

1. Introducción

1.1. Contexto y justificación del Trabajo

Las enfermedades mentales son un problema grave tanto para la sociedad como para los sistemas sanitarios, ya que son enfermedades complejas para las cuales, a día de hoy, no se conoce a ciencia cierta su origen [2]. El desconocimiento existente sobre estas enfermedades provoca una imprecisión considerable en los tratamientos aplicados, lo cual deriva en un gasto enorme para los sistemas sanitarios. Según un informe de la Unión Europea sobre la salud mental en el trabajo [3], solo en España el costo directo de los trastornos mentales osciló entre 150 y 372 millones de euros en 2010. Según este mismo informe, el número de días de baja por enfermedad mental este mismo año ascendieron a 2,78 millones en 2010, lo que supuso un coste de 170,96 millones de euros. Además, de las 17970 muertes registradas en 2010 relacionadas con la salud mental, 312 se las atribuye a las condiciones laborales. Esto último hace pensar que la situación laboral de las personas tiene una relación directa con su salud mental.

El resultado que se pretende obtener con la realización de este trabajo es conocer la situación actual de los trabajadores tecnológicos con respecto a la salud mental y poder detectar, entre los datos proporcionados sobre la situación y entorno laboral de las personas encuestadas, los aspectos que más influyen en su salud mental. Finalmente, se pretende visualizar los resultados obtenidos para que el usuario pueda extraer conocimiento sobre los datos de una forma sencilla e intuitiva.

1.2. Objetivos del Trabajo

El principal objetivo del trabajo es demostrar que existen ciertos factores en la vida personal y laboral de las personas que trabajan en centros tecnológicos que influyen en su salud mental. Concretamente, mediante la realización de este trabajo, se busca:

- Describir los datos de tal forma que se pueda conocer más sobre la prevalencia de enfermedades mentales en trabajadores tecnológicos
- Identificar los factores que más influyen en la salud mental de este tipo de trabajadores
- Visualizar los resultados obtenidos para poder comprenderlos de forma rápida, fácil e intuitiva

1.2.1. Hipótesis (u objetivo principal)

La hipótesis principal de este trabajo es probar si existen ciertos aspectos de la vida personal y laboral de los trabajadores tecnológicos que influyen en la salud mental de los mismos.

1.2.2. Objetivos parciales (o preguntas de investigación)

Las preguntas de investigación que a las que se pretende dar respuesta mediante este trabajo son las siguientes:

- ¿Cuál es la prevalencia de enfermedades mentales entre trabajadores tecnológicos? ¿Cuáles son las más comunes?
- ¿Existen aspectos en la vida de las personas que trabajan en entornos tecnológicos que tienen una influencia significativa en la salud mental de estas personas?
- En caso de que sí, ¿cuáles son estos aspectos y qué nivel de influencia tienen?

1.3. Enfoque y método seguido

Según Oates [4] existen seis estrategias distintas para responder a preguntas de investigación. En este caso, debido a la naturaleza del proyecto, la estrategia más adecuada a seguir es la estrategia experimental. Esta estrategia consiste principalmente en encontrar relaciones de tipo causa-efecto entre los datos, es decir, relacionar factores con consecuencias en base a los datos que se disponen.

Para llevar a cabo este trabajo de investigación, se va a emplear un conjunto de datos público formado por respuestas de trabajadores tecnológicos a un cuestionario sobre salud mental. Estos datos, en su mayoría, al ser respuestas a un cuestionario son datos cualitativos. El análisis de datos se realizará en el lenguaje programación Python [5]. En particular, se van a utilizar las librerías Pandas [6], para la manipulación de los datos, Scikit Learn [7], para llevar a cabo los algoritmos de aprendizaje automático y por último Matplotlib [8] para la visualización de los datos.

Finalmente, en cuanto a la metodología de trabajo, como se ha comentado anteriormente, se va a emplear un conjunto de datos público que recoge aproximadamente 1400 respuestas a un cuestionario realizado en 2016 [9] sobre salud mental a trabajadores tecnológicos. Las preguntas del cuestionario se detallan en el Anexo I.

1.4. Planificación del Trabajo

Para poder dar respuesta a las preguntas de investigación planteadas anteriormente, se va a llevar a cabo un flujo de trabajo clásico de análisis de datos [10]:

1. Definición de la tarea del proyecto de *data mining*

En esta primera fase se van a concretar cuáles son los objetivos que se pretenden alcanzar con el proyecto de minería de datos.

2. Selección de los datos

En un proyecto de minería de datos, una vez definidos los objetivos concretos a alcanzar, se seleccionan los datos a utilizar. En esta fase se seleccionarán de todos los datos de los que se parte, aquellos que se consideran relevantes para el objetivo del trabajo.

3. Preparación de los datos

En esta fase se van a corregir y transformar los datos con la finalidad de aumentar la calidad de los mismos y, por lo tanto, de los resultados obtenidos.

4. Construcción del modelo

Una vez que los datos se encuentren preparados, se comenzará un proceso de búsqueda del modelo que se va a emplear. En esta fase, se explorarán diferentes metodologías y se probarán con los datos para su posterior evaluación y selección del modelo que mejor se adapte a los datos.

5. Evaluación e interpretación del resultado

Una vez construido el modelo en cuestión, se procederá a evaluar los resultados obtenidos en los mismos. La evaluación del modelo se realizará empleando un conjunto de datos distinto al utilizado para construirlo (i.e. conjunto de prueba).

6. Análisis visual

Finalmente, se pretende realizar un análisis visual los resultados que tal forma que se facilite la compresión de los datos y de los resultados del modelo creado. También se pretenderá analizar la prevalencia de enfermedades mentales mediante visualizaciones de los datos.

La metodología de desarrollo que se va a emplear para llevar a cabo este proyecto es la metodología ágil CRISP-DM (*Cross Industry Standard Process for Data Mining*). Esta metodología consiste principalmente en seguir las fases explicadas anteriormente, pero plantear algunas de ellas de manera bidireccional, permitiendo en algunos puntos del proyecto revisar total o parcialmente las fases realizadas anteriormente con el objetivo de perfeccionarlas y, de esta manera, mejorar los resultados [11].

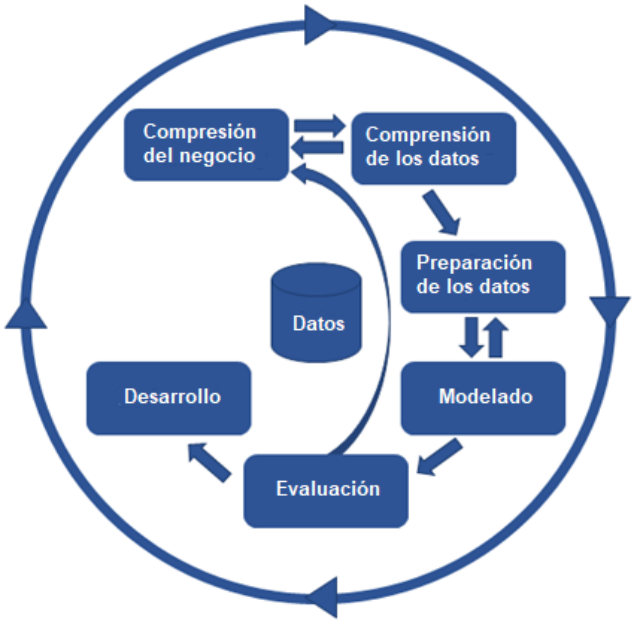


Figura 1. Esquema del Modelo CRISP-DM. Adaptación de [12].

La planificación de este TFM se ha diseñado entorno a los cinco entregables de la asignatura. A continuación, se muestra el diagrama de GANT de este TFM:

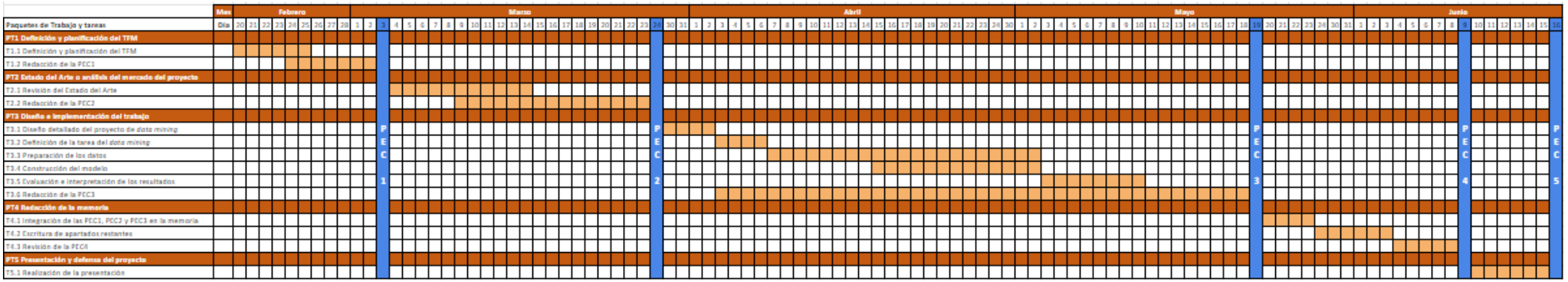


Figura 2. Diagrama de Gantt del TFM

A continuación de proporciona el link al diagrama de GANTT de mayor calidad: diagrama de GANTT.

Los hitos de este TFM, por lo tanto, serán los entregables de la propia asignatura:

Tabla 1. Hitos del TFM

Entregable	Título	Fecha de entrega
1	Definición y planificación del TFM	03/03
2	Estado del Arte o análisis del mercado del proyecto	24/03
3	Diseño e implementación del trabajo	19/05
4	Redacción de la memoria	09/06
5	Presentación y defensa del proyecto	16/06

1.5. Breve sumario de productos obtenidos

Mediante este trabajo ha realizado una descripción de la situación actual de la salud mental de los trabajadores tecnológicos de diferentes países, y se ha creado un sistema de clasificación que permite conocer los principales indicadores de salud mental en este grupo de trabajadores.

1.6. Breve descripción de los otros capítulos de la memoria

La memoria está dividida en los siguientes capítulos:

- Capítulo 2. Estado del arte

En este capítulo se exponen los aspectos más importantes con respecto al estado del arte de la ciencia de datos en la salud mental. En primer lugar, se hace una introducción a la salud mental. A continuación, se explican los avances más destacables que se han llevado a cabo en los últimos años en el campo de la ciencia de datos con respecto a la salud mental. Para finalizar, se destacan algunos estudios de identificación factores de riesgo y predicción de enfermedades mentales mediante cuestionarios.

- Capítulo 3. Diseño e implementación del trabajo

En este apartado se exponen los materiales y métodos empleados para llevar a cabo el estudio. A continuación se explica el proceso seguido para la selección y la preparación de los datos para, finalmente, presentar la construcción del modelo y su evaluación e

interpretación de los resultados. En este apartado se expone también el análisis visual de la prevalencia de enfermedades mentales como el análisis visual del propio modelo.

- Capítulo 4. Conclusiones

En este capítulo se hace un breve resumen de las contribuciones personales al proyecto y se exponen las conclusiones obtenidas acerca del mismo. Además, se hace un análisis de las posibles líneas de trabajo futuras.

- Capítulo 5. Glosario

En el capítulo 5 se ofrece una definición de los términos y acrónimos más relevantes utilizados dentro de la memoria.

- Capítulo 6. Bibliografía

En este apartado se muestra la bibliografía consultada para realización del trabajo.

- Capítulo 7. Anexos

En este capítulo se adjuntan los siguientes anexos:

- Anexo I. Encuesta OSMI 2016
- Anexo II. Clasificación CIE – 10 completa de enfermedades mentales
- Anexo III. Diccionario de variables

2. Estado del arte

2.1. Salud Mental

2.1.1. Introducción

A pesar de no tener una definición oficial, la Organización Mundial de la Salud (OMS) define la salud mental como “un estado de bienestar en el cual el individuo es consciente de sus propias capacidades, puede afrontar las tensiones normales de la vida, puede trabajar de forma productiva y fructífera y es capaz de hacer una contribución a su comunidad” [13]. El concepto de salud mental se emplea como parte de la definición de salud propuesta por la OMS: “La salud es un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades” [13].

La falta de salud mental, en sus múltiples manifestaciones, es uno de los problemas más importantes en las sociedades desarrolladas. Según la OMS, en Europa el 27% de la población adulta (18 – 65 años) ha sufrido por lo menos una vez un trastorno mental en el último año, constituyendo de esta manera, junto con los trastornos ligados al consumo de sustancias, la principal causa de discapacidad del mundo [14]. La consecuencia más devastadora de las enfermedades mentales es el suicidio. Cada año se calcula que se suicidan más de 800.000 personas en el mundo, siendo la segunda causa de muerte entre los jóvenes de 15 a 29 años. Además, hay indicios de que por cada adulto que se suicida hay 20 que lo intentan [15].

A pesar de la enorme prevalencia que existe a nivel mundial, según el Plan de acción sobre salud mental 2013-2020 de la OMS, los sistemas de salud aún no dan una respuesta adecuada a las personas con trastornos mentales. Se estima que en los países de ingresos bajos y medios, entre un 76% y un 85% de las personas con trastornos mentales graves no reciben tratamiento; la cifra es alta también en los países de ingresos elevados: entre un 35% y un 50% [16]. Además, son enfermedades complejas para las cuales, a día de hoy, no se conoce a ciencia cierta su origen [17]. El desconocimiento existente sobre estas enfermedades provoca una imprecisión considerable en los tratamientos aplicados, lo cual deriva en un gasto enorme para los sistemas sanitarios. Según un informe de la Unión Europea sobre la salud mental en el trabajo [18], solo en España el costo directo de los trastornos mentales osciló entre 150 y 372 millones de euros en 2010. Según este mismo informe, el número de días de baja por enfermedad mental este mismo año en España ascendió a los 2,78 millones en 2010, lo que supuso un coste de 170,96 millones de euros. Además, de las 17970 muertes registradas en 2010 relacionadas con la salud mental, 312 se las atribuye a las condiciones laborales, lo cual hace pensar que la situación laboral de las personas tiene una relación directa con la salud mental de las personas.

2.1.2. Clasificación de las enfermedades mentales

Existen infinidad de clasificaciones diferentes para las enfermedades mentales. Actualmente, las más empleadas por los profesionales médicos son dos:

- El capítulo V de la Clasificación Internacional de las Enfermedades (CIE - 10) creada por la OMS [19].
- El Manual diagnóstico y estadístico de los trastornos mentales (DSM - 5) producido por la Asociación Americana de Psiquiatría (*American Psychiatric Association*, APA) [20].

Debido a su internacionalidad, se explica a continuación la clasificación CIE - 10 de las enfermedades mentales. En el Anexo II se adjunta la clasificación CIE – 10 completa de las enfermedades mentales. Esta clasificación, divide las enfermedades mentales en 11 grupos principales:

- F00-F09: Trastornos mentales orgánicos, incluidos los trastornos sintomáticos
- F10-F19: Trastornos mentales y comportamiento debidos al consumo de psicotrópicos
- F20-F29: Esquizofrenia, trastornos esquizotípicos y trastornos delirantes
- F30-F39: Trastornos del humor (afectivos)
- F40-F48: Trastornos neuróticos, trastornos relacionados con el estrés y trastornos somatomorfos
- F50-F59: Síndromes del comportamiento asociados con alteraciones fisiológicas y factores físicos
- F6-F69: Trastornos de la personalidad y del comportamiento en adultos
- F70-F79: Retraso mental
- F80-F89: Trastornos del desarrollo psicológico
- F90-F98: Trastornos emocionales y del comportamiento que aparecen habitualmente en la niñez o en la adolescencia
- F99-F99: Trastornos mentales sin especificar

A continuación, se explican brevemente estos grupos de enfermedades mentales.

F00-F09: Trastornos mentales orgánicos, incluidos los trastornos sintomáticos

Esta agrupación comprende todas aquellas enfermedades mentales cuya etiología es demostrable mediante una enfermedad cerebral, una lesión cerebral o cualquier otro daño

que cause una disfunción cerebral. La demencia (F00-F03) es un tipo de enfermedad mental muy extendida que se incluye en este grupo y se caracteriza por ser una perturbación de múltiples funciones de la corte superior del cerebro, incluyendo la memoria, pensamiento, orientación, comprensión, cálculo, capacidad de aprendizaje, lenguaje y juicio.

F10-F19: Trastornos mentales y comportamiento debidos al consumo de psicotrópicos

Las enfermedades mentales que se encuentran en este grupo son todas aquellas enfermedades mentales que se atribuyen al uso de una o más sustancias psicoactivas, las cuales puede o no que hayan sido prescritas por un médico. Estas sustancias psicoactivas pueden ser, por ejemplo, el alcohol, los opioides, los cannabinoides, sustancias hipnóticas e, incluso, el tabaco.

F20-F29: Esquizofrenia, trastornos esquizotípicos y trastornos delirantes

Este grupo reúne la esquizofrenia, como principal enfermedad mental del grupo, el trastorno esquizotípico, los trastornos delirantes persistentes y un grupo más grande de trastornos psicóticos agudos y transitorios. Estas enfermedades se caracterizan principalmente por provocar una distorsión en la percepción, el pensamiento y las emociones.

F30-F39: Trastornos del humor (afectivos)

Las enfermedades de trastorno de humor son aquellas que, tal y como indica su nombre, se caracterizan por un cambio de afecto o estado de ánimo a la depresión o a la euforia (con o sin ansiedad asociada). Estos trastornos suelen ser recurrentes y el inicio de estos suele estar provocado por situaciones estresantes. Un ejemplo de enfermedad perteneciente a este grupo es el trastorno bipolar.

F40-F48: Trastornos neuróticos, trastornos relacionados con el estrés y trastornos somatomorfos

Las enfermedades incluidas en esta agrupación son aquellas provocadas por una situación estresante para el paciente y que se manifiestan como una mezcla de síntomas, generalmente como una coexistencia entre angustia y depresión. Un ejemplo de este tipo de enfermedad es la agorafobia, es decir, el miedo obsesivo a los lugares abiertos o descubiertos.

F50-F59: Síndromes del comportamiento asociados con alteraciones fisiológicas y factores físicos

Tal y como indica su nombre, esta agrupación incluye todas aquellas enfermedades mentales que están relacionadas con alteraciones fisiológicas y con factores físicos, como, por ejemplo, la anorexia o el insomnio.

F60-F69: Trastornos de la personalidad y del comportamiento en adultos

Este bloque de enfermedades incluye una variedad de enfermedades caracterizadas por un trastorno en las condiciones y patrones de comportamiento de la persona, afectando a la manera en la que tiene de relacionarse consigo mismo y con los demás. Estos patrones tienden a ser estables y a abarcar múltiples dominios de comportamiento y funcionamiento psicológico.

F70-F79: Retraso mental

Son enfermedades provocadas por un desarrollo detenido o incompleto de la mente y se caracterizan principalmente por el deterioro de funciones básicas en la época de desarrollo y que contribuyen al nivel global de la inteligencia, tales como las habilidades cognitivas, del lenguaje, motoras y sociales.

F80-F89: Trastornos del desarrollo psicológico

Las enfermedades contenidas en esta agrupación tienen en común (i) la aparición durante la infancia o niñez; (ii) el deterioro o retraso en el desarrollo de funciones que están relacionadas con la maduración biológica del sistema nervioso central; y (iii) siguen un curso constante sin remisiones ni recaídas. En la mayoría de los casos, las funciones afectadas incluyen el lenguaje, las habilidades visuoespaciales y la coordinación motora.

F90-F98: Trastornos emocionales y del comportamiento que aparecen habitualmente en la niñez o en la adolescencia

Tal y como indica el nombre de la agrupación, son enfermedades mentales caracterizadas por la aparición en la niñez o la adolescencia y que se manifiestan como trastornos en las emociones o en el comportamiento de la persona. Una de las enfermedades más comunes de este grupo es el trastorno de la actividad y de la atención.

F99-F99: Trastornos mentales sin especificar

Es una categoría residual no recomendada, a la cual se recurre cuando no se puede emplear ninguno de los códigos anteriores (F00-F98).

2.1.3. Salud mental y trabajadores tecnológicos

El trabajo juega un papel fundamental en la salud mental de las personas. Mientras que para algunas personas el trabajo es muy beneficioso para su salud mental, para otras es el origen de muchos problemas físicos y mentales. Según la OMS, algunos de los riesgos más comunes para la salud mental en el entorno laboral son [21]:

- Políticas inadecuadas de seguridad y protección de la salud
- Prácticas ineficientes de gestión y comunicación
- Escaso poder de decisión del trabajador o ausencia de control de su área de trabajo
- Bajo nivel de apoyo a los empleados
- Horarios de trabajo rígidos
- Falta de claridad en las áreas u objetivos organizativos
- Trabajo no adecuado a las competencias de la persona
- Carga de trabajo elevada
- Acoso psicológico y la intimidación en el trabajo (*mobbing*)

El Instituto Sindical de Trabajo, Ambiente y Salud (ISTAS) indica que "la salud laboral se construye en un medio ambiente de trabajo adecuado, con condiciones de trabajo justas, donde los trabajadores y trabajadoras puedan desarrollar una actividad con dignidad y donde sea posible su participación para la mejora de las condiciones de salud y seguridad" [22].

Una de las áreas laborales que más riesgo tiene es la industria tecnológica. Por lo general, los trabajadores pertenecientes a esta industria se ven sometidos a altos niveles de presión para innovar y competir constantemente. Esta industria se caracteriza, no solo porque los productos tecnológicos se encuentran en constante cambio, sino también porque las herramientas con las que se trabaja evolucionan a una velocidad enorme. Un estudio que realizó la Universidad de California en el 2015 [23] descubrió que el 49% de los emprendedores de startups tecnológicas de Estados Unidos encuestados sufrían algún tipo de enfermedad mental, donde la depresión constituía la enfermedad más común entre ellos (concretamente, el 30%). Este porcentaje es altísimo teniendo en cuenta que tan solo el 7% de las personas de Estados Unidos sufren depresión [24].

2.2. Ciencia de datos en salud mental

2.2.1. Introducción

En las dos últimas décadas ha habido un aumento enorme en el uso de la analítica de datos en multitud de disciplinas [25]. Uno de los campos en los que se ha comenzado a utilizar

la minería de datos en los últimos años es la investigación médica, principalmente en los ámbitos de la neurociencia y la biomedicina [26]. Recientemente, la psiquiatría ha comenzado a utilizar estas técnicas con el objetivo de entender mejor la composición genética de las enfermedades mentales. Por ejemplo, en [27] presentan varios algoritmos de minería de datos para analizar la expresión genética de la enfermedad del Alzheimer. Otro tipo de datos que se está empezando a utilizar para analizar las enfermedades mentales es el de las imágenes médicas. En [28], se centran en aplicar algoritmos sobre imágenes fMRI del cerebro para identificar biomarcadores que permitan predecir la enfermedad de la esquizofrenia. Por otro lado, IBM ha creado un sistema de predicción de enfermedades mentales, el cual es capaz de evaluar patrones de habla en grabaciones de personas hablando sobre sí mismos mediante técnicas de Procesamiento del Lenguaje Natural (PLN), alcanzando una precisión del 80% [29].

A pesar de haber obtenido en estos estudios resultados positivos, tanto los datos genéticos como las imágenes médicas y las grabaciones de voz tienen una desventaja importante: son datos que requieren un costo y un tiempo elevado para su recolección, además de que el procesamiento de este tipo de datos es muy complejo [26].

También existen algunos productos comerciales que aplican la Inteligencia Artificial (IA) a la salud mental. Un ejemplo de esto es el Alfred Health [30], un software que contiene un sistema de soporte a la decisión basado en *Deep Learning* cuyo objetivo principal es predecir la respuesta a tratamientos de enfermedades mentales para facilitar la toma de decisiones a los médicos. Actualmente se encuentra en fase de desarrollo, donde se están centrando únicamente en la respuesta a tratamientos para la depresión, pero su idea es escalar Alfred Health a todas las enfermedades mentales para poder ampliar la utilidad clínica. Otro ejemplo de caso de éxito es el de Quartet Health [31]; un sistema de IA capaz de detectar problemas de salud mental en su fase temprana mediante la monitorización y análisis de historias clínicas y patrones de comportamiento de los pacientes. También es capaz de recomendar un seguimiento preventivo en casos en los que los pacientes hayan recibido una mala noticia como, por ejemplo, un diagnóstico de una enfermedad física grave.

A continuación, se presentan una selección de artículos científicos enfocados en la aplicación de la minería de datos a la salud mental. Concretamente, se hace un repaso de estudios de minería de datos actuales enfocados a la identificación de factores de riesgo de enfermedades mentales y a la predicción de salud mental mediante cuestionarios. Finalmente, se presentan un artículo científico de minería de datos en el cual se ha empleado el conjunto de datos “OSMI Mental Health in Tech Survey 2016” como base del estudio.

2.2.2. Modelos de identificación de factores de riesgo y predicción de enfermedades mentales mediante cuestionarios

Existen diferentes estudios actuales que se centran en la identificación de factores de riesgo y predicción de enfermedades mentales mediante el análisis de cuestionarios.

En [32] crearon dos modelos predictivos capaces de calcular el riesgo de padecer una enfermedad mental. Los datos empleados para generar los modelos fueron respuestas a un cuestionario de 30 pacientes diferentes. Este cuestionario estaba dividido en 4 secciones principales: información demográfica (12 preguntas), factores biológicos (4 preguntas), factores psicológicos (5 preguntas) y, finalmente, factores medioambientales (5 preguntas). Para generar los modelos predictivos de enfermedad mental utilizaron, por un lado, el clasificador de Naïve Bayes y, por otro lado, un árbol de decisión generado por el algoritmo C4.5. Los modelos predictivos diseñados fueron formulados y simulados mediante el software WEKA [33]. Se observó que el árbol de decisión ofrecía una mayor precisión que el clasificador de Naïve Bayes; el modelo creado con el árbol de decisión era capaz de predecir el riesgo de salud mental con una precisión del 83,3%, mientras que el clasificador de Naïve Bayes con una precisión del 76,6%.

El autor del artículo [34] analizó los factores de riesgo de ideación e intento de suicidio entre la comunidad filipinoamericana (i.e. americanos de ascendencia filipina) mediante la técnica de *Random Forest*. Los datos empleados procedían de un estudio epidemiológico sobre la comunidad filipinoamericana, titulado *Filipino American Community Epidemiological Study (FACES)* [35], el cual contenía información sobre la salud, el consumo de alcohol, el estado anímico, los síntomas físicos, el origen cultural e información sociodemográfica de las personas. Los resultados mostraron que los predictores más importantes para la ideación de suicidio eran el padecer un trastorno depresivo, el sufrir un trastorno por abuso de sustancias y los años vividos en Estados Unidos. En cuanto a los factores más relevantes para el intento de suicidio se obtuvieron dos principales: el número de familiares y la existencia de conflictos familiares.

En [36] crearon un modelo de clasificación para enfermedades crónicas y demostró la aplicación de este método utilizando como caso de uso la enfermedad de la depresión. Los datos utilizados para la creación del modelo fueron la encuesta “U.S. Behavioral Risk Factor Surveillance System (BRFSS)”; una encuesta telefónica que se realiza cada año en Estados Unidos y que pretende recoger información sobre conductas y condiciones de salud de los ciudadanos [37]. Construyeron varios modelos de clasificación sobre estos datos, incluyendo un Árbol de Decisión C4.5, un *Random Forest*, un modelo de Perceptrón Multicapa, un clasificador de tipo Adaboost y otro mediante el algoritmo de *Support Vector Machines*. De estos modelos, se seleccionó como modelo final el Árbol de Decisión debido a la facilidad de

interpretación que ofrece. Con este modelo se obtuvo una precisión de clasificación del 80%-82% y un AUC (*Area Under the Curv*e, es decir, el área bajo la curva ROC) del 0,70-0,72. Este modelo seleccionó la variable "*childhood experience living with mentally ill*" (haber vivido con enfermos mentales en la infancia) como la variable más determinante a la hora de sufrir una enfermedad mental. Con menor relevancia, en el árbol se incluían también las variables "*Limited Usual Activity*" (actividad -física- habitual limitada) y "*Childhood sexually touched*" (tocamientos sexuales en la infancia).

Por último, en [38] se analizó el conjunto de datos "OSMI Mental Health in Tech Survey 2016" con el objetivo de identificar los principales factores predictivos que hacen que un trabajador tecnológico recurra a los servicios sanitarios de salud mental en busca de tratamiento. Para ello, se llevaron a cabo varios contrastes de hipótesis utilizando el test de Chi-cuadrado y cálculos de probabilidades mediante el cálculo de Odds Ratio (OR) para cada uno de los posibles predictores y los enfrentó contra el tratamiento aplicado. Las variables predictoras que resultaron tener una asociación significativa con el tratamiento aplicado se seleccionaron para llevar a cabo un modelo de regresión de tipo probit. Los resultados de este análisis mostraron que los predictores más importantes son (i) tener familiares con enfermedades mentales; (ii) tener interferencia laboral; (iii) disponer de programas en la empresa de cuidado de la salud mental; (iv) tener consecuencias negativas en el trabajo por tener una enfermedad mental y (v) la edad avanzada. El modelo generado obtuvo como índice AUC de 0,9.

A continuación, se muestra una tabla de resumen de los estudios analizados.

Tabla 2. Comparación de los estudios analizados

Autor(es)	Año de publicación	Método	Resultado obtenido
Idowu et al.	2018	Árbol de Decisión C4.5 y Naïve Bayes	Árbol de Decisión C4.5: 83,3% Naïve Bayes: 76,6%
Kuroki, Y.	2015	*Random Forest*	No indicado
Sunmoo YOON et al.	2014	Árbol de Decisión	Precisión clasificación del 80%-82% y AUC del 0,70-0,72
Patel, P.	2018	Modelo probit (Regresión Logística)	AUC de 0,9

3. Diseño e implementación del trabajo

3.1. Materiales y métodos

En este apartado se va a describir, en primer lugar, el conjunto de datos que se va a emplear como punto de partida del estudio. A continuación, se describen las herramientas que se van a emplear para llevar a cabo dicho estudio: lenguaje de programación, librerías y softwares.

3.1.1. Conjunto de datos

Para llevar a cabo este estudio, se va a utilizar un conjunto de datos público formado por respuestas de trabajadores tecnológicos a un cuestionario sobre salud mental. Concretamente, es un cuestionario que lleva a cabo anualmente la organización sin ánimo de lucro OSMI (*Open Sourcing Mental Ilness*) [39] con el objetivo de facilitar datos acerca de este tema para analizar la salud mental dentro de lugares de trabajo tecnológicos y la prevalencia de trastornos de salud mental dentro de la industria tecnológica. Estos datos se encuentran disponibles en Kaggle [40] y los análisis de datos llevados a cabo mediante esta plataforma los emplea esta organización para concienciar a la población sobre esta problemática y mejorar las condiciones de los trabajadores tecnológicos con trastornos mentales [9]. En concreto, los datos del cuestionario con el que se va a trabajar corresponden al cuestionario llevado a cabo en el año 2016. En el Anexo I se adjunta dicho cuestionario.

Este conjunto de datos está formado por 63 variables, donde cada una de ellas se refiere a una de las preguntas del cuestionario, y 1433 observaciones, es decir, trabajadores del ámbito tecnológico que han respondido a estas preguntas.

3.1.2. Lenguaje de programación, librerías y softwares

Para llevar a cabo este estudio, se va a utilizar el lenguaje de programación Python [5], concretamente la versión 3.7.2. El motivo por el que se ha elegido este lenguaje de programación es porque actualmente Python es uno de los lenguajes más empleados en el mundo del análisis de datos [41]. Los principales motivos por los que la comunidad científica está eligiendo este lenguaje para llevar a cabo proyectos de análisis de datos son [41]:

- Es un lenguaje fácil y rápido de aprender y entender
- Es un lenguaje de programación escalable
- Ofrece multitud de librerías enfocadas al análisis de datos
- Dispone de una variedad de opciones para la visualización de datos

En cuanto a las librerías empleadas, a continuación, se presenta una tabla con el nombre de la librería, la versión utilizada y en qué punto del estudio se emplea.

Tabla 3 . Librerías empleadas, versión y motivo

Librería	Versión	Utilidad
Pandas	0.24.2	Manipulación de datos
Scikit Learn	0.20.3	Aprendizaje automático
Matplotlib	3.0.3	Visualización de gráficas
Pydotplus	2.0.2	Visualización de árboles de decisión

El desarrollo del código se ha realizado en Jupyter Notebook [42]. Jupyter es un entorno de trabajo interactivo que permite crear código de manera dinámica... Las principales características de Juypter son las siguientes [43]:

- Posee una avanzada interfaz web que permite combinar código fuente, textos, fórmulas, figuras y multimedia en un único documento.
- Aunque el lenguaje de programación fundamental en Jupyter Notebook es Python, esta aplicación también es compatible con más de 40 lenguajes.
- Los documentos realizados en Jupyter Notebook se pueden exportar a diferentes formatos estáticos como, por ejemplo, HTML.
- Permite el acceso desde cualquier lugar sin necesidad de instalación de otros servicios, ya que funciona como cliente servidor.

3.1.3. Modelo de minería de datos

3.1.3.1. Introducción

En el ámbito del aprendizaje automático existen principalmente dos tipos de algoritmos: los algoritmos de aprendizaje supervisado y los algoritmos de aprendizaje no supervisado. Por un lado, el aprendizaje supervisado se emplea cuando los datos de los que se parte están formados por variables de entrada (x) y una variable de salida o resultado (Y). Cuando se tiene este tipo de variable en los datos, normalmente se dice que los datos se encuentran etiquetados. Este tipo de algoritmo se utiliza con el objetivo de crear una función que indique la relación entre la variable de salida y las diferentes variables de entrada [44].

Es por esto, que en estadística a las variables de entrada se les suele denominar variables independientes y a la variable de salida, la variable dependiente.

Dependiendo de la tipología de la variable de salida, los problemas de aprendizaje supervisado se pueden dividir a su vez en dos tipos: problemas de clasificación -cuando la variable de salida es una categoría- y de regresión -cuando la variable de salida de tipo numérico-.

Por otro lado, el aprendizaje no supervisado se da cuando únicamente se dispone de variables de entrada (x) en el conjunto de datos y no variables de salida. Es decir, los datos no se encuentran etiquetados. El objetivo principal de los algoritmos de aprendizaje no supervisado es modelar la distribución de los datos, por lo que tiene un carácter exploratorio. Los problemas de aprendizaje no supervisado se pueden dividir en grupos: problemas de *clustering* -permite descubrir agrupaciones de registros entre los datos- y de asociación -cuando el objetivo es identificar reglas que describen los datos- [45].

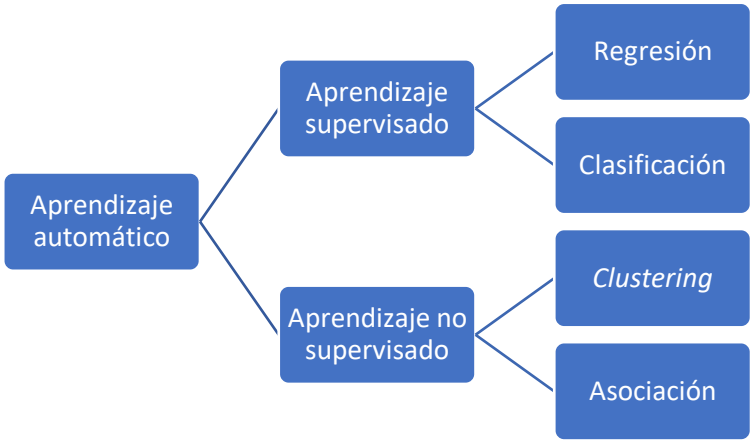


Figura 3. Tipos de aprendizaje automático

3.1.3.2. Aprendizaje supervisado: problemas de clasificación

Dada la naturaleza de los datos de partida y el objetivo del estudio, el método de aprendizaje automático más conveniente es el de tipo supervisado, ya que disponemos en el conjunto de datos la variable que queremos predecir: si el trabajador ha sido diagnosticado de alguna enfermedad mental por un profesional. Concretamente, se trata de un problema de clasificación binaria, ya que la variable respuesta puede tomar únicamente los valores “Yes” y “No”.

Existen multitud de métodos de clasificación en el aprendizaje supervisado. A continuación, se proporciona una breve introducción a algunos de los algoritmos de clasificación más comunes [44], [46], [47].

Regresión logística

Tal y como se ha explicado anteriormente, la regresión se emplea para tareas de aprendizaje supervisado cuando la variable dependiente es de tipo numérico. Existen diferentes tipos de regresión. Uno de los criterios de clasificación de los algoritmos de regresión es según el tipo de función f(X) generada mediante el algoritmo; regresión lineal, si la función f(X) es lineal y regresión no lineal, si la f(X) es no lineal. La regresión logística es el equivalente de la regresión lineal para problemas de clasificación. La idea detrás de la regresión logística es la de encontrar la relación entre las variables de entrada y la probabilidad de pertenecer a una clase. Mediante este algoritmo, se pretende encontrar la función logística que modela la probabilidad de que una observación pertenezca a una clase o a la otra.

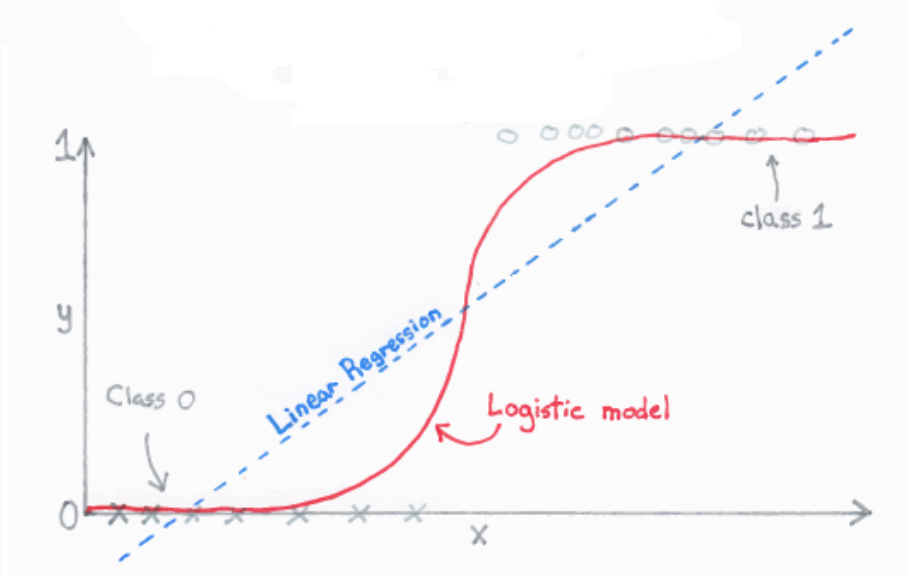


Figura 4 . Representación gráfica de la regresión logística [48]

Naive Bayes

El algoritmo de Naive Bayes se basa en el teorema de Bayes asumiendo la independencia entre las variables predictoras y que la contribución de las variables predictoras es igual. El teorema de Bayes proporciona una manera de calcular la probabilidad posterior P(c|x) teniendo en cuenta P(c), P(x) y P(x|c). El teorema de Bayes se puede resumir pendiente la siguiente fórmula:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

donde:

- P(c|x): probabilidad posterior de una clase (c) dado una variable (x). Esto es lo que se busca calcular.
- P(x|c): probabilidad posterior de la variable predictora dada la clase.
- P(c): probabilidad previa de la clase.
- P(x): probabilidad previa de la variable predictora.

Aunque el algoritmo de clasificación de Naive Bayes se basa en este teorema, el algoritmo de clasificación es mucho más complejo.

K vecinos más cercanos

Este algoritmo separa los elementos del conjunto de datos en grupos según la distancia que los separa y clasifica cada dato nuevo en el grupo que le corresponde según tenga k vecinos más cerca de un grupo o de otro. Es un algoritmo no paramétrico y de aprendizaje vago (i.e. el uso de los datos de entrenamiento se pospone hasta el momento en el que se hace la consulta, no tiene tiempo de entrenamiento pero requiere más tiempo en la predicción), y se puede emplear también en tareas que requieran técnicas de regresión.

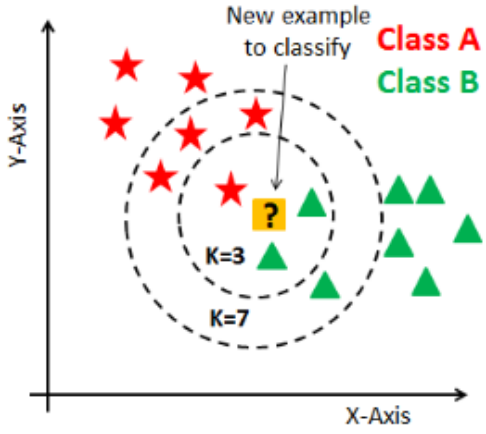


Figura 5. Representación del algoritmo K vecinos más cercanos [49]

Árbol de decisión

El algoritmo de árbol de decisión construye modelos de clasificación o regresión en forma de árbol. El algoritmo descompone el conjunto de datos en subconjuntos más pequeños en base a ciertas características creando una estructura de decisión en base a las variables más relevantes para la clasificación de resultados. Una de sus mayores ventajas de este algoritmo es la interpretabilidad.

Existe también un algoritmo denominado bosques aleatorios (*Random Forest*, en inglés) que selecciona de manera aleatoria una cantidad de variables con las cuales se construyen árboles de decisión individuales y los fusiona para obtener una predicción más precisa y estable.

Máquinas de vectores de soporte (*Support Vector Machine*)

Es un algoritmo de aprendizaje supervisado que se utiliza tanto en problemas de regresión como de clasificación. Se basa en el concepto de planos de decisión (hiperplanos) que separan los objetos pertenecientes a diferentes clases.

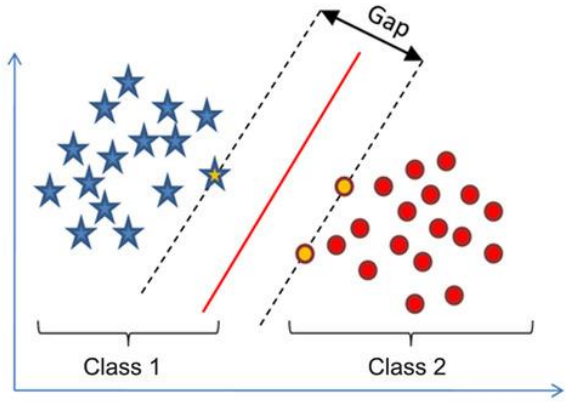


Figura 6. Representación del algoritmo Support Vector Machine [50]

3.1.3.3. Modelo seleccionado

Tras analizar los algoritmos más comunes de aprendizaje supervisado para la clasificación, se ha seleccionado el árbol de decisión como modelo para llevar a cabo este estudio. Los motivos principales por los que se ha elegido este algoritmo son los siguientes [51]:

- Ofrece una alta interpretabilidad y los resultados son fáciles de entender, lo cual es muy interesante en este caso porque un posible usuario final de esta herramienta puede ser una persona con un perfil no técnico (e.g. psicólogos, salud pública, recursos humanos, etc.).
- Es muy robusto frente a los valores atípicos u *outliers*, en inglés.
- No requiere un preprocesado exhaustivo de los datos.
- El coste de predicción de los datos sigue una función logarítmica con respecto al número de variables necesarias para entrenar el modelo, lo cual va a proporcionar un coste de predicción menor que otros modelos.

Existen diferentes métodos de construcción de árboles de decisión, siendo los más comunes ID3, C4.5, CART (*Classification And Regression Trees*), CHAID (*Chi Square Automatic Interaction Detector*) y LMDT (*Linear Machine Decision Tree)* [52]. De estos algoritmos se ha seleccionado CART para llevar a cabo el estudio principalmente por los siguientes dos motivos:

- CART puede utilizar las mismas variables en diferentes partes del árbol, lo cual permite revelar interdependencias complejas entre las variables. Esto es muy interesante en este caso ya que las variables del conjunto de datos de partida, al recoger datos personales y laborales de una misma persona, pueden estar muy relacionadas entre sí [53].
- CART no requiere una selección previa de variables, sino que identifica las variables más significativas del conjunto de datos por sí mismo y descarta las que lo son menos [54].

Como se ha comentado anteriormente, una de las principales ventajas del modelo de árbol de decisión es su fácil entendimiento y su interpretabilidad, ya que permite mostrar el modelo en forma de árbol de tal forma que se puede fácilmente extraer conocimiento de él. Este tipo de diagramas tienen una estructura de árbol similar a un diagrama de flujo donde cada uno de sus nodos representa una variable y cada rama una regla de decisión sobre esa variable. Estos nodos, siguiendo estas reglas de decisión dividen los datos en diferentes subconjuntos [55]. En la representación del árbol de decisión de la librería que se va a emplear, cada uno de los nodos muestra la siguiente información:

- **Criterio** (i.e. variable y valor de la variable) por el cual se está dividiendo el conjunto de datos en dos.
- **Gini**: índice de Gini. Mide el grado de “pureza” del nodo con respecto a las clases, es decir, la proporción de datos de cada clase que se encuentra en el conjunto de datos del nodo. Cuanto mayor es el índice de Gini menor es la pureza del nodo. La intensidad de los colores de los nodos se determina siguiendo el valor de índice de Gini; cuando más intenso es el color, menor es el índice de Gini y, por lo tanto, mayor es la pureza del nodo. Por el contrario, a menor intensidad del color del nodo, mayor es el índice de Gini y menor es su pureza.
- **Samples**: número de observaciones que se encuentran en el nodo.
- **Value**: probabilidades de cada categoría. Es decir, del número de observaciones (i.e. **samples**) que se encuentran en el nodo, cuantas pertenecen a cada una de las categorías.

- **Class**: muestra la predicción que el nodo va a realizar en ese punto. La clase se puede obtener del listado de valores (i.e. **Value**); el valor de la clase que más se da entre las observaciones de ese nodo se elige como clase del nodo. Los nodos se colorean con un color u otro según este criterio.

A pesar de las ventajas de este modelo, existen también ciertas desventajas asociadas a este tipo de modelo. La principal desventaja que presenta es la facilidad que tiene de realizar un sobreajuste de los datos (*overfitting,* en inglés). Mediante este término se hace referencia a la incapacidad de un modelo de aprendizaje automático de generalizar correctamente. Un algoritmo de aprendizaje automático debe ser capaz de predecir resultados de un conjunto de datos no vistos anteriormente. Sin embargo, cuando un modelo se sobreentrena, el algoritmo se ajusta perfectamente a las características de los datos de entrenamiento, pero es incapaz de predecir correctamente un conjunto de datos no visto previamente.

Existen principalmente dos técnicas para evitar el sobreajuste de los datos en los árboles de decisión [56]. En primer lugar, la denominada técnica de *pre-pruning* consiste en parar el crecimiento del árbol antes de que haya crecido completamente. Por otro lado, la técnica de *post-pruning* permite al árbol que crezca en su totalidad sin ninguna restricción de tamaño para, una vez completado, podarlo [57]. La poda consiste en reducir el tamaño del árbol de decisión mediante la eliminación de hojas que proporcionan poca capacidad de clasificar instancias [58].

Aunque la librería que se va a emplear para la generación de los árboles de decisión no permite aplicar estas técnicas como tal, se puede evitar el sobreajuste de los datos modificando el número mínimo de muestras requeridas para que un nodo sea un nodo final y la profundidad máxima del árbol [51]. De hecho, en la propia documentación de los árboles de decisión (*DecisionTreeClassifier*) de la librería se ofrecen una serie de consejos para evitar que esto ocurra. Los principales consejos que se ofrecen son [51]:

1. Tener una proporción correcta entre las observaciones y el número de variables, ya que este tipo de algoritmo tiende a sobreajustar los datos si dispone de pocas muestras en un espacio de dimensionalidad elevada.
2. Limitar la profundidad del árbol (parámetro *max_depth)*, lo cual para el crecimiento del árbol en el punto especificado evitando que el sobreajuste del modelo. Se recomienda comenzar limitando la profundidad del árbol a 3 y aumentarla para ver cómo se comporta el árbol.
3. Usar los parámetros de la función del modelo que indican el número mínimo de muestras para dividir un nodo interno del árbol (parámetro *min_samples_split)* y el del

número mínimo de muestras requeridas para que un nodo sea un nodo final o también conocido como nodo hoja (parámetro *min_samples_leaf*). El efecto generalmente que causa en el árbol el aumento de estos dos parámetros es que puede aumentar el índice de Gini en cada nodo (i.e. disminuir su pureza) ya que tiene que considerar más muestras en cada uno de los nodos. Pero al tener que considerar más muestras es capaz de generalizar mejor. Normalmente en modelos de clasificación se suelen obtener buenos resultados estipulando el parámetro *min_samples_leaf* a 1 aunque se recomienda comenzar con *min_samples_leaf* igual a 5.

Además, tal y como se explica en la documentación de esta función [51], los datos que se empleen para crear el modelo deben cumplir dos condiciones:

1. No deben tener *missing values.*
2. No admite variables categóricas directamente, sino que tienen que ser codificadas antes de su uso.

Los resultados obtenidos en los modelos generados de árboles de decisión se van a contrastar con modelos generados mediante *Random Forest*. Este algoritmo, tal y como se ha explicado anteriormente, crea múltiples de árboles de decisión y los fusiona para mejorar la precisión del modelo. Los motivos por los cuales se ha decidido complementar los resultados obtenidos con los árboles de decisión con el método de *Random Forest* es porque este método es una versión avanzada de los árboles de decisión y, puesto que combinan los resultados obtenidos con más de un árbol, normalmente tiende a ser más preciso [59].

Al igual que en el caso de los árboles de decisión, se va a tratar de crear el modelo más adecuado mediante el ajuste de parámetros. Para crear los modelos se va a emplear la función *RandomForest* de Scikit Learn. Los criterios que se van a seguir para el ajuste de parámetros son los siguientes [60], [61]:

- Los principales parámetros a modificar deben ser el número de estimadores (parámetro *n_estimators)*, es decir, el número de árboles del bosque y el tamaño de los subconjuntos aleatorios a considerar cuando se divide un nodo (parámetro *max_features*). Cuanto mayor es el primer parámetro, mejores resultados se obtienen ya que tiene en cuenta los resultados de más árboles, pero el coste computacional aumenta también. En cuanto al segundo, cuanto menor es este parámetro, mayor es la reducción de la varianza pero también aumenta el sesgo.
- Normalmente se obtienen buenos resultados estipulando el tamaño de los subconjuntos (parámetro *max_features*) a la raíz cuadrada del número de clases. En este caso, como el número de clases a predecir es igual a dos (“Yes” / “No”), se va a

probar a estipular este parámetro a la raíz cuadrada de 2. Este es el valor predeterminado que presenta la función de la librería que se va a emplear, por lo que no se va a modificar.

- También se suelen obtener buenos resultados no limitando la profundidad máxima del árbol (parámetro *max_depth)* y estipulando el parámetro que hace referencia al número mínimo de muestras requeridas para dividir un nodo interno a 2 (parámetro *min_samples_split)*. Este es el valor predeterminado que presenta la función, por lo que tampoco se va a modificar.

La base de la medida de la calidad de un modelo de clasificación es saber si el modelo etiquetará correctamente objetos nuevos [62]. Para ello, los modelos se van a evaluar utilizando la medida de la precisión del modelo, es decir, la proporción de éxitos (i.e. operaciones de clasificación correctas) que ha sido capaz de obtener el modelo. Esta precisión se calcula de la siguiente manera [62]:

$$Precisión = \frac{Nº\ casos\ de\ éxito}{Nº\ casos\ totales}$$

Para realizar esta evaluación, se va a dividir el conjunto de datos en dos subconjuntos: el de entrenamiento, con el cual se construirá el modelo de clasificación, y el de prueba, el cual se empleará para calcular la precisión del modelo. Normalmente, cuanto mayor es el conjunto de entrenamiento, mejor es el clasificador y cuanto mayor es el conjunto de prueba, más cerca de la realidad se encuentra la precisión que obtenemos [62]. La proporción entre el conjunto de datos de entrenamiento y test que se ha seleccionado va a ser una de las más comunes empleadas en algoritmos de aprendizaje automático [63]; 80:20.

3.1.3.4. Alternativa a la librería Scikit Learn

El hecho de que la librería Scikit Learn no permita trabajar variables categóricas directamente es una limitación muy importante. El mayor problema de esto es que la codificación de las variables categóricas, bien por empleando la técnica de *Label Encoder* o la de *One Hot Encoding*, puede llegar a dificultar enormemente la interpretabilidad del árbol de decisión.

Al buscar una alternativa para esta librería, se ha detectado una falta de librerías en Python que permitan crear árboles de decisión con variables categóricas. La mayoría de las

librerías encontradas o bien no admiten datos categóricos directamente como ocurre con Scikit Learn o bien aceptan datos categóricos pero la propia función de codificación los transforma a numéricos.

Sin embargo, tras detectar esta limitación, se ha encontrado una librería que sí que permite crear árboles de decisión con variables categóricas directamente. Esta librería se llama Orange [64] y además de ser una librería de Python es un programa informático que permite llevar a cabo procesos completos de análisis de datos sin tener que programar. De hecho, lo que la librería Orange permite es manipular los componentes de este programa utilizando Python. Cabe destacar que la última versión de esta librería (la versión 3), la cual utiliza Python 3, no tiene implementada la función de graficar árboles de decisión, por lo que en caso de que se quisiese emplear esta librería para mostrar árboles de decisión se tendría que utilizar la versión 2, la cual utiliza Python 2.

Por lo tanto, además de crear los modelos mediante la librería Scikit Learn, se van a crear los mismos modelos utilizando el software Orange y se van a contrastar los resultados obtenidos.

3.2. Selección de los datos

Analizando las preguntas del cuestionario, se puede observar que existen multitud de preguntas de diferente índole. Muchas de las respuestas no aportan información relevante o son subjetivas (i.e. dos personas en la misma situación puede que respondan a la misma pregunta de forma distinta), por lo que no se van a emplear en el modelo. Los criterios por los que se van a descartar variables son los siguientes:

- Respuestas al cuestionario que sean opiniones o sentimientos.
- Respuestas al cuestionario que sean texto libre (menos la variable del sexo. A continuación se detallará el motivo).
- Respuestas al cuestionario que puedan tomar más de un valor.
- Repuestas al cuestionario que, por su contexto y contenido, se considera que no pueden ser predictoras de enfermedades mentales (e.g. el tipo de baja que elegiría el trabajador en caso de que se le diagnosticase una enfermedad mental no tiene carácter predictivo).

A continuación, se muestra una tabla con el listado de las respuestas al cuestionario seleccionadas a priori y si se va a emplear en el estudio o no.

Tabla 4. Selección de variables para el estudio

Nº	Pregunta del cuestionario	¿Incluida en el análisis?	Motivo de exclusión
1	Are you self-employed?	Sí	
2	How many employees does your company or organization have?	Sí	
3	Is your employer primarily a tech company/organization?	Sí	
4	Is your primary role within your company related to tech/IT?	Sí	
5	Does your employer provide mental health benefits as part of healthcare coverage?	Sí	
6	Do you know the options for mental health care available under your employer-provided coverage?	Sí	
7	Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?	Sí	
8	Does your employer offer resources to learn more about mental health concerns and options for seeking help?	Sí	
9	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?	Sí	
10	If a mental health issue prompted you to request a medical leave from work, asking for that leave would be:	No	Se considera que no tiene carácter predictivo
11	Do you think that discussing a mental health disorder with your employer would have negative consequences?	No	Es una opinión
12	Do you think that discussing a physical health issue with your employer would have negative consequences?	No	Es una opinión
13	Would you feel comfortable discussing a mental health disorder with your coworkers?	No	Es un sentimiento
14	Would you feel comfortable discussing a mental health disorder with your direct supervisor(s)?	No	Es un sentimiento
15	Do you feel that your employer takes mental health as seriously as physical health?	No	Es un sentimiento
16	Have you heard of or observed negative consequences for co-workers who have been open about mental health issues in your workplace?	No	Se considera que no tiene carácter predictivo
17	Do you have medical coverage (private insurance or state-provided) which includes treatment of mental health issues?	Sí	
18	Do you know local or online resources to seek help for a mental health disorder?	Sí	
19	If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to clients or business contacts?	No	Se considera que no tiene carácter predictivo

20	If you have revealed a mental health issue to a client or business contact, do you believe this has impacted you negatively?	No	Es una opinión
21	If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to coworkers or employees?	No	Se considera que no tiene carácter predictivo
22	If you have revealed a mental health issue to a coworker or employee, do you believe this has impacted you negatively?	No	Es una opinión
23	Do you believe your productivity is ever affected by a mental health issue?	No	Es una opinión
24	If yes, what percentage of your work time (time performing primary or secondary job functions) is affected by a mental health issue?	No	Se considera que no tiene carácter predictivo
25	Do you have previous employers?	Sí	
26	Have your previous employers provided mental health benefits?	Sí	
27	Were you aware of the options for mental health care provided by your previous employers?	Sí	
28	Did your previous employers ever formally discuss mental health (as part of a wellness campaign or other official communication)?	Sí	
29	Did your previous employers provide resources to learn more about mental health issues and how to seek help?	Sí	
30	Was your anonymity protected if you chose to take advantage of mental health or substance abuse treatment resources with previous employers?	Sí	
31	Do you think that discussing a mental health disorder with previous employers would have negative consequences?	No	Es una opinión
32	Do you think that discussing a physical health issue with previous employers would have negative consequences?	No	Es una opinión
33	Would you have been willing to discuss a mental health issue with your previous co-workers?	No	Se considera que no tiene carácter predictivo
34	Would you have been willing to discuss a mental health issue with your direct supervisor(s)?	No	Se considera que no tiene carácter predictivo
35	Did you feel that your previous employers took mental health as seriously as physical health?	No	Es un sentimiento
36	Did you hear of or observe negative consequences for co-workers with mental health issues in your previous workplaces?	No	Se considera que no tiene carácter predictivo
37	Would you be willing to bring up a physical health issue with a potential employer in an interview?	No	Se considera que no tiene carácter predictivo
38	Why or why not?	No	Texto libre

39	Would you bring up a mental health issue with a potential employer in an interview?	No	Se considera que no tiene carácter predictivo
40	Why or why not?	No	Texto libre
41	Do you feel that being identified as a person with a mental health issue would hurt your career?	No	Es un sentimiento
42	Do you think that team members/co-workers would view you more negatively if they knew you suffered from a mental health issue?	No	Es una opinión
43	How willing would you be to share with friends and family that you have a mental illness?	No	Se considera que no tiene carácter predictivo
44	Have you observed or experienced an unsupportive or badly handled response to a mental health issue in your current or previous workplace?	No	Se considera que no tiene carácter predictivo
45	Have your observations of how another individual who discussed a mental health disorder made you less likely to reveal a mental health issue yourself in your current workplace?	No	Se considera que no tiene carácter predictivo
46	Do you have a family history of mental illness?	Sí	
47	Have you had a mental health disorder in the past?	Sí	
48	Do you currently have a mental health disorder?	No	Es más bien una opinión. Es mejor utilizar la variable de diagnóstico por un profesional.
49	If yes, what condition(s) have you been diagnosed with?	No	Puede tomar más de un valor
50	If maybe, what condition(s) do you believe you have?	No	Es una opinión
51	Have you been diagnosed with a mental health condition by a medical professional?	Sí	
52	If so, what condition(s) were you diagnosed with?	No	Puede tomar más de un valor
53	Have you ever sought treatment for a mental health issue from a mental health professional?	No	Se considera que no tiene carácter predictivo
54	If you have a mental health issue, do you feel that it interferes with your work when being treated effectively?	No	Es una opinión
55	If you have a mental health issue, do you feel that it interferes with your work when NOT being treated effectively?	No	Es una opinión
56	What is your age?	Sí	
57	What is your gender?	Sí	
58	What country do you live in?	Sí	
59	What US state or territory do you live in?	Sí	
60	What country do you work in?	Sí	
61	What US state or territory do you work in?	Sí	

62	Which of the following best describes your work position?	No	Puede tomar más de un valor
63	Do you work remotely?	Sí	

3.3. Preparación de los datos

En esta fase se van a preprocesar los datos con la finalidad de, por un lado, aumentar su calidad y, por lo tanto, de los resultados obtenidos en el estudio (e.g. corrección de valores erróneos, tratamiento de datos faltantes, etc.). Por otro lado, se van a transformar los datos para que se encuentren en el formato adecuado para poder emplearlos en el modelo. Para poder elegir los tipos de correcciones y transformaciones requeridas es necesario conocer los datos de los que se parte. Para ello, se va a llevar a cabo un análisis exploratorio de las variables o *Exploratory Data Analysis* (EDA) en inglés. El análisis exploratorio de las variables en un proceso crítico en cualquier análisis de datos ya que permite explorar los datos con los que se va a trabajar con el objetivo de familiarizarse con ellos, descubrir patrones, encontrar anomalías, entre otras cosas, con la ayuda de resúmenes estadísticos y representaciones gráficas [65].

A pesar de que este análisis se realice antes -y también después- de llevar a cabo el preprocesamiento de los datos para identificar las correcciones pertinentes, en la memoria se va a mostrar únicamente este análisis con los datos preprocesados ya que son los que definitivamente se van a emplear en el modelo y, por lo tanto, los que puedan ser de interés para el lector.

3.3.1. Corrección de datos y tratamiento de *missing values*

La corrección de datos consiste en procesar los datos para eliminar aquellos que sean erróneos o redundantes [10]. Este es un paso fundamental en el proceso de la analítica de los datos puesto que este tipo de datos pueden suponer un gran problema no solo en el desarrollo del estudio sino también en los resultados obtenidos, dando lugar a conclusiones erróneas.

En este caso, las variables que requieren una corrección son las siguientes:

- What is your gender?

  Esta variable toma valores muy distintos, puesto que es una respuesta de tipo texto libre. Existen algunas respuestas distintas que se refieren a lo mismo (e.g. “Male” y “M”), por lo que se van a homogeneizar este tipo de respuestas en “Male” y “Female”. Por otro lado, todas aquellas respuestas distintas de hombre y mujer se van a agrupar bajo la categoría “Others”.

- What is your age?

  Esta variable puede tomar los siguientes valores: [3, 15, 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 61, 62, 63, 65, 66, 70, 74, 99, 323]. Los valores 3, 99 y 323 se deben lógicamente a un error, puesto que es imposible que existan trabajadores con esas edades. Puesto que la función que se va a emplear para crear el modelo no admite *missing values* en el conjunto de datos y tan solo hay una observación por cada uno de estos valores, se van a excluir a estos tres trabajadores del estudio.

Como se ha podido observar en el análisis exploratorio de las variables, existen bastantes variables que tienen datos faltantes (o *missing values,* en inglés). Se van a tratar los datos de tal forma que no haya ningún dato faltante en el conjunto de datos ya que el modelo que se va a emplear lo requiere. En cada uno de los casos se va a valorar si la mejor opción es eliminar las observaciones o la imputación de los valores. A continuación, se muestra el listado de variables con su correspondiente número de datos faltantes:

self_employed	0
n_employees	287
tech_company	287
tech_it_role	1170
employer_mental_health_benefits	287
employer_mental_health_options	420
employer_discussed_mental_health	287
employer_offer_resources	287
employer_anonymity_protected	287
medical_coverage	1146
know_local_or_online_resources	1146
previous_employers	0
previous_employers_mental_health_benefits	169
previous_employers_aware_mental_health_options	169
previous_employer_discussed_mental_health	169
previous_employers_provide_resources	169
previous_employer_anonymity_protected	169
family_history_mental_illness	0
past_mental_health_disorder	0
diagnosed_by_medical_professional	0
age	0
gender	0
country	0
us_state	593
country_work	0
us_state_work	582
work_remotely	0

Figura 7. Missing values de las variables

En primer lugar, se puede observar que algunas variables tienen exactamente 287 datos faltantes. Revisando las preguntas del cuestionario a las que pertenecen dichas respuestas se llega a la conclusión de que se tratan de respuestas al cuestionario de trabajadores autónomos a preguntas sobre la empresa en la que trabaja (e.g. si la empresa donde trabaja ofrece ayudas para cuidar la salud mental). Para evitar que haya datos faltantes en estas variables se van a rellenar estos *missing values* con un "SE" (*Self Employed*) para indicar de que se trata de un trabajador autónomo.

Una vez corregida la situación de estas observaciones, se aprecia que existen 5 variables con relación a trabajos anteriores que tienen exactamente 169 datos incompletos. Estos datos incompletos corresponden a los trabajadores que no han tenido previamente un trabajo. Al igual que en el caso anterior, estos *missing values* se van a rellenar con un "NPE" (*No Previous Employer*) para los trabajadores que no han tenido previamente un trabajo.

Hasta ahora, se tiene un conjunto de datos con 1420 observaciones (i.e. trabajadores) y 27 variables (i.e. preguntas del cuestionario). Llegados a este punto, se detectan dos variables que no tienen prácticamente ningún dato (1143 datos faltantes de 1420 observaciones). Estas variables son las correspondientes a las preguntas del cuestionario "Do you have medical coverage (private insurance or state-provided) which includes treatment of mental health issues" y "Do you know local or online resources to seek help for a mental health disorder?". Debido al elevado número de datos faltantes que contienen estas variables se van a eliminar del conjunto de datos.

Por otro lado, puesto que el enfoque del estudio es global, no es necesario que el participante viva en Estados Unidos ni que trabaje ahí, por lo que las variables "What US state or territory do you live in?" y "What US state or territory do you work in?" se van a eliminar también.

En este punto, las variables que tienen algún dato faltante son las correspondientes a las siguientes preguntas del cuestionario:

- "Is your primary role within your company related to tech/IT?"
- "Do you know the options for mental health care available under your employer-provided coverage?"

Con respecto a la pregunta "Is your primary role within your company related to tech/IT?", el porcentaje de datos faltantes de esta variable es muy elevado; 1167 de 1430 observaciones, es decir, casi el 82% de los registros no tienen un valor. Por este motivo, no se va a tener en cuenta esta variable en el modelo.

Finalmente, con respecto a la variable de la pregunta del cuestionario "Do you know the options for mental health care available under your employer-provided coverage?" puesto

que el porcentaje de datos faltantes es bastante bajo (9,3%) se van a eliminar estas observaciones del conjunto de datos.

Por lo tanto, el conjunto de datos resultante no tiene ningún dato faltante y está formado por 1297 observaciones y 22 variables.

3.3.2. Análisis exploratorio de las variables

En este apartado se va a mostrar un análisis exploratorio de las variables del conjunto de datos tras la corrección de los datos y el tratamiento de los *missing* values.

El análisis exploratorio será distinto para las variables numéricas y para las categóricas o dicotómicas. En el caso de las variables numéricas (en este caso, únicamente se tiene la edad) se va a mostrar una descripción general de los valores numéricos que toma, para a continuación, mostrar su histograma y diagrama de caja. En el caso de las variables categóricas y dicotómicas se van a mostrar sus posibles valores y su diagrama de barras. En el Anexo III se ofrece un diccionario con los nombres de las variables y sus correspondientes preguntas del cuestionario para facilitar la comprensión del análisis.

- Variables numéricas (edad)

A continuación, se muestra una descripción sobre los valores que toma la variable de la edad. En esta tabla se ofrece la siguiente información:

- count: cantidad de valores excluyendo los valores vacíos
- mean: media de los valores
- std: desviación estándar de los valores
- min: valor mínimo
- 25%: percertil 25 de los valores
- 50%: percertil 25 o mediana de los valores
- 75%: percertil 75 de los valores
- max: valor máximo

Tabla 5 . Descripción de la variable edad

Indicador	Valor
count	1297
mean	34,3
std	8,2
min	15
25%	29
50%	33
75%	39
max	74

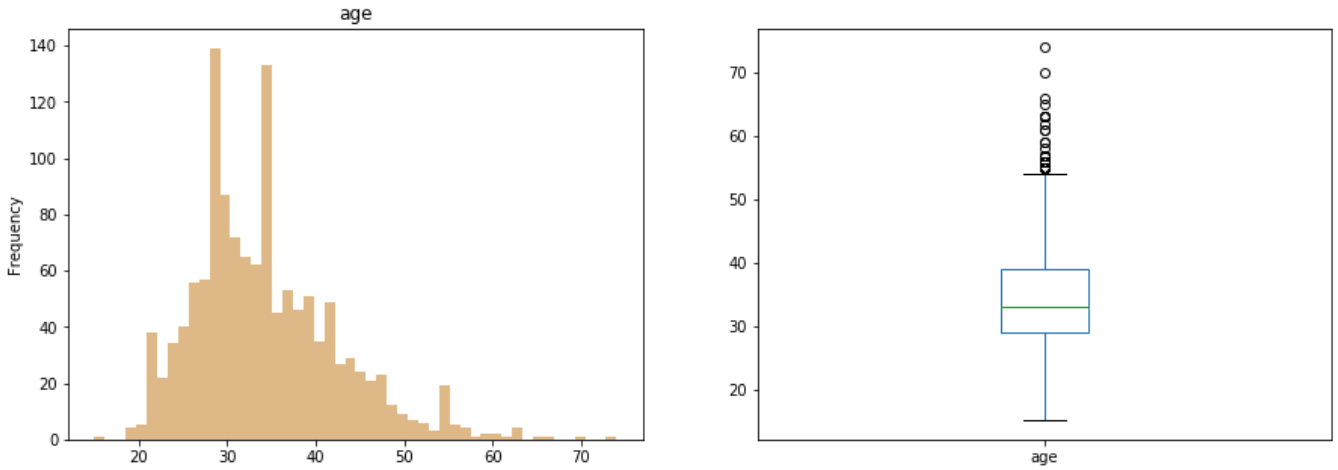


Figura 8. Diagrama de barras y boxplot de la variable "edad"

A continuación, se muestra una tabla donde se resumen las posibles categorías de las variables categóricas:

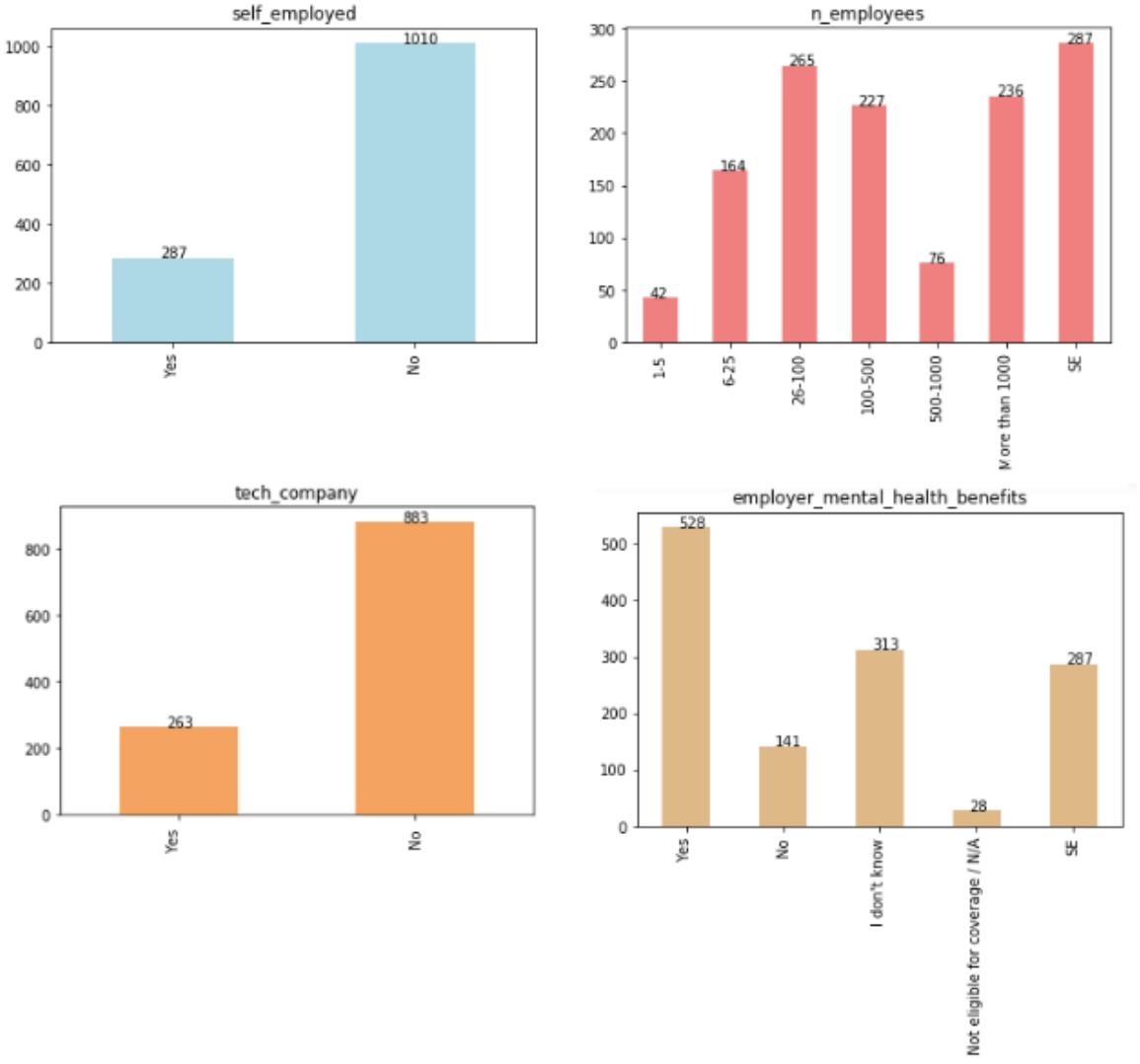
Tabla 6. Posibles categorías de las variables categóricas

Nombre de la variable	Categorías
self_employed	“Yes”, “No”
n_employees	“1-5'”, “6-25”, “26-100”, “100-500”, “500-1000”, “More than 1000”, “SE”
tech_company	“Yes”, “No”
employer_mental_health_benefits	“Yes”, “I don’t know”, “No”, “Not eligible for coverage / N/A”, “SE”
employer_mental_health_options	“Yes”, “No”, “I am not sure”, “SE”
employer_discussed_mental_health	“Yes”, “No”, “I don’t know”, “SE”
employer_offer_resources	“Yes”, “No”, “I don’t know”, “SE”
employer_anonymity_protected	“Yes”, “No”, “I don’t know”, “SE”
previous_employers	“Yes”, “No”
previous_employers_mental_health_benefits	“Yes, they all did”, “No, none did”, “Some did”, “I don’t know”, “NPE”
previous_employers_aware-_mental_health_options	“Yes, I was aware of all of them”, “No, I only became aware later”, “I was aware of some”, “N/A (not currently aware)”, “NPE”
previous_employer_discussed_mental_health	“Yes, they all did”, “None did”, “Some did”, “I don’t know”, “NPE”
previous_employers_provide_resources	“Yes, they all did”, “None did”, “Some did”, “NPE”

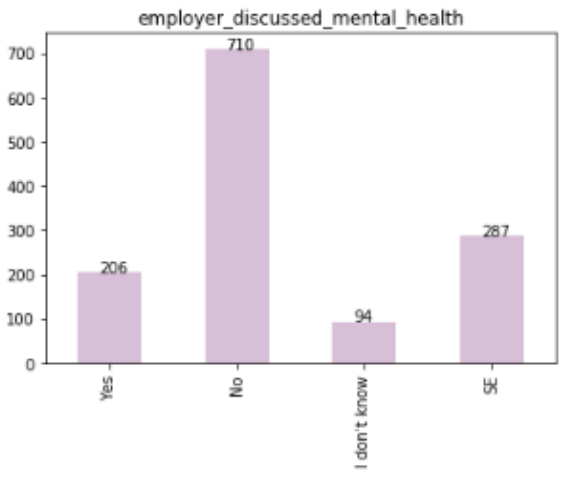
previous_employer_anonymity_protected	“Yes, always”, “No”, “Sometimes”, “I don’t know”, “NPE”
family_history_mental_illness	“Yes”, “No”, “I don’t know”
past_mental_health_disorder	“Yes”, “No”, “Maybe”
diagnosed_by_medical_professional	“Yes”, “No”
gender	“Male”, “Female”, “Others”
work_remotely	“Always”, “Never”, “Sometimes”
country	“United Kingdom”, “United States of America”, “Canada”, “Germany”, “Netherlands”, “Czech Republic”, “Lithuania”, “Australia”, “France”, “Venezuela”, “Poland”, “Belgium”, “Brazil”, “Denmark”, “Sweden”, “Russia”, “Spain”, “India”, “Mexico”, “Switzerland”, “Norway”, “Argentina”, “Ireland”, “Italy”, “Finland”, “Colombia”, “Costa Rica”, “Vietnam”, “Bulgaria”, “New Zealand”, “South Africa”, “Slovakia”, “Austria”, “Bangladesh”, “Algeria”, “Pakistan”, “Afghanistan”, “Greece”, “Romania”, “Other”, “Brunei”, “Japan”, “Iran”, “Hungary”, “Israel”, “Ecuador”, “Bosnia and Herzegovina”, “China”, “Chile”, “Guatemala”, “Taiwan”, “Serbia”, “Estonia”
country_work	“United Kingdom”, “United States of America”, “Canada”, “Germany”, “Netherlands”, “Czech Republic”, “Lithuania”, “Australia”, “France'', “Venezuela”, “Poland”, “Belgium”, “Brazil”, “Denmark”, “Sweden”, “Russia”, “Spain”,

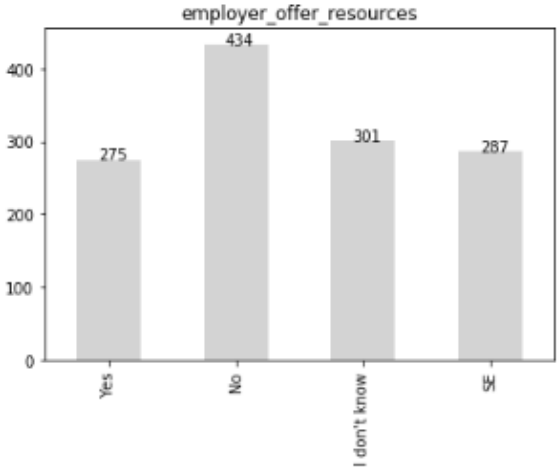
	“India”, “United Arab Emirates”, “Mexico”, “Switzerland”, “Norway''Argentina”, “Ireland”, “Italy”, “Finland”, “Turkey”, “Colombia”, “Costa Rica''Vietnam”, “Bulgaria”, “New Zealand”, “South Africa”, “Slovakia”, “Austria''Bangladesh”, “Pakistan”, “Afghanistan”, “Greece”, “Other”, “Romania”, “Brunei”, “Iran”, “Hungary”, “Israel”, “Japan”, “Ecuador”, “Bosnia and Herzegovina”, “China”, “Chile”, “Guatemala”, “Serbia”, “Estonia”

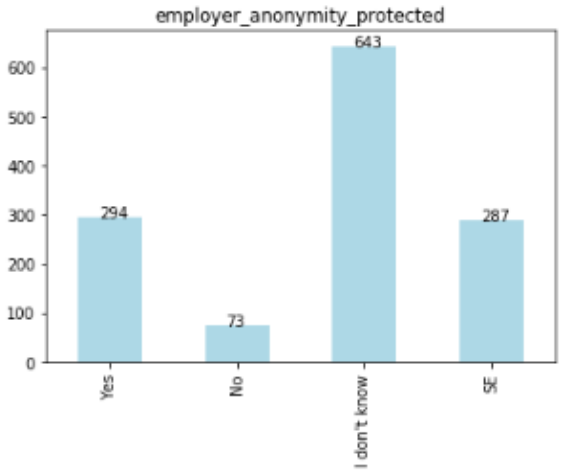
Finalmente, se muestran para las variables categóricas su correspondiente diagrama de barras.

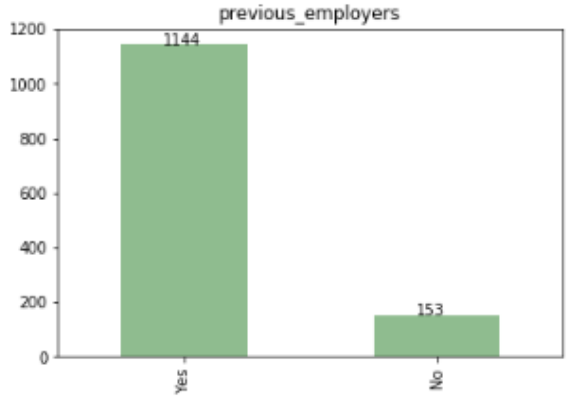


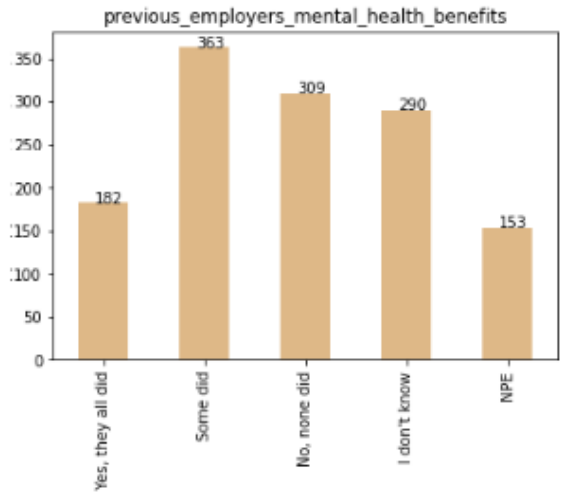

employer_mental_health_options
307
352
351
287
Yes
No
I am not sure
SE

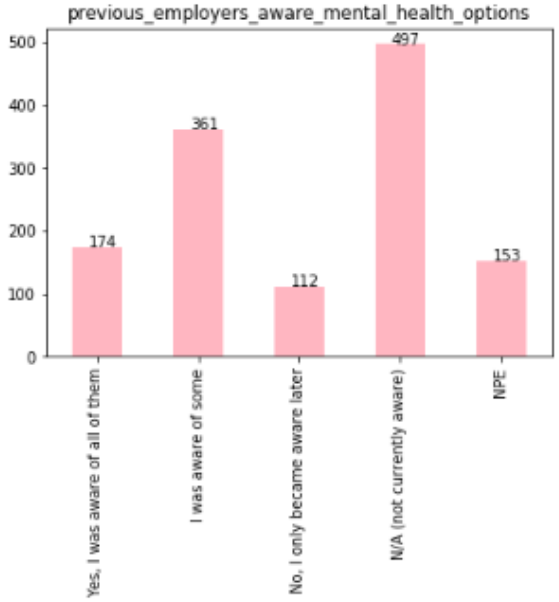

employer_discussed_mental_health
206
710
94
287
Yes
No
I don't know
SE

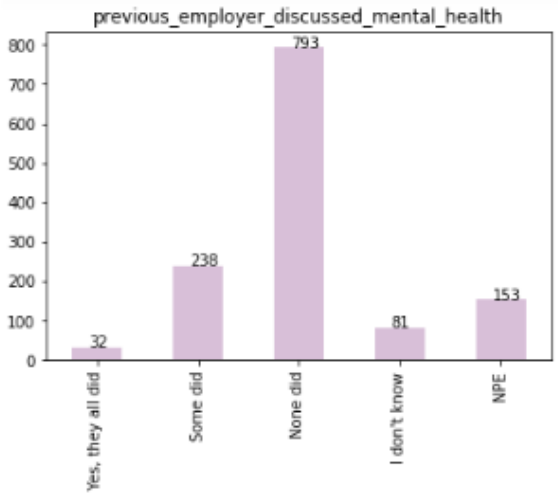

employer_offer_resources
275
434
301
287
Yes
No
I don't know
SE

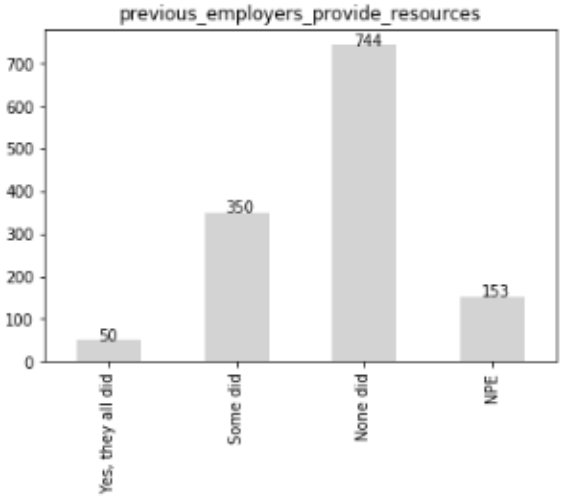

employer_anonymity_protected
294
73
643
287
Yes
No
I don't know
SE

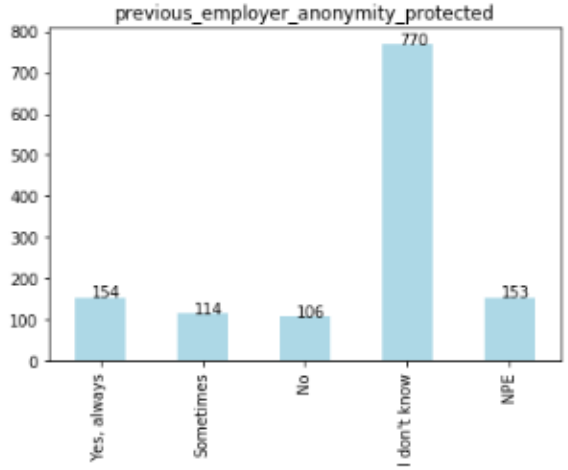

previous_employers
1144
153
Yes
No

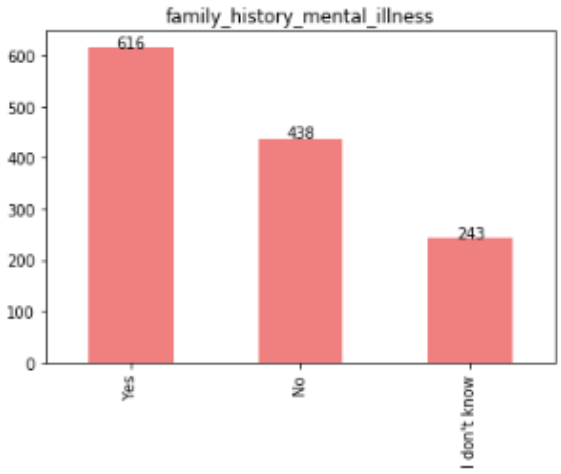

previous_employers_mental_health_benefits
182
363
309
290
153
Yes, they all did
Some did
No, none did
I don't know
NPE

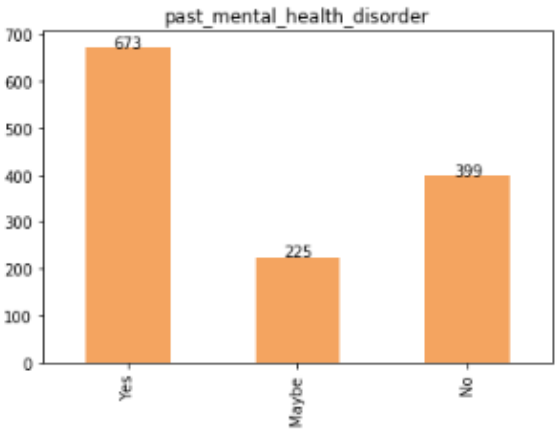

previous_employers_aware_mental_health_options
174
361
112
497
153
Yes, I was aware of all of them
I was aware of some
No, I only became aware later
N/A (not currently aware)
NPE

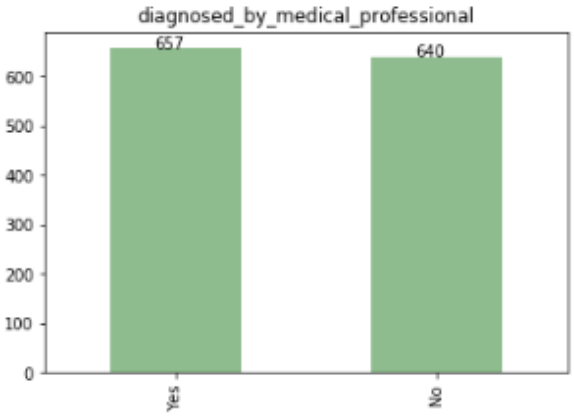

previous_employer_discussed_mental_health
32
238
793
81
153
Yes, they all did
Some did
None did
I don't know
NPE

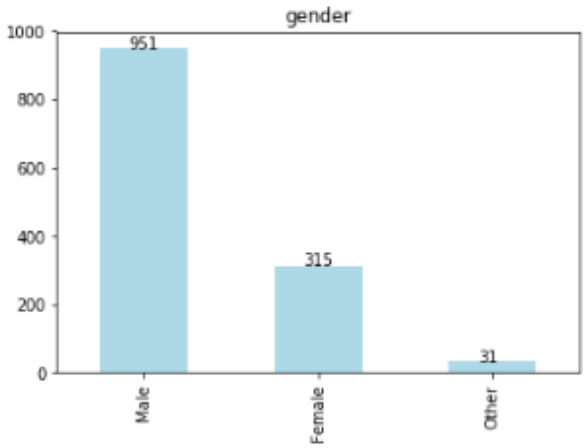

previous_employers_provide_resources
50
350
744
153
Yes, they all did
Some did
None did
NPE

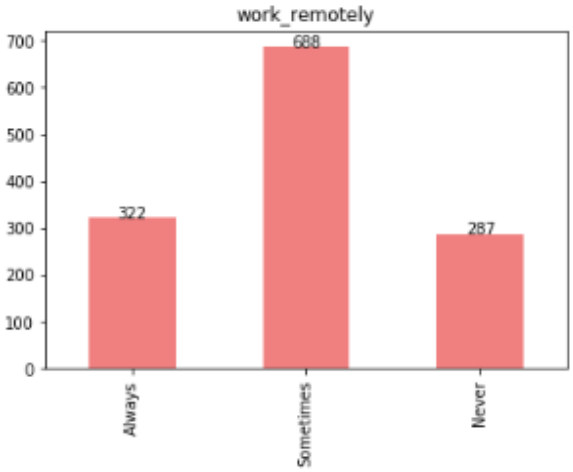

previous_employer_anonymity_protected
154
114
106
770
153
Yes, always
Sometimes
No
I don't know
NPE

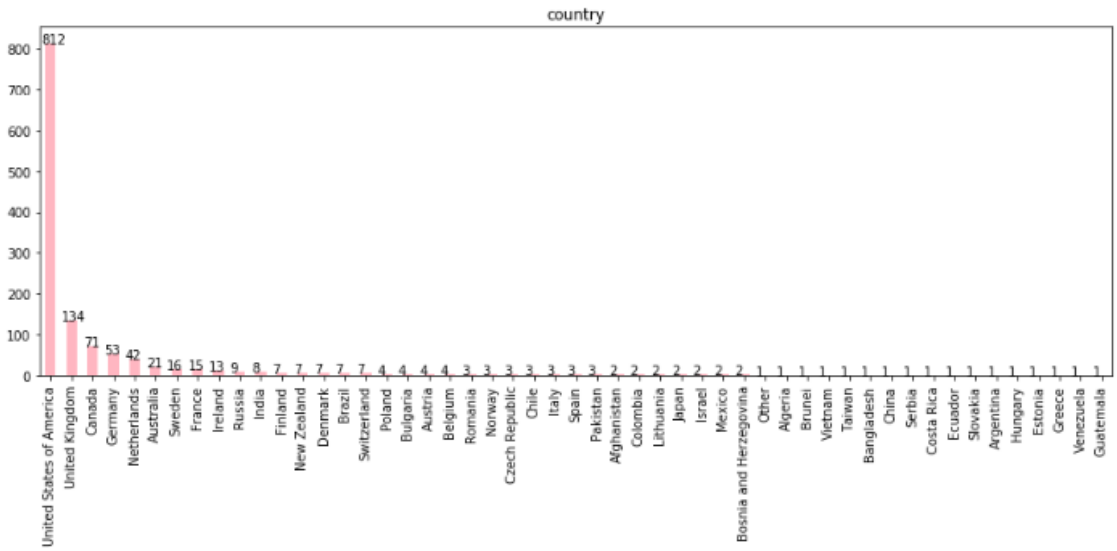

family_history_mental_illness
616
438
243
Yes
No
I don't know


past_mental_health_disorder
673
225
399
Yes
Maybe
No


diagnosed_by_medical_professional
657
640
Yes
No


gender
951
315
31
Male
Female
Other


work_remotely
688
322
287
Always
Sometimes
Never


country
812
134
71
53
42
21
16
15
13
9
8
7
United States of America
United Kingdom
Canada
Germany
Netherlands
Australia
Sweden
France
Ireland
Russia
India
Finland

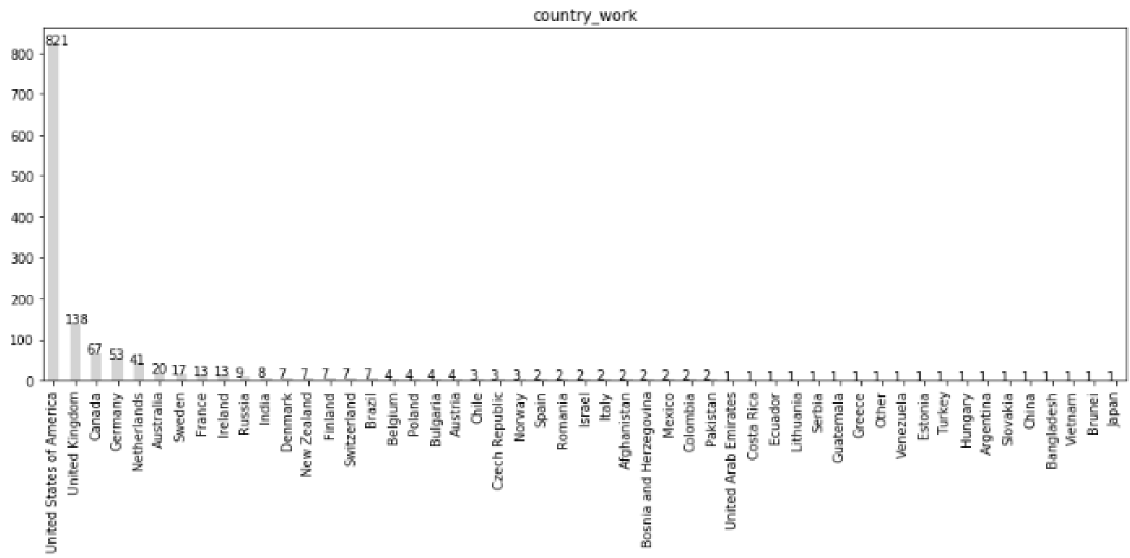


Figura 9. Diagramas de barras de las variables categóricas

3.3.3. Transformación de datos

La transformación de los datos consiste en modificar los datos de tal forma que se encuentren en el formato adecuado para poder emplearlos en el modelo. Existen diferentes tipos de transformaciones, dependiendo de los datos de origen y de los modelos a emplear.

En este caso concreto, la función de la librería que se va a emplear para crear el modelo requiere que los valores de las variables sean numéricos, por lo que es necesario realizar una transformación de valores categóricos a numéricos en todas las variables categóricas.

Existen diferentes técnicas para la transformación de variables categóricas a numéricas. La más común, es asignar a cada una de las categorías posibles un número comenzando desde el 1. Esta técnica se suele conocer como *Label Encoder*. Sin embargo, aunque es muy común utilizarla, existen ciertas fuentes [66], [67] que indican que este método de transformación de variables categóricas puede dar resultados erróneos, ya que asume que los valores numéricos asignados a cada categoría indican un orden, cuando esto no es así. Por este motivo, se acostumbra a emplear la técnica *One Hot Encoding* [68], la cual consiste en dividir cada una de las columnas categóricas del conjunto de datos en tantas columnas como posibles categorías tiene. En cada una de estas columnas se le asigna un 1 en caso de que el valor que tome la observación sea el correspondiente a esa columna y un 0 en caso de que no. Veamos un ejemplo: imaginemos que tenemos una tabla con los siguientes datos y que queremos codificar la columna de “País de origen”:

Tabla 7. Datos de ejemplo para la codificación de variables

País de origen	Número de habitantes (millones)
España	46,72
Reino Unido	66,04
Francia	66,99

Si empleásemos la primera técnica descrita, obtendríamos lo siguiente:

Tabla 8. Datos de ejemplo codificados mediante Label Encoder

País de origen	Pais_encoded	Número de habitantes (millones)
España	1	46,72
Reino Unido	2	66,04
Francia	3	66,99

Mientras que si utilizásemos la técnica de *One Hot Encoding*, obtendríamos el siguiente resultado:

Tabla 9. Datos de ejemplo codificados mediante One Hot Encoding

País de origen	Pais_España	Pais_Reino_Unido	País_Francia	Número de habitantes (millones)
España	1	0	0	46,72
Reino Unido	0	1	0	66,04
Francia	0	0	1	66,99

En este estudio se van a emplear las dos técnicas para codificar las variables categóricas y se van a comparar los resultados para analizar si efectivamente hay una diferencia significativa en los resultados del modelo.

A continuación, se muestra el diccionario de categorías obtenido mediante la primera técnica:

Tabla 10. Diccionario de categorías (Label Encoder)

Nombre de la variable	Codificación de las categorías
self_employed	“No” = 0, “Yes” = 1
n_employees	“1-5” = 0, “100-500” = 1, “26-100” = 2, “500-1000” = 3, “6-25” = 4, “More than 1000” = 5, “SE” = 6
tech_company	“No” = 0, “Yes” = 1
employer_mental_health_benefits	"I don”t know" = 0, “No” = 1, “Not eligible for coverage / N/A” = 2, “SE” = 3, “Yes” = 4
employer_mental_health_options	“I am not sure” = 0, “No” = 1, “SE” = 2, “Yes” = 3
employer_discussed_mental_health	"I don”t know" = 0, “No” = 1, “SE” = 2, “Yes” = 3
employer_offer_resources	"I don”t know" = 0, “No” = 1, “SE” = 2, “Yes” = 3
employer_anonymity_protected	"I don”t know" = 0, “No” = 1, “SE” = 2, “Yes” = 3
previous_employers	“No” = 0, “Yes” = 1
previous_employers_mental_health_benefits	"I don”t know" = 0, “NPE” = 1, “No, none did” = 2, “Some did” = 3, “Yes, they all did” = 4
previous_employers_aware-_mental_health_options	“I was aware of some” = 0, “N/A (not currently aware)” = 1, “NPE” = 2, “No, I only became aware later” = 3, “Yes, I was aware of all of them” = 4
previous_employer_discussed_mental_health	"I don”t know" = 0, “NPE” = 1, “None did” = 2, “Some did” = 3, “Yes, they all did” = 4
previous_employers_provide_resources	“NPE” = 0, “None did” = 1, “Some did” = 2, “Yes, they all did” = 3
previous_employer_anonymity_protected	"I don”t know" = 0, “NPE” = 1, “No” = 2, “Sometimes” = 3, “Yes, always” = 4
family_history_mental_illness	"I don”t know" = 0, “No” = 1, “Yes” = 2
past_mental_health_disorder	“Maybe” = 0, “No” = 1, “Yes” = 2

diagnosed_by_medical_professional	“No” = 0, “Yes” = 1
gender	“Female” = 0, “Male” = 1, “Other” = 2
work_remotely	“Always” = 0, “Never” = 1, “Sometimes” = 2
country	“Afghanistan” = 0, “Algeria” = 1, “Argentina” = 2, “Australia” = 3, “Austria” = 4, “Bangladesh” = 5, “Belgium” = 6, “Bosnia and Herzegovina” = 7, “Brazil” = 8, “Brunei” = 9, “Bulgaria” = 10, “Canada” = 11, “Chile” = 12, “China” = 13, “Colombia” = 14, “Costa Rica” = 15, “Czech Republic” = 16, “Denmark” = 17, “Ecuador” = 18, “Estonia” = 19, “Finland” = 20, “France” = 21, “Germany” = 22, “Greece” = 23, “Guatemala” = 24, “Hungary” = 25, “India” = 26, “Ireland” = 27, “Israel” = 28, “Italy” = 29, “Japan” = 30, “Lithuania” = 31, “Mexico” = 32, “Netherlands” = 33, “New Zealand” = 34, “Norway” = 35, “Other” = 36, “Pakistan” = 37, “Poland” = 38, “Romania” = 39, “Russia” = 40, “Serbia” = 41, “Slovakia” = 42, “Spain” = 43, “Sweden” = 44, “Switzerland” = 45, “Taiwan” = 46, “United Kingdom” = 47, “United States of America” = 48, “Venezuela” = 49, “Vietnam” = 50
country_work	“Afghanistan” = 0, “Argentina” = 1, “Australia” = 2, “Austria” = 3, “Bangladesh” = 4, “Belgium” = 5, “Bosnia and Herzegovina” = 6, “Brazil” = 7, “Brunei” = 8, “Bulgaria” = 9, “Canada” = 10, “Chile” = 11, “China” = 12, “Colombia” = 13, “Costa Rica” = 14, “Czech Republic” = 15, “Denmark” = 16, “Ecuador” = 17, “Estonia” = 18, “Finland” = 19, “France” = 20, “Germany” = 21, “Greece” = 22, “Guatemala” = 23, “Hungary” = 24, “India” = 25, “Ireland” = 26, “Israel” = 27, “Italy” = 28, “Japan” = 29, “Lithuania” = 30, “Mexico” = 31, “Netherlands” = 32, “New Zealand” = 33, “Norway” = 34, “Other” = 35, “Pakistan” = 36, “Poland” = 37, “Romania” = 38, “Russia” = 39, “Serbia” = 40, “Slovakia” = 41, “Spain” = 42, “Sweden” = 43, “Switzerland” = 44, “Turkey” = 45, “United Arab Emirates” = 46, “United Kingdom” = 47, “United States of America” = 48, “Venezuela” = 49, “Vietnam” = 50

3.4. Construcción del modelo

Cuando se lleva a cabo un modelo de clasificación, es importante que el conjunto de datos se encuentre balanceado, es decir, que la proporción entre observaciones de cada

clase de la variable dependiente en el conjunto de datos esté igualada. En este caso, se ha seleccionado como variable dependiente del modelo la variable “Have you been diagnosed with a mental health condition by a medical professional?”. Puesto que el número de observaciones tras el tratamiento de *missing values* de la categoría “Yes” de esta variable es de 657 y de la categoría “No” de 640, no se considera que es necesario llevar a cabo un proceso de balanceo. Existen diferentes técnicas para balancear conjuntos de datos como, por ejemplo, el submuestreo que consiste en eliminar observaciones de la clase mayoritaria con el objetivo de igualar los tamaños de las clases [69].

Puesto que existen multitud de combinaciones de parámetros, se van a crear varios árboles de decisión con parámetros distintos de tal forma que nos permita conocer cuál es la mejor combinación. Estos parámetros han sido explicados anteriormente en el apartado 3.1.3.3. Además, cada uno de estos árboles se va a modelar con los dos conjuntos de datos codificados que se han creado. A continuación, se muestra una tabla de resumen de los diferentes árboles de decisión creados con sus correspondientes parámetros.

Tabla 11. Árboles de decisión creados y sus parámetros

Número de modelo	**Límite de profundidad (*max_depth*)**	**Mínimo de muestras para dividir nodo interno (*min_samples_split*)**	**Mínimo de muestras requeridas para nodo final (*min_samples_leaf*)**
1	Sin límite (predet.)	2 (predet.)	1 (predet.)
2	Sin límite (predet.)	2 (predet.)	5
3	Sin límite (predet.)	5	1 (predet.)
4	Sin límite (predet.)	5	5
5	3	2 (predet.)	1 (predet.)
6	3	2 (predet.)	5
7	3	5	1 (predet.)

8	3	5	5
9	5	2 (predet.)	1 (predet.)
10	5	2 (predet.)	5
11	5	5	1 (predet.)
12	5	5	5

Además de crear estos 12 árboles, se van a crear también modelos de aprendizaje supervisado mediante *Random Forest*, cuyos resultados se contrastarán con los obtenidos con los árboles de decisión. A continuación, se muestra una tabla de resumen de los diferentes modelos *Random Forest* creados con sus correspondientes parámetros. El parámetro que seleccionado se explica en el apartado 3.1.3.3.

Tabla 12. Modelos Random Forest credos y sus parámetros

Número de modelo	Número de estimadores (*n_estimators*)
13	10 (predet.)
14	50
15	100
16	200

3.5. Evaluación e interpretación del resultado del modelo

A continuación, se muestra una tabla de resumen de diferentes árboles de decisión creados con sus correspondientes características y la precisión obtenida con cada uno de ellos. En el apartado 3.1.3.3 se ha explicado qué es la precisión en este contexto.

Tabla 13. Precisión obtenida con los árboles de decisión

Número de modelo	Límite de profundidad	Mínimo de muestras para dividir nodo interno	Mínimo de muestras requeridas para nodo final	Precisión con Label Encoder	Precisión con One Hot Encoding	Precisión media
1	Sin límite (predet.)	2 (predet.)	1 (predet.)	0.8	0.7846	0.7923
2	Sin límite (predet.)	2 (predet.)	5	0.8692	0.8192	0.8442
3	Sin límite (predet.)	5	1 (predet.)	0.7962	0.7731	0.7846
4	Sin límite (predet.)	5	5	0.8692	0.8192	0.8442
5	3	2 (predet.)	1 (predet.)	0.8769	0.8808	0.8788
6	3	2 (predet.)	5	0.8808	0.8808	0.8808
7	3	5	1 (predet.)	0.8769	0.8808	0.8788
8	3	5	5	0.8808	0.8808	0.8808
9	5	2 (predet.)	1 (predet.)	0.8615	0.8769	0.8692
10	5	2 (predet.)	5	0.8692	0.8731	0.8711
11	5	5	1 (predet.)	0.8615	0.8731	0.8673
12	5	5	5	0.8692	0.8731	0.8711
				0.8592	0.8512	**0.8552**

No existe un criterio universal para determinar la bondad de la precisión de un modelo ya que es algo completamente dependiente de la aplicación para la cual se está empleando el modelo [70]. En este caso, se puede considerar que la precisión obtenida con los modelos creados es buena: todas ellas se encuentran entre 0.7846 y 0.8808. Es decir, el modelo que menor precisión ha conseguido ha clasificado correctamente el 78% de las muestras del

conjunto de datos de prueba, mientras que el mejor modelo ha clasificado correctamente el 88% de las muestras.

En cuanto a la diferencia entre los modelos creados los diferentes métodos de codificación, se puede observar que prácticamente se ha obtenido la misma precisión con cada uno de ellos: 0.8592 con el conjunto de datos codificado mediante *Label Encoder* y 0.8512 con el codificado por *One Hot Encoding*. Tampoco hay una diferencia significativa en la precisión con respecto a los parámetros del modelo escogido, aunque se puede apreciar que se obtiene una precisión ligeramente mayor cuando la profundidad del árbol de decisión es igual a 3.

Como se ha comentado anteriormente, se han creado también modelos mediante *Random Forest*. A continuación, se muestra una tabla de resumen de diferentes modelos creados con sus correspondientes características y la precisión obtenida con cada uno de ellos.

Tabla 14. Precisión obtenida con Random Forest

Número de modelo	Número de estimadores	Precisión con Label Encoder	Precisión con One Hot Encoding	Precisión media
13	10 (predet.)	0.8615	0.85	0.8557
14	50	0.8769	0.8885	0.8827
15	100	0.8769	0.8846	0.8807
16	200	0.8769	0.8808	0.8788
		0.8730	0.8759	**0.8745**

Se puede observar que la precisión obtenida en los modelos de *Random Forest* es elevada también; la precisión de todos los modelos se encuentra entre 0.85 y 0.8846. Si se compara la precisión media de todos los modelos de *Random Forest* generados, se puede apreciar que es ligeramente mayor que la media de la precisión de los árboles de decisión (0.8745 frente 0.8552) pero no es una diferencia notable. En cuanto a la diferencia entre los modelos generados mediante las dos técnicas de codificación, ocurre como en el caso de los árboles de decisión; la diferencia es mínima. Finalmente, en cuanto a la diferencia entre modelos teniendo en cuenta el número de estimadores, se puede observar que el mejor

resultado se obtiene cuando el número de estimadores es igual a 200, aunque en este caso tampoco hay grandes diferencias.

Analizando las precisiones obtenidas con los diferentes parámetros, se va a seleccionar el modelo número 6 para analizar los resultados obtenidos con él. El motivo por el que se ha seleccionado este modelo es porque, aunque no existen grandes diferencias entre los modelos creados, presenta una de las precisiones más altas (0.8808) con los dos tipos de codificaciones. Además, es de los modelos más simples que se ha construido: es un árbol de decisión con profundidad de 3 nodos, el mínimo de muestras para dividir un nodo interno es de 2 y el mínimo de muestras requeridas para nodo final es de 5.

A continuación, se muestran las variables que, según este modelo, tienen más efecto sobre la variable dependiente. Se muestran únicamente aquellas que el modelo ha decidido que tienen una importancia mayor que 0. La importancia de una variable se calcula como la media de descenso en la impuridad con respecto a la clase a predecir que provoca esa variable en el conjunto de datos.

- Modelo creado con datos codificados mediante *Label Encoder*

	importance
past_mental_health_disorder	0.973347
employer_mental_health_options	0.012176
gender	0.006155
age	0.004472
family_history_mental_illness	0.003850

- Modelo creado con datos codificados mediante *One Hot Encoding*

	importance
past_mental_health_disorder_Yes	0.940169
past_mental_health_disorder_No	0.029458
gender_Male	0.009558
employer_mental_health_options_Yes	0.006749
family_history_mental_illness_No	0.005905
employer_mental_health_options_No	0.005560
previous_employers_mental_health_benefits_Yes, they all did	0.002601

Tal y como se puede apreciar, ambos modelos consideran de mayor impacto prácticamente las mismas variables. En ambos casos, la más importante es, con diferencia, si el trabajador ha sufrido anteriormente una enfermedad mental. Por otro lado, el primer modelo considera que el conocimiento de cuidados mentales ofrecidos por la empresa donde trabaja es la segunda más importante, mientras que el segundo modelo considera haber respondido “Si” o “No” a esta pregunta en 4º y 6º lugar, respectivamente. Por otro lado, el primer modelo considera el sexo del trabajador el tercer factor más importante y el segundo modelo también, concretamente el ser hombre. En el 5º de ambos modelos, se encuentra la pregunta del cuestionario que hace referencia a si el trabajador tiene o no familiares con enfermedades mentales. En el segundo modelo concretamente haber respondido “No” a esta pregunta. Los únicos factores donde no coindicen ambos modelos son, en la edad del trabajador (el primer modelo lo considera importante en 4º lugar) y en el hecho de que empresas de trabajos previos hayan ofrecido al trabajador ayudas con relación a la salud mental. Concretamente, el haber respondido que todas sus anteriores empresas ofrecían ayudas, el segundo modelo lo considera importante en el 7º y último lugar, mientras que el primero no lo considera importante. En el siguiente apartado se analizarán estas conclusiones en más detalle mediante el análisis visual de los árboles obtenidos.

3.6. Análisis visual

Se denomina Visualización Analítica (o *Visual Analytics*, en inglés) a “la ciencia del razonamiento analítico facilitada por interfaces visuales” [71]. El razonamiento analítico es la capacidad de detectar patrones dentro de los datos y de comprenderlos en detalle observando representaciones visuales de grandes cantidades de los mismos.

Uno de los criterios más extendidos para clasificar los tipos de visualizaciones es la dimensionalidad de la visualización, es decir, el número de variables que la visualización es capaz de mostrar. La visualización univariada -una única dimensión- es la forma más simple de análisis visual de los datos y su objetivo principal es el de conocer la distribución, la tendencia central y la extensión de una única variable. Por otro lado, el objetivo principal de la visualización multivariada -dos o más dimensiones- es el de permitir el análisis de la relación y/o interacción entre diferentes variables [23], [72].

A continuación, se muestran dos representaciones visuales de los datos. Por un lado, se analiza la prevalencia de enfermedades mentales en trabajadores tecnológicos mediante diferentes representaciones visuales de los datos. Por otro lado, se analiza visualmente el modelo creado.

3.6.1. Análisis de la prevalencia de enfermedades mentales en trabajadores tecnológicos

En este apartado, se va a analizar la prevalencia de enfermedades mentales mediante gráficas univariables para, a continuación, comparar algunas variables de interés mediante visualizaciones multivariables. Para ello, se va a utilizar el conjunto de datos preprocesado antes de la codificación de los datos (ver apartado 3.3).

En primer lugar, se va a estudiar la prevalencia de enfermedades mentales entre los trabajadores tecnológicos mediante un gráfico de barras que va a permitir conocer la cantidad de trabajadores que sufren alguna enfermedad mental. Concretamente, si han sido diagnosticados o no por un profesional.

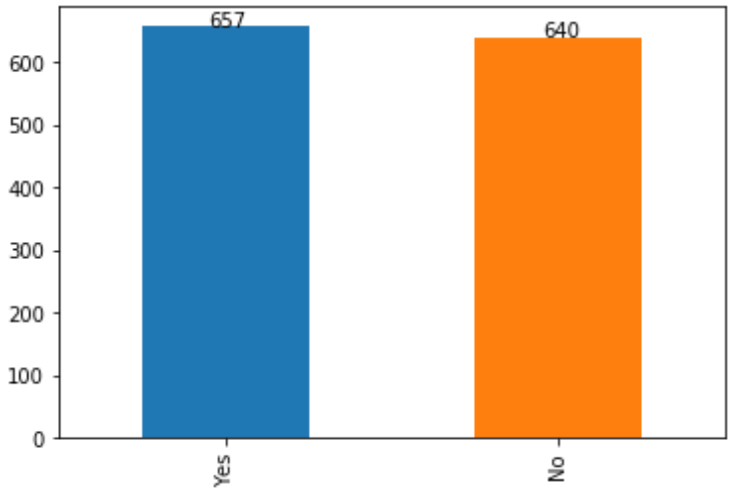


Figura 10. Cantidad de trabajadores diagnosticados por un profesional

Como se puede observar, 657 de los trabajadores encuestados han sido diagnosticados de alguna enfermedad mental por un profesional, mientras que 640 no han sido diagnosticados de ninguna enfermedad mental (bien porque han sido atendidos por un especialista y éste ha determinado que la persona no tenía una enfermedad mental o bien porque directamente no han sido nunca atendidos por un profesional).

A continuación, se va a entrar más en detalle en la tipología de las enfermedades mentales sufridas por estos trabajadores y, para ello, se va a presentar un diagrama de barras que muestra la cantidad de veces que se ha diagnosticado cada una de las enfermedades entre este grupo de personas. Aunque una misma persona haya podido ser diagnosticada de más de una enfermedad mental, estas enfermedades se van a contabilizar por separado para poder detectar cuales son las más comunes. A continuación, se muestra un diagrama de

barras horizontal con las 10 enfermedades mentales diagnosticadas por profesionales más comunes y el número de veces que han sido diagnosticadas, ordenadas de mayor a menor.

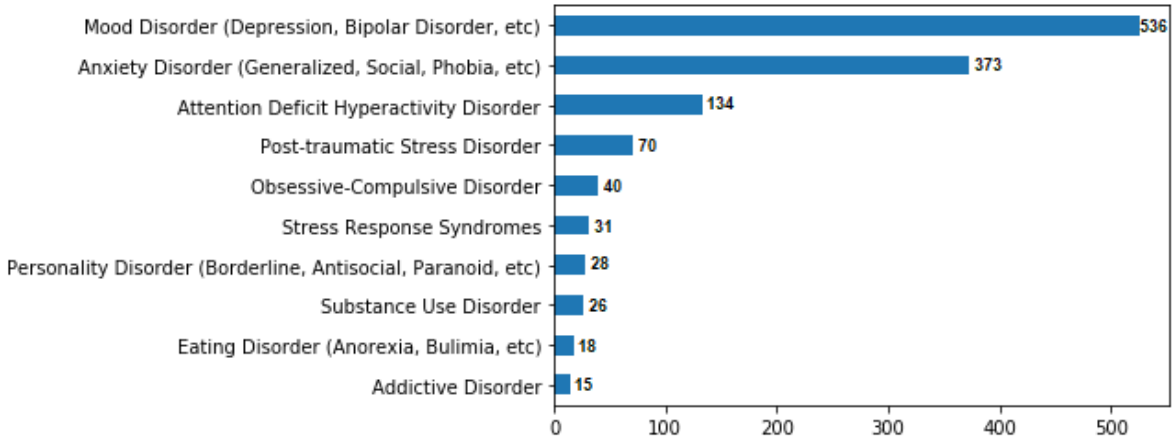


Figura 11. Las 10 enfermedades mentales más comunes

Como se puede observar en la gráfica, las enfermedades mentales más comunes diagnosticadas entre los trabajadores tecnológicos son el trastorno emocional y el trastorno de ansiedad, con 536 y 373 diagnósticos respectivamente. Le siguen el déficit de atención e hiperactividad (134 diagnósticos), el trastorno de estrés postraumático (70 diagnósticos), el síndrome de estrés postraumático (40 diagnósticos), el trastorno de la personalidad (31 diagnósticos), los trastornos por consumo de sustancias (26 diagnósticos), los trastornos alimenticios (18 diagnósticos) y, finalmente, los trastornos adictivos (15 diagnósticos).

A continuación se van a analizar el número de enfermedades mentales distintas que ha podido tener un mismo trabajador para conocer qué es lo más común entre este perfil de personas. Estas enfermedades se han podido dar a la vez o no.

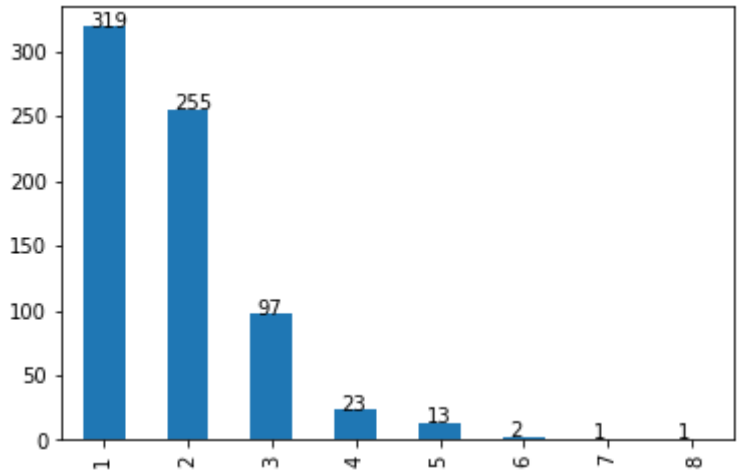


Figura 12. Número de enfermedades mentales por trabajador

Se puede apreciar que lo más común es haber sido diagnosticado únicamente de una enfermedad mental, aunque una gran cantidad de los trabajadores encuestados ha tenido dos enfermedades mentales. Se puede observar que el número máximo de enfermedades de las que ha sido diagnosticado un mismo trabajador son 8.

Finalmente, se va a analizar la prevalencia de enfermedades mentales desde el punto de vista sociodemográfico. Es decir, se va a representar la prevalencia de estas enfermedades mediante gráficas diferentes gráficas organizadas por sexo, edad y país.

A continuación, se muestra la prevalencia de enfermedades mentales por sexo. Tal y como se ha explicado en el apartado 3.3, los datos en relación al sexo del trabajador se han reorganizado en tres categorías: “Male”, “Female” y “Others”. La gráfica seleccionada para este tipo de visualización es una gráfica circular donde se muestra, para cada sexo, el porcentaje de personas que han sido diagnosticadas de alguna enfermedad mental por un profesional y las que no.

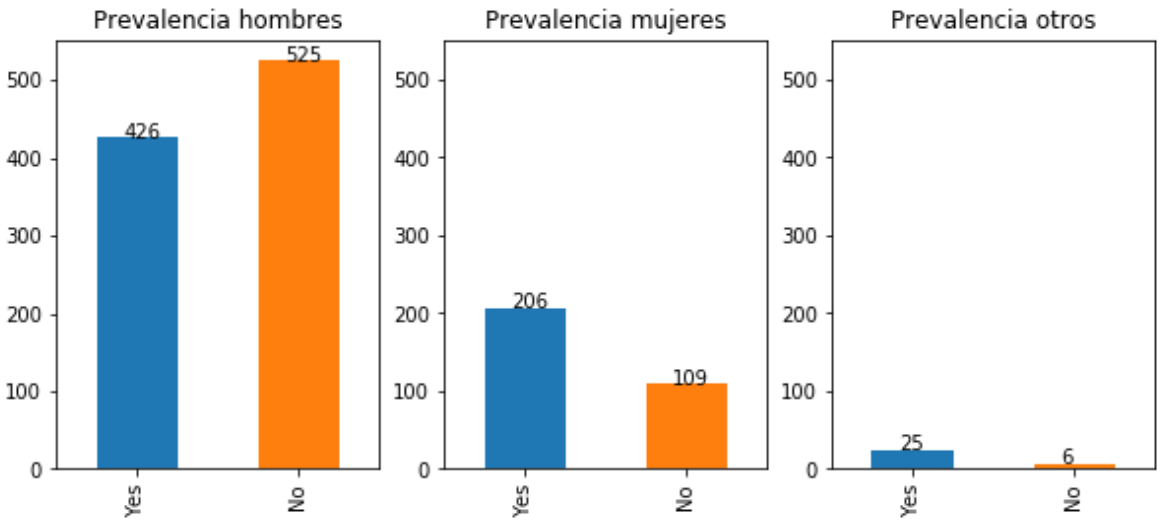


Figura 13 . Prevalencia de enfermedades mentales en hombres, mujeres y otros

En el caso de los hombres, una mayor parte de ellos no han sido diagnosticados de alguna enfermedad mental (426 lo han sido frente a 525 que no). Sin embargo, en el caso de las mujeres la prevalencia es mucho mayor; el número de mujeres diagnosticadas de alguna enfermedad mental prácticamente duplica a las que no lo han sido. Finalmente, en el caso de las personas bajo la categoría de sexo “Others” la proporción es mayor aún; 25 personas han sido diagnosticadas, mientras que tan solo 6 no lo han sido.

Por otro lado, se va a analizar el impacto de la edad en el padecimiento de alguna enfermedad mental mediante dos gráficas de boxplot; una que muestra la distribución de la edad los trabajadores diagnosticados de alguna enfermedad mental y otra para los que no han digo diagnosticados nunca de alguna enfermedad mental.

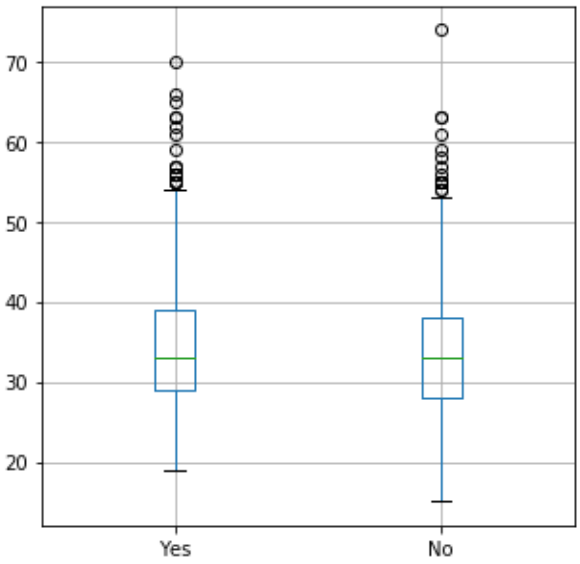


Figura 14. Diagnóstico de enfermedades mentales por edad

Se puede apreciar que la distribución es muy similar en ambos casos. La mediana de edad se encuentra alrededor de los 33 años tanto en los pacientes que han sido diagnosticados de una enfermedad mental como en los que no. Se puede apreciar que la mitad intermedia en el caso de los trabajadores diagnosticados de alguna enfermedad mental tiene una edad ligeramente superior a los que no, pero no existen grandes diferencias. Por otro lado, se observa que tanto el valor mínimo como el máximo en el caso de los trabajadores no diagnosticados toma valores menores que en el caso de los diagnosticados.

Finalmente, se analiza la prevalencia de enfermedades mentales por país de residencia y país donde se trabaja para poder comparar cuales son los países donde existe más problema de enfermedades mentales. Tal y como se ha visto en el análisis exploratorio de las variables (apartado 3.3.2), el número de trabajadores que han respondido a la encuesta no es el mismo para cada uno de los países, ni de residencia ni de trabajo. De hecho, existen grandes diferencias.

Por este motivo, únicamente se van a analizar la prevalencia de enfermedades mentales en los países en los que tienen un mínimo de participantes de ese país. Ya que, si por ejemplo se tiene una única persona de un país y esta persona ha respondido que sí ha sido diagnosticado de una enfermedad mental, no es representativo decir que la prevalencia de enfermedades mentales en ese país es del 100%. Aunque existen diferentes teorías, una muy común es la [73], donde se calcula que el mínimo de muestras necesarias para que la prevalencia calculada sea representativa es de 30, por lo que se van a analizar la prevalencia de los países donde por lo menos tienen 30 participantes de dicho país.

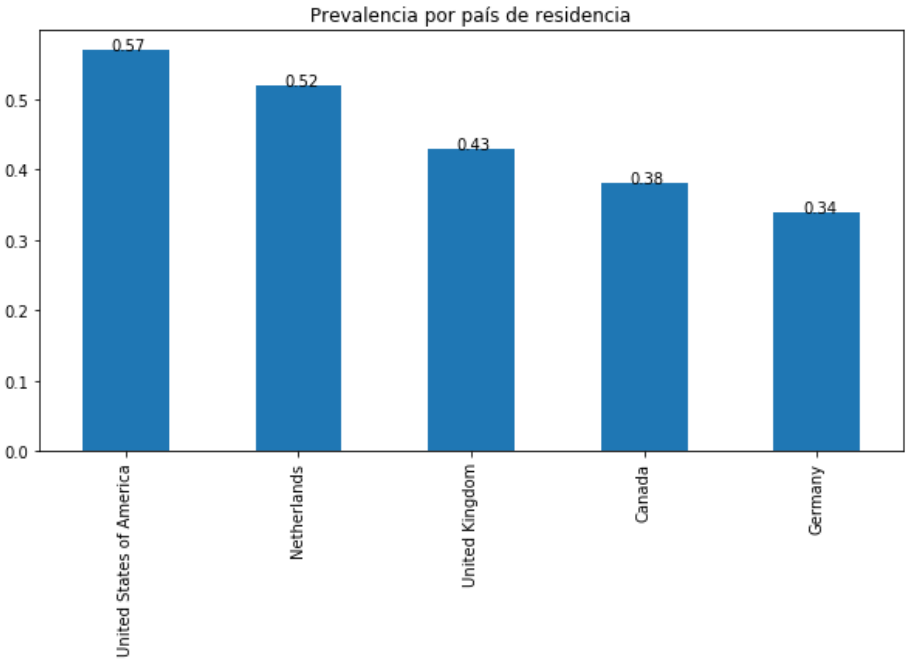


Figura 15. Prevalencia de enfermedades mentales por país de residencia

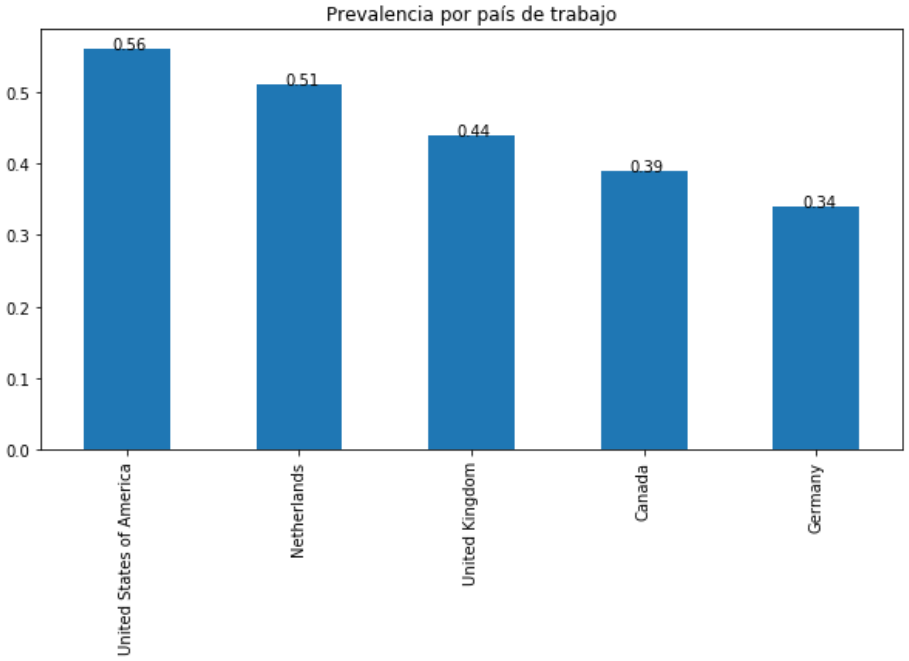


Figura 16. Prevalencia de enfermedades mentales por país de trabajo

Se puede observar que los países con mayor proporción de trabajadores tecnológicos con enfermedades mentales residiendo en ellos son, descendentemente: Estados Unidos, Holanda, Reino Unido, Canadá y Alemania. Por país de trabajo, todos los países se

mantienen en sus puestos y la prevalencia obtenida en ellos es prácticamente igual, lo cual hace pensar que la mayor parte de los trabajadores encuestados viven y trabajan en el mismo país.

3.6.2. Análisis visual del modelo

En esta sección se va a mostrar la representación gráfica del árbol de decisión seleccionado en el apartado 3.5. En primer lugar, se va a mostrar el árbol generado con el conjunto de datos codificados mediante Label Encoder y, a continuación, el árbol creado con los datos codificados con la técnica One Hot Encoding.

3.6.2.1. Análisis visual del modelo generado con datos codificados mediante Label Encoder

La representación gráfica del árbol de decisión seleccionado creado con los datos codificados mediante la técnica Label Encoder se muestra a continuación.

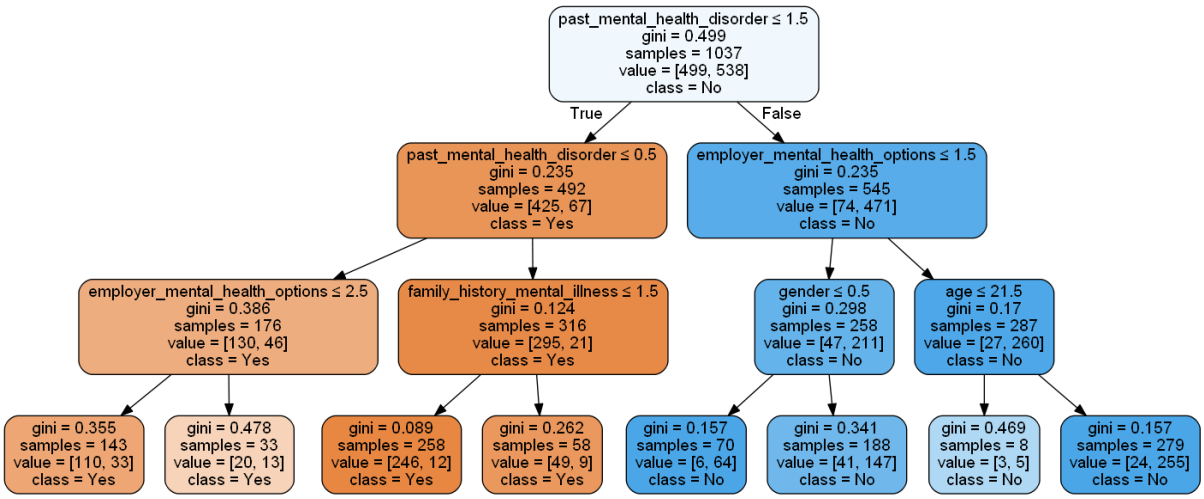


Figura 17. Árbol de decisión con datos codificados mediante Label Encoder

Con el objetivo de reducir la explicación del árbol, se van a seleccionar dos ramas del árbol que llevan a un nodo final con mayor índice de Gini para cada caso (i.e. diagnosticado por un profesional si / no) y se van a explicar. Para poder interpretar este árbol, es necesario recurrir al diccionario de categorías del apartado 3.3.3. Las ramas seleccionadas son las que se muestran a continuación:

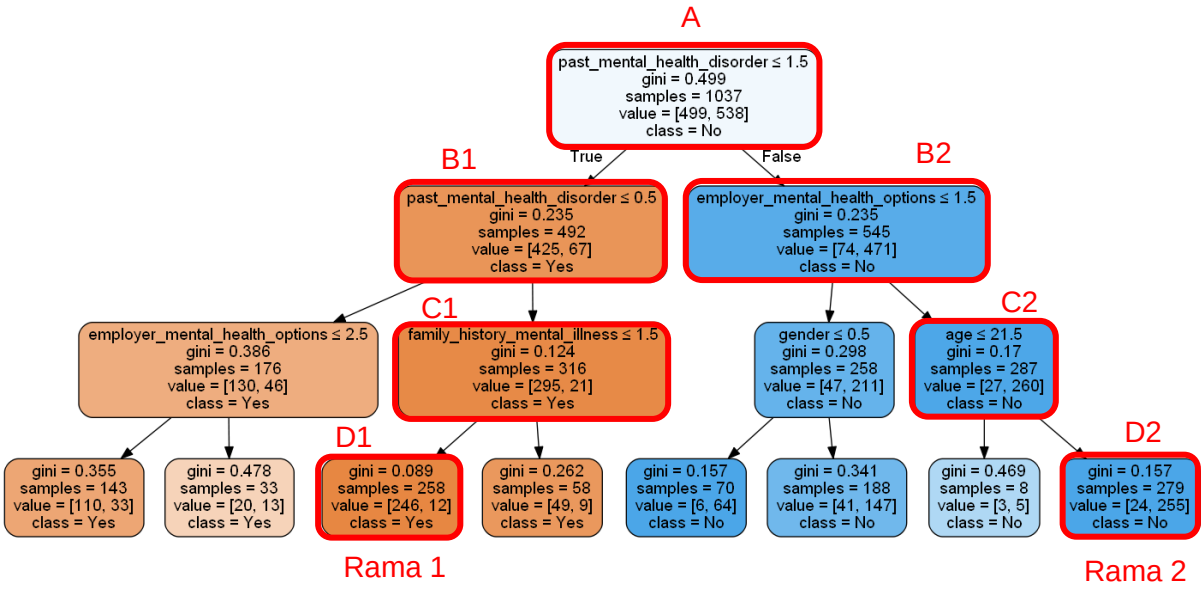


Figura 18. Árbol de decisión con datos codificados mediante Label Encoder (selección)

Se puede observar que el nodo raíz (nodo A) contiene 1037 muestras. Estas muestras son todas las del conjunto de entrenamiento (ver apartado 3.1.3.3). El hecho de que la clase (parámetro **Class**) de este nodo sea “No” indica que hay una mayoría de trabajadores en el conjunto de entrenamiento que no han sido diagnosticados de una enfermedad mental. Sin embargo, el índice de Gini es bastante elevado (0.499), lo cual indica que en este primer nodo no hay muchos más trabajadores que no han sido diagnosticados de una enfermedad mental que los que sí lo han sido. Esto lo podemos comprobar analizando el parámetro **Value**, el cual nos indica que en este nodo hay 499 trabajadores no diagnosticados y 538 que sí.

La rama número 1, corresponde a una decisión de la clase “Yes” (i.e. haber sido diagnosticado de alguna enfermedad mental). Esta rama de decisión se podría interpretar de la siguiente manera: “si el trabajador no ha tenido una enfermedad mental en el pasado (nodo A), no sabe si ha tenido o no ha tenido familiares con enfermedades mentales (nodos B1 y C1), entonces es muy probable que tenga una enfermedad mental (nodo D1)”.

Por otro lado, la rama número 2, se podría interpretar tal que así: “si el trabajador ha sufrido anteriormente una enfermedad mental (nodo A), sí dispone en la empresa de ayuda con respecto a la salud mental o es autónomo (nodo B2) y es mayor de 21 años y medio (nodo C2), lo más probable es que no tenga una enfermedad mental (nodo D2)”.

Como se puede apreciar, la interpretación de este árbol de decisión es bastante difícil e imprecisa puesto que las variables categóricas se encuentran codificadas.

3.6.2.2. Análisis visual del modelo generado con datos codificados mediante One Hot Encoding

Por otro lado, la representación gráfica obtenida del árbol de decisión seleccionado creado con los datos codificados mediante la técnica One Hot Encoding es la siguiente:

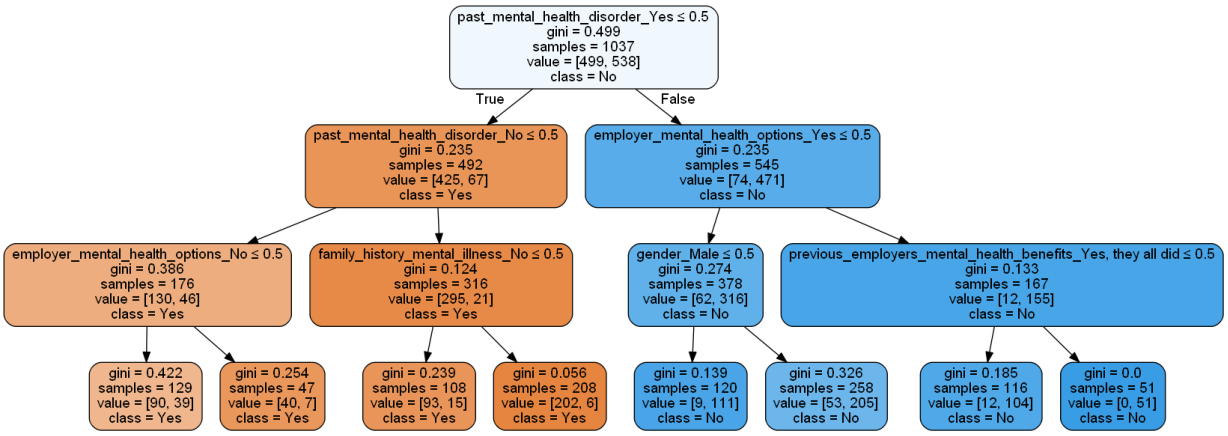


Figura 19. Árbol de decisión con datos codificados mediante One Hot Encoding

A continuación, se va a analizar el modelo creado con los datos codificados mediante la técnica de *One Hot Encoding*. Al igual que en el caso anterior, se van a seleccionar dos ramas.

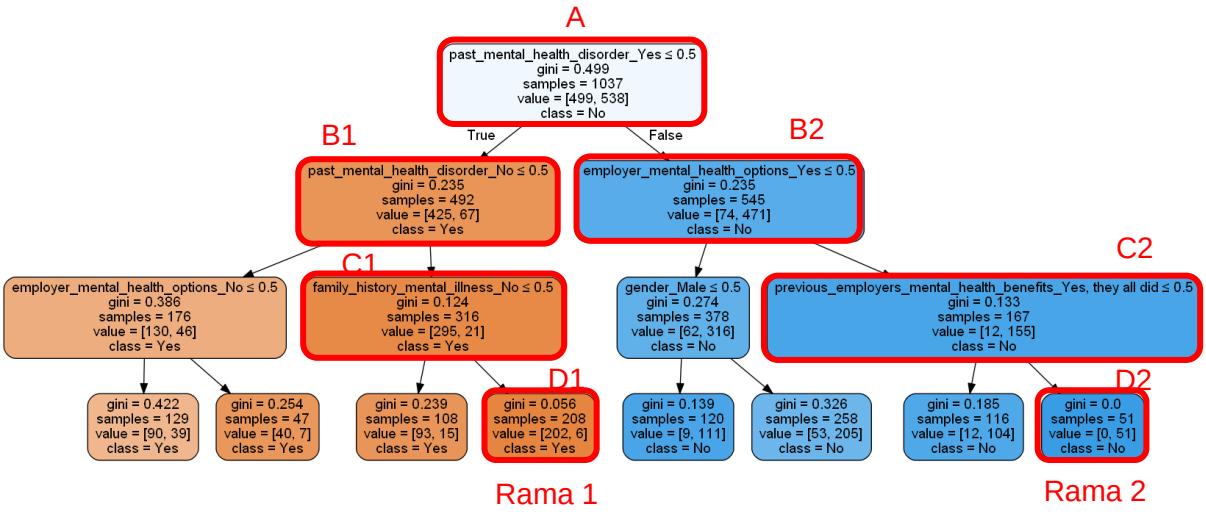


Figura 20. Árbol de decisión con datos codificados mediante One Hot Encoding (selección)

Al igual que en el caso anterior, el nodo raíz (nodo A) contiene 1037 muestras correspondientes a todas las muestras del conjunto de entrenamiento del modelo (ver

apartado 3.1.3.3). Puesto que los datos de los que se parte para generar ambos modelos son los mismos, la clase del nodo raíz en este caso también es “No” y el índice de Gini es exactamente el mismo.

En este caso, debido al tipo de codificación que se ha llevado a cabo en las variables categóricas, los valores de todas estas variables pueden tomar valores de 0 o 1. Por lo tanto, que por ejemplo un nodo tenga un criterio de que una variable sea <0.5 significará que la variable sea igual a 0.

La rama número 1, corresponde a una decisión de la clase “Yes” (i.e. haber sido diagnosticado de alguna enfermedad mental). Esta rama de decisión se podría interpretar de la siguiente manera: “si el trabajador no ha tenido anteriormente una enfermedad de salud mental en el pasado (nodo A y B1) y no ha tenido familiares con enfermedades mentales (nodo C1), es muy probable que sufra una enfermedad mental (nodo D1)”.

Por otro lado, la rama número 2, se podría interpretar tal que así: “si el trabajador ha sufrido anteriormente una enfermedad mental (nodo A), sí dispone en la empresa de ayuda con respecto a la salud mental (nodo B2), las empresas donde ha trabajado anteriormente no ofrecían ayudas con respecto a la salud mental (nodo C2), lo más probable es que no tenga una enfermedad mental (nodo D2)”.

Al igual que en el caso anterior, la interpretación de este árbol es muy complicada e inexacta debido a la codificación de las variables categóricas.

3.7. Resultados utilizando el software Orange

Tal y como se ha explicado en el apartado 3.1.3.4, se va a crear el mismo modelo mediante el software Orange con el objetivo de ver cómo hubiese quedado utilizando esta librería. Este árbol tendrá las mismas características que el árbol seleccionado en el apartado 3.5 (*max_depth* = 3, *min_samples_split* = 2 y *min_samples_leaf* = 5). La precisión obtenida con este modelo en Orange es de 0.84, muy similar a los resultados obtenidos con Scikit Learn (0.88).

.

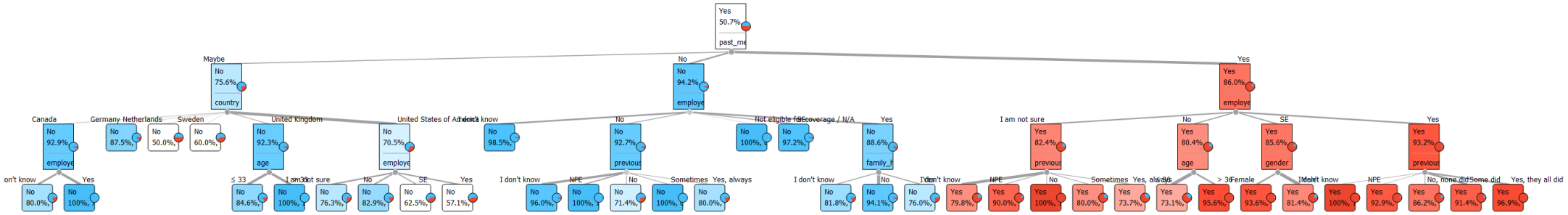


Figura 21. Árbol de decisión creado con Orange

(Link a la imagen con mayor calidad)

El árbol de decisión creado con Orange clasifica a los trabajadores tecnológicos teniendo en cuenta las siguientes variables:

- Si ha tenido anteriormente una enfermedad mental
- El país en el que vive
- Si la empresa donde trabaja ofrece ayudas con respecto a la salud mental
- Si el trabajador conoce las opciones de cuidado de salud mental que ofrece la empresa donde trabaja
- Si las empresas donde ha trabajado anteriormente protegían la anonimidad de aquellos que recurrían a ayudas sobre la salud mental
- Si las empresas donde ha trabajado anteriormente ofrecían ayudas con respecto a la salud mental
- La edad del trabajador
- El sexo del trabajador
- Si tiene familiares con enfermedades mentales

Se puede apreciar que muchas de las variables consideradas de alto impacto coinciden con las detectadas mediante la librería Scikit Learn.

Este árbol, al tener en cuenta las variables categóricas sin necesidad de codificarlas, es mucho más fácil interpretar los resultados obtenidos. Además, la interpretación se convierte más exacta puesto que las decisiones se toman en base a las categorías y no a rangos de valores numéricos asignados a las categorías.

4. Conclusiones

A lo largo de este trabajo se han llevado a cabo diferentes tareas con el propósito de contribuir a la investigación de la salud mental de los trabajadores tecnológicos, concretamente en el análisis de la prevalencia de enfermedades mentales y en la detección de los principales factores de riesgo para el padecimiento de enfermedades mentales –tanto personales como laborales- de este tipo de trabajadores. A continuación, se detallan las principales contribuciones.

En primer lugar, se ha llevado a cabo un análisis del estado del arte de la ciencia de datos en la salud mental. Para ello, se han analizado los siguientes aspectos: (i) la salud mental: una introducción, clasificación de las principales enfermedades mentales y una breve explicación de cada una de ellas; (ii) los avances más destacables que se han llevado a cabo en los últimos años en el campo de la ciencia de datos con respecto a la salud mental; (iii) estudios recientes de identificación de factores de riesgo y predicción de enfermedades mediante cuestionarios.

Una vez analizado el estado del arte, se ha construido un modelo de clasificación con el objetivo de detectar los factores tanto personales como laborales de los trabajadores tecnológicos que mayor impacto tienen en el padecimiento de alguna enfermedad mental. Para ello, en primer lugar, se ha analizado el conjunto de datos empleado en la realización del trabajo y diferentes modelos de aprendizaje supervisado para tareas de clasificación. Una vez seleccionado el modelo que más se adecua a los objetivos del trabajo, se ha hecho una selección y una preparación de los datos. Concretamente, se han corregido algunas variables del conjunto de datos, se ha hecho un tratamiento de *missing values* y, finalmente, se ha llevado a cabo una tarea de transformación de los datos para poder emplearlos en el modelo escogido.

Una vez tratados los datos, se ha construido el modelo seleccionado. Más específicamente, tras haber analizado en la selección del modelo los parámetros más comunes a modificar y haber estudiado el efecto que causa su modificación, se han creado varios modelos con diferentes parámetros para detectar el que mejores resultados proporciona. Para ello, se ha llevado a cabo una evaluación e interpretación de los resultados.

Finalmente, se realizado un análisis visual de la prevalencia de enfermedades mentales entre trabajadores tecnológicos desde diferentes puntos de vista y, por otro lado, un análisis visual del modelo creado anteriormente. El resultado obtenido se ha contrastado con un modelo creado mediante el software Orange.

En cuanto a los logros de los objetivos planteados al comienzo del trabajo, se considera que se han logrado parcialmente. Por un lado, el análisis de la prevalencia de enfermedades mentales en trabajadores tecnológicos ha sido satisfactorio y se ha

llegado a conclusiones relevantes. Por otro lado, en la tarea de detección de factores de riesgo, a pesar de haber tenido que codificar las variables categóricas, con ambas técnicas de codificación se ha llegado a conclusiones similares. Sin embargo, la interpretación visual del modelo ha sido muy compleja como consecuencia de la codificación de las variables y no se ha podido llegar unas conclusiones claras.

Esto ha sido una consecuencia del hecho de que la librería empleada para llevar a cabo el trabajo (Scikit Learn) no permitía trabajar con datos categóricos directamente, sino que han tenido que ser codificados. Puesto que la mayoría de las variables del conjunto de datos de partida eran variables categóricas por ser respuestas a un cuestionario, se han tenido que codificar la mayoría de ellas.

El mayor error en este trabajo ha sido suponer que una de las librerías más comunes en la programación de modelos de aprendizaje automático de Python iba a ser lo suficientemente flexible y completa como para poder llevar a cabo el trabajo planteado inicialmente. En relación con esto, la principal lección aprendida mediante la realización de este trabajo es que no se debe dar nada por hecho y, antes de comenzar a programar es necesario cerciorarse que las herramientas seleccionadas cumplen con los requisitos necesarios.

Independientemente de esto, se ha detectado una carencia de librerías de Python que permitan crear árboles de decisión con variable categóricas. La mayoría de las librerías encontradas o bien no admiten datos categóricos directamente como ocurre con Scikit Learn, o bien no aceptan datos categóricos pero la propia función de codificación los transforma a numéricos tal y como se ha hecho manualmente en este trabajo.

Sin embargo, sí que se ha encontrado una librería que Python que permite crear árboles de decisión con variables categóricas directamente: Orange. Además de ser una librería de Python es un programa informático que permite llevar a cabo procesos completos de análisis de datos sin tener que programar. De hecho, lo que la librería Orange permite es manipular los componentes de este programa utilizando Python. Cabe destacar que la última versión de esta librería (la versión 3), la cual utiliza Python 3, no tiene implementada la función de graficar árboles de decisión, por lo que en caso de que se quisiese emplear esta librería para mostrar árboles de decisión se tendría que utilizar la versión 2, la cual utiliza Python 2.

En cuanto a la planificación planteada inicialmente, se había propuesto desarrollar el trabajo siguiendo el flujo de trabajo clásico de un proyecto de análisis de datos y se ha cumplido fielmente. Con respecto a la metodología, se propuso seguir la metodología ágil CRISP-DM, enfocado principalmente en crear un flujo de trabajo bidireccional en el paso de la preparación de los datos al modelado de los mismos. Esta bidireccionalidad se ha dado en el trabajo y ha sido completamente necesario, ya que una vez seleccionado el modelo más adecuado a las necesidades del trabajo y

analizadas las limitaciones de la librería seleccionada, ha sido necesario retroceder a la preparación de los datos para modificarlos de tal forma que estuviesen en el formato correcto.

Finalmente, en cuanto a las líneas futuras del trabajo, una de ellas puede ser incluir en el modelo predictivo todas las variables subjetivas para analizar el problema también desde el punto de vista de la subjetividad de los trabajadores. Con el objetivo de simplificar el trabajo, se descartaron también todas aquellas variables recogían datos en formato de texto libre, por lo que podría ser interesante también procesar estas variables mediante técnicas de Procesamiento del Lenguaje Natural (PLN) e incluirlas en el modelo. También podría ser enriqueceros realizar los modelos separando las variables en factores personales y factores laborales. Para terminar, de cara a crear un producto final que lo puedan utilizar diferentes usuarios para extraer conocimiento sobre este tema, se podría crear una aplicación web con, por un lado, el análisis visual de prevalencia de enfermedades mentales y, por otro lado, los resultados del modelo.

5. Glosario

- **Análisis de datos**: proceso de inspección, limpieza, transformación y modelado de los datos con el objetivo de descubrir información útil, extraer conclusiones y dar soporte a la toma de decisiones.
- **Análisis visual**: la ciencia del razonamiento analítico facilitada por interfaces visuales.
- **Aprendizaje automático**: método de análisis de datos que automatiza la construcción de modelos analíticos. Es una rama de la inteligencia artificial basada en la idea de que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones con mínima intervención humana.
- **Aprendizaje no supervisado**: técnica de aprendizaje automático empleada cuando únicamente se dispone de variables de entrada (x) en el conjunto de datos y no variables de salida y cuyo objetivo principal es modelar la distribución de los datos.
- **Aprendizaje supervisado**: técnica de aprendizaje automático empleada cuando los datos están formados por variables de entrada (x) y una variable de salida o resultado (Y) y cuyo objetivo principal es crear una función que indique la relación entre la variable de salida y las diferentes variables de entrada.
- **Árbol de decisión**: modelo de predicción de aprendizaje supervisado en el cual dado un conjunto de datos se fabrican diagramas de construcciones lógicas que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.
- **Bosques aleatorios**: algoritmo de aprendizaje supervisado en el cual se selecciona de manera aleatoria una cantidad de variables con las cuales se construyen árboles de decisión individuales y se fusionan para obtener una predicción más precisa y estable.
- **Ciencia de datos**: campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados.
- **Conjunto de entrenamiento**: subconjunto de datos empleado para construir el modelo.
- **Conjunto de prueba**: subconjunto de los datos utilizado para probar el modelo creado con el conjunto de entrenamiento.
- **Enfermedad mental**: alteración de tipo emocional, cognitivo y/o comportamiento, en que quedan afectados procesos psicológicos básicos como son la emoción, la motivación, la cognición, la conciencia, la conducta, la percepción, la sensación, el

aprendizaje, el lenguaje, etc. Lo que dificulta a la persona su adaptación al entorno cultural y social en que vive y crea alguna forma de malestar subjetivo.

- **Factor de riesgo**: toda circunstancia o situación que aumenta las probabilidades de una persona de contraer una enfermedad o cualquier otro problema de salud.
- ***Label Encoder****:* técnica de codificación de variables categóricas que consiste en asignar a cada una de las categorías posibles un número comenzando desde el 1.
- **Minería de datos**: proceso de descubrimiento patrones y relaciones en grandes cantidades de datos.
- ***One Hot Encoding***: técnica de codificación de variables categóricas que consiste en dividir cada una de las columnas categóricas del conjunto de datos en tantas columnas como posibles categorías y a cada una de estas columnas, asignarle un 1 en caso de que el valor que tome la observación sea el correspondiente a esa columna y un 0 en caso de que no.
- **Poda**: técnica de reducción del tamaño del árbol de decisión mediante la eliminación de hojas que proporcionan poca capacidad de clasificar instancias.
- ***Post-prunning***: técnica empleada para evitar el sobreajuste de los datos en los árboles de decisión y que consiste en permitir al árbol que crezca en su totalidad sin ninguna restricción de tamaño para, una vez completado, podarlo.
- **Precisión**: medida de calidad de un modelo de clasificación calculado como la proporción de operaciones de clasificación correctas que ha sido capaz de obtener el modelo.
- **Prevalencia**: proporción de individuos de un grupo o una población que presentan una característica o evento determinado en un momento o en un período determinado.
- ***Pre-prunning***: técnica empleada para evitar el sobreajuste de los datos en los árboles de decisión y que consiste en parar el crecimiento del árbol antes de que haya crecido completamente.
- **Salud mental**: estado de bienestar en el cual el individuo es consciente de sus propias capacidades, puede afrontar las tensiones normales de la vida, puede trabajar de forma productiva y fructífera y es capaz de hacer una contribución a su comunidad.
- **Sobreajuste**: incapacidad de un modelo de aprendizaje automático de generalizar correctamente.
- **Trabajador tecnológico**: persona que desarrolla su trabajo en la industria tecnológica.

6. Bibliografía

[1] Organización Mundial de la Salud (OMS), *Invertir en Salud Mental*. 2004.

[2] H. M. L. Castro, «Estigma y enfermedad mental: un punto de vista histórico-social», p. 10.

[3] S. Leka y A. Jain, «Mental Health in the Workplace in Europe», p. 43.

[4] *Researching Information Systems and Computing - Briony J Oates - Google Libros*. .

[5] «Welcome to Python.org». [En línea]. Disponible en: https://www.python.org/. [Accedido: 19-may-2019].

[6] «Python Data Analysis Library — pandas: Python Data Analysis Library». [En línea]. Disponible en: https://pandas.pydata.org/. [Accedido: 01-jun-2019].

[7] «scikit-learn: machine learning in Python — scikit-learn 0.21.2 documentation». [En línea]. Disponible en: https://scikit-learn.org/stable/. [Accedido: 01-jun-2019].

[8] *Matplotlib: Python plotting — Matplotlib 3.0.3 documentation*. .

[9] «OSMI Mental Health in Tech Survey 2016 | Kaggle». [En línea]. Disponible en: https://www.kaggle.com/osmi/mental-health-in-tech-2016. [Accedido: 27-abr-2019].

[10] Ramon Sangüesa i Solé, *El proceso de descrubimiento a partir de los datos*. Universidad Oberta de Catalunya (UOC).

[11] Rodríguez y Díaz, *Metodología para el Desarrollo de Proyectos de Minería de Datos CRISP-DM*. 2004.

[12] D. M. Bielostotzky, «Ingeniería de Requerimientos en Proyectos de Ciencia de Datos.», *Medium*, 18-feb-2019. .

[13] «OMS | Salud mental: un estado de bienestar», *WHO*. [En línea]. Disponible en: https://www.who.int/features/factfiles/mental_health/es/. [Accedido: 23-mar-2019].

[14] «Mental health: data and resources», *WHO Europe*, 23-mar-2019. [En línea]. Disponible en: http://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health/data-and-resources. [Accedido: 23-mar-2019].

[15] «OMS | 10 datos sobre la salud mental», *WHO*. [En línea]. Disponible en: https://www.who.int/features/factfiles/mental_health/mental_health_facts/es/. [Accedido: 23-mar-2019].

[16] «OMS | Plan de acción sobre salud mental 2013-2020».

[17] H. M. L. Castro, «ESTIGMA Y ENFERMEDAD MENTAL: UN PUNTO DE VISTA HISTORICO-SOCIAL», p. 10.

[18] S. Leka y A. Jain, «MENTAL HEALTH IN THE WORKPLACE», p. 40.

[19] «International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) Version for 2010». [En línea]. Disponible en: https://web.archive.org/web/20140622030723/http://apps.who.int/classifications/icd10/browse/2010/en#/V. [Accedido: 23-mar-2019].

[20] American Psyquiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM–5)*. .

[21] «OMS | Salud mental en el lugar de trabajo», *WHO*. [En línea]. Disponible en: http://www.who.int/mental_health/in_the_workplace/es/. [Accedido: 23-mar-2019].

[22] «Salud laboral | ISTAS». [En línea]. Disponible en: https://istas.net/salud-laboral. [Accedido: 23-mar-2019].

[23] Michael A. Freeman, «Are Entrepreneurs “Touched with Fire”?», 17-abr-2015.

[24] Laura A. Pratt y Debra J. Brody, «NCHS Data Brief - Depression in the U.S. Household Population, 2009-2012», No. 172, dic. 2014.

[25] J. F. Dipnall *et al.*, «Fusing Data Mining, Machine Learning and Traditional Statistics to Detect Biomarkers Associated with Depression», *PLOS ONE*, vol. 11, n.º 2, p. e0148195, feb. 2016.

[26] S. G. Alonso *et al.*, «Data Mining Algorithms and Techniques in Mental Health: A Systematic Review», *J. Med. Syst.*, vol. 42, n.º 9, p. 161, jul. 2018.

[27] B. L. Quéau, O. Shafiq, y R. Alhajj, «Analyzing Alzheimer's disease gene expression dataset using clustering and association rule mining», en *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, 2014, pp. 283-290.

[28] R. GeethaRamani y K. Sivaselvi, «Data mining technique for identification of diagnostic biomarker to predict Schizophrenia disorder», en *2014 IEEE International Conference on Computational Intelligence and Computing Research*, 2014, pp. 1-8.

[29] C. M. Corcoran *et al.*, «Prediction of psychosis across protocols and risk cohorts using automated language analysis», *World Psychiatry*, vol. 17, n.º 1, pp. 67-75, 2018.

[30] «Aifred Health». [En línea]. Disponible en: https://aifredhealth.com/index.html. [Accedido: 23-mar-2019].

[31] Q. Health, «Quartet Health», *Quartet Health*. [En línea]. Disponible en: https://www.quartethealth.com/. [Accedido: 23-mar-2019].

[32] P. A. Idowu, «A Predictive Model for the Risk of Mental Illness in Nigeria Using Data Mining».

[33] «Weka 3 - Data Mining with Open Source Machine Learning Software in Java». [En línea]. Disponible en: https://www.cs.waikato.ac.nz/ml/weka/. [Accedido: 23-mar-2019].

[34] Y. Kuroki, «Risk factors for suicidal behaviors among Filipino Americans: a data mining approach», *Am. J. Orthopsychiatry*, vol. 85, n.º 1, pp. 34-42, ene. 2015.

[35] D. Takeuchi, «Filipino American Community Epidemiological Study (FACES), 1995-1999: Version 1». Inter-university Consortium for Political and Social Research, 08-ago-2011.

[36] S. YOON, B. TAHA, y S. BAKKEN, «Using a Data Mining Approach to Discover Behavior Correlates of Chronic Disease: A Case Study of Depression», *Stud. Health Technol. Inform.*, vol. 201, pp. 71-78, 2014.

[37] «CDC - BRFSS». [En línea]. Disponible en: https://www.cdc.gov/brfss/. [Accedido: 23-mar-2019].

[38] Patel, P., «Perceived Workplace Factors and their Influence on Self-Reported Mental Health Service Seeking Among Technology Workers», oct. 2018.

[39] «OSMI Home :: Open Sourcing Mental Illness - Changing how we talk about mental health in the tech community - Stronger Than Fear». [En línea]. Disponible en: https://osmihelp.org/. [Accedido: 19-may-2019].

[40] «Kaggle: Your Home for Data Science». [En línea]. Disponible en: https://www.kaggle.com/. [Accedido: 19-may-2019].

[41] «Is Python the most popular language for data science? - Maruti Techlabs». [En línea]. Disponible en: https://www.marutitech.com/python-data-science/. [Accedido: 19-may-2019].

[42] «Project Jupyter | Home». [En línea]. Disponible en: https://jupyter.org/. [Accedido: 19-may-2019].

[43] «Jupyter notebook: documenta y ejecuta código desde el navegador». [En línea]. Disponible en: https://blog.desdelinux.net/jupyter-notebook/. [Accedido: 19-may-2019].

[44] «Supervised and Unsupervised Machine Learning Algorithms». [En línea]. Disponible en: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/. [Accedido: 27-abr-2019].

[45] «Types of Machine Learning Algorithms You Should Know». [En línea]. Disponible en: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861. [Accedido: 27-abr-2019].

[46] «7 Types of Classification Algorithms - Analytics India Magazine». [En línea]. Disponible en: https://www.analyticsindiamag.com/7-types-classification-algorithms/. [Accedido: 27-abr-2019].

[47] «Essential Classification Algorithms Explained | Kaggle». [En línea]. Disponible en: https://www.kaggle.com/anniepyim/essential-classification-algorithms-explained. [Accedido: 27-abr-2019].

[48] «Supervised Machine Learning: Classification – Towards Data Science». [En línea]. Disponible en: https://towardsdatascience.com/supervised-machine-learning-classification-5e685fe18a6d. [Accedido: 27-abr-2019].

[49] «KNN Classification using Scikit-learn (article) - DataCamp». [En línea]. Disponible en: https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn. [Accedido: 19-may-2019].

[50] «Digg Data • Support Vector Machine without tears». [En línea]. Disponible en: https://diggdata.in/post/94066544971/support-vector-machine-without-tears. [Accedido: 19-may-2019].

[51] «1.10. Decision Trees — scikit-learn 0.21.1 documentation». [En línea]. Disponible en: https://scikit-learn.org/stable/modules/tree.html. [Accedido: 19-may-2019].

[52] Ramon Sangüesa i Solé, *Clasificación: árboles de decisión*. Universidad Oberta de Catalunya (UOC).

[53] «Handbook of Statistical Analysis and Data Mining Applications - 1st Edition». [En línea]. Disponible en: https://www.elsevier.com/books/handbook-of-statistical-analysis-and-data-mining-applications/nisbet/978-0-12-374765-5. [Accedido: 08-jun-2019].

[54] Roman Timofeev, «Classification and Regression Trees (CART) Theory and Applications».

[55] «Decision Tree Classification in Python (article) - DataCamp». [En línea]. Disponible en: https://www.datacamp.com/community/tutorials/decision-tree-classification-python. [Accedido: 19-may-2019].

[56] «Overfitting of decision tree and tree pruning, How to avoid overfitting in data mining», *T4Tutorials*. [En línea]. Disponible en: https://t4tutorials.com/overfitting-of-decision-tree-and-tree-pruning-in-data-mining/. [Accedido: 02-jun-2019].

[57] N. S. Patel y S. Upadhyay, «Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA», 2012.

[58] «Decision tree pruning», *Wikipedia*. 17-mar-2019.

[59] «Machine Learning in Python - PyImageSearch». [En línea]. Disponible en: https://www.pyimagesearch.com/2019/01/14/machine-learning-in-python/. [Accedido: 19-may-2019].

[60] «1.11. Ensemble methods — scikit-learn 0.21.1 documentation». [En línea]. Disponible en: https://scikit-learn.org/stable/modules/ensemble.html#forest. [Accedido: 19-may-2019].

[61] M. B. Fraj, «In Depth: Parameter tuning for Random Forest», *All things AI*, 21-dic-2017. .

[62] Luis Carlos Molina Félix y Ramon Sangüesa i Solé, *Evaluación de modelos*. Universidad Oberta de Catalunya (UOC).

[63] «How to split a dataset – Beyond the lines». [En línea]. Disponible en: https://www.beyondthelines.net/machine-learning/how-to-split-a-dataset/. [Accedido: 08-jun-2019].

[64] «Orange – Data Mining Fruitful & Fun». [En línea]. Disponible en: https://orange.biolab.si/. [Accedido: 19-may-2019].

[65] «What is Exploratory Data Analysis? – Towards Data Science». [En línea]. Disponible en: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15. [Accedido: 19-may-2019].

[66] R. Shaikh, «Choosing the right Encoding method-Label vs OneHot Encoder», *Towards Data Science*, 09-nov-2018. [En línea]. Disponible en: https://towardsdatascience.com/choosing-the-right-encoding-method-label-vs-onehot-encoder-a4434493149b. [Accedido: 29-abr-2019].

[67] S. Srinidhi, «Label Encoder vs. One Hot Encoder in Machine Learning», *Medium*, 30-jul-2018. .

[68] «sklearn.preprocessing.OneHotEncoder — scikit-learn 0.20.3 documentation». [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. [Accedido: 29-abr-2019].

[69] «Por qué es importante trabajar con datos balanceados para clasificación». [En línea]. Disponible en: http://amsantac.co/blog/es/2016/09/20/balanced-image-classification-r-es.html. [Accedido: 19-may-2019].

[70] «What is a good classification accuracy in data mining? | Data Mining Blog - www.dataminingblog.com». [En línea]. Disponible en: http://www.dataminingblog.com/what-is-a-good-classification-accuracy-in-data-mining/. [Accedido: 08-jun-2019].

[71] «The state of visual analytics: Views on what visual analytics is and where it is going | UMIACS». [En línea]. Disponible en: https://www.umiacs.umd.edu/publications/state-visual-analytics-views-what-visual-analytics-and-where-it-going. [Accedido: 19-may-2019].

[72] «Better Understand Your Data in R Using Visualization (10 recipes you can use today)». [En línea]. Disponible en: https://machinelearningmastery.com/data-visualization-in-r/. [Accedido: 19-may-2019].

[73] «Roscoe, J.T. (1975) Fundamental Research Statistics for the Behavioral Sciences [by] John T. Roscoe. Holt, Rinehart and Winston, New York. - References - Scientific Research Publishing». [En línea]. Disponible en: https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1691933. [Accedido: 08-jun-2019].

7. Anexos

Anexo I. Encuesta OSMI 2016

1. Are you self-employed?
    - Yes
    - No

1. How many employees does your company or organization have?
    - 1 - 5
    - 6 – 25
    - 26 – 100
    - 100 – 500
    - 500 - 1000
    - More than 1000

2. Is your employer primarily a tech company/organization?
    - Yes
    - No

3. Is your primary role within your company related to tech/IT?
    - Yes
    - No

4. Does your employer provide mental health benefits as part of healthcare coverage?
    - Yes
    - No
    - I don’t know

5. Do you know the options for mental health care available under your employer-provided coverage?
    - Yes
    - No
    - I am not sure

6. Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?
    - Yes
    - No
    - I don't know

7. Does your employer offer resources to learn more about mental health concerns and options for seeking help?
    - Yes
    - No
    - I don't know

8. Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?
    - Yes
    - No
    - I don't know

9. If a mental health issue prompted you to request a medical leave from work, asking for that leave would be:
    - Very easy
    - Somewhat easy
    - Neither easy nor difficult
    - Somewhat difficult
    - Very difficult
    - I don't know

10. Do you think that discussing a mental health disorder with your employer would have negative consequences?
    - Yes
    - No
    - Maybe

11. Do you think that discussing a physical health issue with your employer would have negative consequences?
    - Yes

- No
- Maybe

12. Would you feel comfortable discussing a mental health disorder with your coworkers?
    - Yes
    - No
    - Maybe

13. Would you feel comfortable discussing a mental health disorder with your direct supervisor(s)?
    - Yes
    - No
    - Maybe

14. Do you feel that your employer takes mental health as seriously as physical health?
    - Yes
    - No
    - I don't know

15. Have you heard of or observed negative consequences for co-workers who have been open about mental health issues in your workplace?
    - Yes
    - No

16. Do you have medical coverage (private insurance or state-provided) which includes treatment of mental health issues?
    - Yes
    - No

17. Do you know local or online resources to seek help for a mental health disorder?
    - Yes, I know several
    - No, I don't know any
    - I know some

18. If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to clients or business contacts?
    - Yes, always
    - Sometimes, if it comes up
    - No, because it doesn't matter
    - No, because it would impact me negatively
    - Not applicable to me

19. If you have revealed a mental health issue to a client or business contact, do you believe this has impacted you negatively?
    - Yes
    - No
    - I am not sure

20. If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to coworkers or employees?
    - Yes, always
    - Sometimes, if it comes up
    - No, because it doesn't matter
    - No, because it would impact me negatively
    - Not applicable to me

21. If you have revealed a mental health issue to a coworker or employee, do you believe this has impacted you negatively?
    - Yes
    - No
    - I'm not sure
    - Not applicable to me

22. Do you believe your productivity is ever affected by a mental health issue?
    - Yes
    - No
    - Unsure
    - Not applicable to me

23. If yes, what percentage of your work time (time performing primary or secondary job functions) is affected by a mental health issue?
    - 1 – 25%
    - 26 – 50%
    - 51 – 75%
    - 76 – 100%

24. Do you have previous employers?
    - Yes
    - No

25. Have your previous employers provided mental health benefits?
    - Yes, they all did
    - Some did
    - No, none did
    - I don't know

26. Were you aware of the options for mental health care provided by your previous employers?
    - Yes, I was aware of all of them
    - I was aware of some
    - No, I only became aware later

27. Did your previous employers ever formally discuss mental health (as part of a wellness campaign or other official communication)?
    - Yes, they all did
    - Some did
    - None did
    - I don't know

28. Did your previous employers provide resources to learn more about mental health issues and how to seek help?
    - Yes, they all did
    - Some did
    - None did

29. Was your anonymity protected if you chose to take advantage of mental health or substance abuse treatment resources with previous employers?
    - Yes, always
    - Sometimes
    - No
    - I don't know

30. Do you think that discussing a mental health disorder with previous employers would have negative consequences?
    - Yes, all of them
    - Some of them
    - None of them
    - I don't know

31. Do you think that discussing a physical health issue with previous employers would have negative consequences?
    - Yes, all of them
    - Some of them
    - None of them

32. Would you have been willing to discuss a mental health issue with your previous co-workers?
    - Yes, at all of my previous employers
    - Some of my previous employers
    - No, at none of my previous employers

33. Would you have been willing to discuss a mental health issue with your direct supervisor(s)?
    - Yes, at all of my previous employers
    - Some of my previous employers
    - No, at none of my previous employers
    - I don't know

34. Did you feel that your previous employers took mental health as seriously as physical health?
    - Yes, they all did

- Some did
- None did
- I don't know

35. Did you hear of or observe negative consequences for co-workers with mental health issues in your previous workplaces?
    - Yes, all of them
    - Some of them
    - None of them

36. Would you be willing to bring up a physical health issue with a potential employer in an interview?
    - Yes
    - No
    - Maybe

37. Why or why not? ……..
38. Would you bring up a mental health issue with a potential employer in an interview?
    - Yes
    - No
    - Maybe

39. Why or why not? ……..
40. Do you feel that being identified as a person with a mental health issue would hurt your career?
    - Yes, it has
    - Yes, I think it would
    - Maybe
    - No, I don't think it would
    - No, it has not

41. Do you think that team members/co-workers would view you more negatively if they knew you suffered from a mental health issue?
    - Yes, I think they would
    - Yes, they do

- Maybe
- No, I don't think they would
- No, they do not

42. How willing would you be to share with friends and family that you have a mental illness?
    - Somewhat open
    - Very open
    - Somewhat not open
    - Neutral
    - Not applicable to me (I do not have a mental illness)
    - Not open at all

43. Have you observed or experienced an unsupportive or badly handled response to a mental health issue in your current or previous workplace?
    - Yes, I observed
    - Yes, I experienced
    - Maybe/Not sure
    - No

44. Have your observations of how another individual who discussed a mental health disorder made you less likely to reveal a mental health issue yourself in your current workplace?
    - Yes
    - Maybe
    - No

45. Do you have a family history of mental illness?
    - Yes
    - No
    - I don't know

46. Have you had a mental health disorder in the past?
    - Yes
    - Maybe
    - No

47. Do you currently have a mental health disorder?

- Yes
- Maybe
- No

48. If yes, what condition(s) have you been diagnosed with?

- Mood Disorder (Depression, Bipolar Disorder, etc)
- Anxiety Disorder (Generalized, Social, Phobia, etc)
- Attention Deficit Hyperactivity Disorder
- Post-traumatic Stress Disorder
- Obsessive-Compulsive Disorder
- Substance Use Disorder
- Personality Disorder (Borderline, Antisocial, Paranoid, etc)
- Stress Response Syndromes
- Addictive Disorder
- Eating Disorder (Anorexia, Bulimia, etc)
- Dissociative Disorder
- Psychotic Disorder (Schizophrenia, Schizoaffective, etc)
- Other

49. If maybe, what condition(s) do you believe you have?

- Mood Disorder (Depression, Bipolar Disorder, etc)
- Anxiety Disorder (Generalized, Social, Phobia, etc)
- Attention Deficit Hyperactivity Disorder
- Post-traumatic Stress Disorder
- Obsessive-Compulsive Disorder
- Substance Use Disorder
- Personality Disorder (Borderline, Antisocial, Paranoid, etc)
- Stress Response Syndromes
- Addictive Disorder
- Eating Disorder (Anorexia, Bulimia, etc)
- Dissociative Disorder
- Psychotic Disorder (Schizophrenia, Schizoaffective, etc)
- Other

50. Have you been diagnosed with a mental health condition by a medical professional?
    - Yes
    - No

51. If so, what condition(s) were you diagnosed with?
    - Mood Disorder (Depression, Bipolar Disorder, etc)
    - Anxiety Disorder (Generalized, Social, Phobia, etc)
    - Attention Deficit Hyperactivity Disorder
    - Post-traumatic Stress Disorder
    - Obsessive-Compulsive Disorder
    - Substance Use Disorder
    - Personality Disorder (Borderline, Antisocial, Paranoid, etc)
    - Stress Response Syndromes
    - Addictive Disorder
    - Eating Disorder (Anorexia, Bulimia, etc)
    - Dissociative Disorder
    - Psychotic Disorder (Schizophrenia, Schizoaffective, etc)
    - Other

52. Have you ever sought treatment for a mental health issue from a mental health professional?
    - Yes
    - No

53. If you have a mental health issue, do you feel that it interferes with your work when being treated effectively?
    - Not applicable to me
    - Sometimes
    - Rarely
    - Never
    - Ofter

54. If you have a mental health issue, do you feel that it interferes with your work when NOT being treated effectively?
    - Not applicable to me

- Sometimes
- Rarely
- Never
- Ofter

55. What is your age? …….
56. What is your gender? …….
57. What country do you live in? …….
58. What US state or territory do you live in? …….
59. What country do you work in? …….
60. What US state or territory do you work in? …….
61. Which of the following best describes your work position?
    - Back-end Developer
    - Front-end Developer
    - Supervisor/Team Lead
    - DevOps/SysAdmin
    - Other
    - Support
    - One-person shop
    - Designer
    - Dev Evangelist/Advocate
    - Executive Leadership
    - Sales
    - HR
62. Do you work remotely?
    - Sometimes
    - Never
    - Always

Anexo II. Clasificación CIE – 10 completa de enfermedades mentales

(F00-F09) Trastornos mentales orgánicos, incluidos los trastornos sintomáticos

- (F00) Demencia en la enfermedad de Alzheimer
- (F01) Demencia vascular
    - (F01.1) Demencia multi-infartica
- (F02) Demencia en otras enfermedades clasificadas
    - (F02.0) Demencia en la enfermedad de Pick
    - (F02.1) Demencia en la enfermedad de Creutzfeldt-Jakob
    - (F02.2) Demencia en la enfermedad de Huntington
    - (F02.3) Demencia en la enfermedad de Parkinson
    - (F02.4) Demencia en el VIH
- (F03) Demencia sin especificar
- (F04) Síndrome amnésico orgánico, no inducido por alcohol o por otros psicotrópicos
- (F05) Delírium, no inducido por alcohol o por otros psicotrópicos
- (F06) Otros trastornos mentales debidos a daños neuronales, disfunciones y enfermedades físicas
- (F06.0) Alucinosis Orgánica
- (F06.1) Trastorno catatónico, Orgánico
- (F06.2) Trastorno delirante [esquizofreniforme], Orgánico.
- (F06.3) Trastornos del humor [afectivos], Orgánico
- (F06.4) Trastorno de ansiedad, Orgánico
- (F06.5) Trastorno disociativo, Orgánico
- (F07) Trastornos de personalidad y comportamiento debido a enfermedades neuronales, daños y disfunciones
- (F08) Trastorno narcisista de la personalidad.

- (F09) Trastornos mentales orgánicos o sintomáticos sin especificar

(F10-F19) Trastornos mentales y de comportamiento debidos al consumo de psicotrópicos

- Nota: las siguientes condiciones son subtipos de cada código en el intervalo F10-19:
    - (F1x.0) Intoxicación aguda
    - (F1x.1) Uso perjudicial
    - (F1x.2) Síndrome de dependencia
    - (F1x.3) Síndrome de abstinencia
    - (F1x.4) síndrome de abstinencia con delirium
    - (F1x.5) Trastorno psicótico
    - (F1x.6) Trastorno Amnésico
    - (F1x.7) Trastorno psicótico residual
    - (F1x.8) Otro trastorno mental del comportamiento.
    - (F1x.9) Trastorno mental o del comportamiento no especificado.
- (F10) Trastornos mentales y de comportamiento debidos al consumo de alcohol
- (F11) Trastornos mentales y de comportamiento debidos al consumo de opioides
- (F12) Trastornos mentales y de comportamiento debidos al consumo de cannabinoides
- (F13) Trastornos mentales y de comportamiento debidos al consumo de sedantes o hipnóticos
- (F14) Trastornos mentales y de comportamiento debidos al consumo de cocaína
- (F15) Trastornos mentales y de comportamiento debidos al consumo de otros estimulantes, incluyendo la cafeína
- (F16) Trastornos mentales y de comportamiento debidos al consumo de alucinógenos
- (F17) Trastornos mentales y de comportamiento debidos al consumo de tabaco

- (F18) Trastornos mentales y de comportamiento debidos al consumo de disolventes volátiles
- (F19) Trastornos mentales y de comportamiento debidos al consumo de múltiples drogas y otros psicotrópicos

(F20-29) Esquizofrenia, trastornos esquizotípicos y trastornos delirantes

- (F20) Esquizofrenia
    - (F20.0) Esquizofrenia paranoide
    - (F20.1) Esquizofrenia hebefrénica
    - (F20.2) Esquizofrenia catatónica
    - (F20.3) Esquizofrenia indiferenciada
    - (F20.4) Depresión post-esquizofrénica
    - (F20.5) Esquizofrenia residual
    - (F20.6) Esquizofrenia simple
    - (F20.8) Otras esquizofrenias
    - (F20.9) Esquizofrenia no especificada
- (F21) Trastorno esquizotípico
- (F22) Trastornos delirantes persistentes
    - (F22.0) Trastorno delirante
- (F23) Trastornos psicóticos agudos y transitorios
    - (F23.0) Trastorno psicótico polimórfico agudo sin síntomas de esquizofrenia
    - (F23.1) Trastorno psicótico polimórfico agudo con síntomas de esquizofrenia
    - (F23.2) Trastorno psicótico agudo estilo esquizofrenia
    - (F23.3) Otros trastornos psicóticos agudos predominantemente delirantes
    - (F23.8) Otros trastornos psicóticos agudos y transitorios
    - (F23.9) Trastornos psicóticos agudo y transitorios sin especificar
- (F24) Trastorno de ideas delirantes inducidas

- (F25) Trastornos esquizoafectivos
    - (F25.0) Trastorno esquizoafectivo, tipo maníaco
    - (F25.1) Trastorno esquizoafectivo, tipo depresivo
    - (F25.2) Trastorno esquizoafectivo, tipo mixto
    - (F25.8) Otros trastornos esquizoafectivos
    - (F25.9) Trastorno esquizoafectivo sin especificar
- (F28) Otros trastornos psicóticos no orgánicos
- (F29) Psicosis no orgánica sin especificar

(F30-39) Trastornos del humor (afectivos)

- (F30) Episodio maníaco
    - (F30.0) Hipomanía
- (F31) Trastorno bipolar afectivo
- (F32) Episodio depresivo
    - (F32.0) Episodio depresivo leve
    - (F32.1) Episodio depresivo moderado
- (F33) Trastorno depresivo recurrente
    - (F33.0) Trastorno depresivo recurrente, episodio actual leve
    - (F33.1) Trastorno depresivo recurrente, episodio actual moderado
    - (F33.2) Trastorno depresivo recurrente, episodio actual grave sin síntomas psicóticos
    - (F33.3) Trastorno depresivo recurrente, episodio actual grave con síntomas psicóticos
    - (F33.4) Trastorno depresivo recurrente actualmente en remisión
- (F34) Trastornos afectivos persistentes
    - (F34.0) Ciclotimia
    - (F34.1) Distimia
- (F38) Otros trastornos afectivos

- (F39) Trastorno afectivo sin especificar

(F40-49) Trastornos neuróticos, trastornos relacionados con el estrés y trastornos somatomorfos

- (F40) Trastornos fóbicos de ansiedad
    - (F40.0) Agorafobia
- (F41) Otros trastornos de ansiedad
    - (F41.0) Trastorno de pánico (ansiedad episódica paroxismal)
    - (F41.1) Trastorno de ansiedad generalizada
    - (F41.2) Trastorno mixto ansioso-depresivo
- (F42) Trastorno obsesivo-compulsivo
- (F43) Reacción al stress grave y trastornos de adaptación
    - (F43.0) Reacción al stress agudo
    - (F43.1) Trastorno post-traumático del stress
    - (F43.2) Trastorno de adaptación
- (F44) Trastorno de conversión disociativo
    - (F44.0) Amnesia disociativa
    - (F44.1) Fuga disociativa
- (F45) Trastorno somatomorfo
    - (F45.0) Trastorno de somatización
- (F48) Otras neurosis
    - (F48.0) Neurastenia

(F50-59) Síndromes del comportamiento asociados con alteraciones fisiológicas y factores físicos

- (F50) Trastornos de la ingestión de alimentos
    - (F50.0) Anorexia nerviosa
    - (F50.2) Bulimia nerviosa
    - (F50.3) Bulimia nerviosa atípica

        - (F50.4) Hiperfagia asociada a otros trastornos psicológicos
        - (F50.5) Vómitos asociados a otros trastornos psicológicos
        - (F50.8) Otros trastornos de la conducta alimentaria
        - (F50.9) Trastornos de la conducta alimentaria no especificado
- (F51) Trastornos del sueño no orgánicos
    - (F51.0) Insomnio no orgánico
    - (F51.1) Hipersomnio no orgánico
    - (F51.2) Trastorno del reloj biológico no orgánico
    - (F51.3) Sonambulismo
    - (F51.4) Terror nocturno
    - (F51.5) Pesadillas
- (F52) Disfunción sexual no ocasionada por trastornos ni enfermedades orgánicas
    - (F52.4) Eyaculación precoz
    - (F52.8) Hipersexualidad
- (F53) Trastornos mentales y de comportamiento asociados con el puerperio no clasificados
    - (F53.0) Trastornos mentales suaves y de comportamiento asociados con el puerperio no clasificados
    - (F53.1) Trastornos mentales severos y de comportamiento asociados con el puerperio no clasificados
- (F54) Factores psicológicos y de comportamiento asociados con los desórdenes o enfermedades clasificados
- (F55) Abuso de sustancias que no producen dependencia
- (F59) Síndromes de comportamiento sin especificar asociados con perturbaciones psicológicas y factores físicos

(F60-69) Trastornos de la personalidad y del comportamiento en adultos

- (F60) Trastorno de personalidad específico
    - (F60.0) Trastorno paranoide de la personalidad

    - (F60.1) Trastorno esquizoide de la personalidad
    - (F60.2) Trastorno disocial de la personalidad
    - (F60.3) Trastorno de inestabilidad emocional de la personalidad
    - (F60.4) Trastorno histriónico de la personalidad
    - (F60.5) Trastorno anancástico de la personalidad
    - (F60.6) Trastorno ansioso o por evitación de la personalidad
    - (F60.7) Trastorno dependiente de la personalidad
    - (F60.8) Otros trastornos de personalidad específicos
    - (F60.9) Trastorno de personalidad, sin especificar
- (F61) Trastornos de personalidad mixtos y otros
- (F62) Cambios de personalidad duraderos, no atribuibles a enfermedades o daños cerebrales
- (F63) Trastornos impulsivos y de hábito
    - (F63.0) Ludopatía patológica
    - (F63.1) Piromanía patológica
    - (F63.2) Cleptomanía patológica
    - (F63.3) Tricotilomanía
    - (F65.0) Fetichismo
    - (F65.2) Exhibicionismo
    - (F65.3) Voyeurismo
    - (F65.4) Pedofilia
    - (F65.5) Sadomasoquismo
    - (F65.6) Múltiples trastornos de preferencia sexual
    - (F65.8) Otros trastornos de preferencia sexual
- (F66) Trastornos psicológicos y de comportamiento asociados con el desarrollo y la orientación sexual
    - (F66.0) Trastorno de maduración sexual

        - (F66.1) Orientación sexual egodistónica
        - (F66.2) Trastorno relacional sexual
        - (F66.8) Otros trastornos de la pulsión
        - (F66.9) Trastornos de la pulsión, sin especificar
    - (F68) Otros trastornos de la personalidad y el comportamiento en adultos
        - (F68.0) Elaboración de síntomas físicos por razones psicológicas
        - (F68.1) Producción intencionada o ficción de síntomas o incapacidades, físicas o psicológicas
        - (F68.8) Otros trastornos específicos de personalidad o comportamiento en adultos
    - (F69) Trastornos de la personalidad y el comportamiento en adultos sin especificar

(F70-79) Retraso mental

- (F70) Retraso mental leve
- (F71) Retraso mental moderado
- (F72) Retraso mental severo
- (F73) Retraso mental profundo
- (F78) Otros retrasos mentales
- (F79) Retrasos mentales sin especificar

(F80-89) Trastornos del desarrollo psicológico

- (F80) Trastornos específicos del lenguaje y del habla
    - (F80.0) Trastorno específico de la articulación del habla
    - (F80.1) Trastorno expresivo del lenguaje
    - (F80.2) Trastorno receptivo del lenguaje
    - (F80.3) Afasia adquirida con epilepsia (Landau-Kleffner)
    - (F80.8) Otros trastornos del desarrollo del lenguaje y el habla
    - (F80.9) Trastornos del desarrollo del lenguaje y el habla sin especificar

- (F81) Trastornos de desarrollo específicos de habilidades académicas
    - (F81.0) Trastorno específico de la lectura
    - (F81.1) Agrafía
    - (F81.2) Trastornos específicos de habilidades aritméticas
    - (F81.3) Trastornos mixtos de habilidades escolares
    - (F81.8) Otros desórdenes del desarrollo de habilidades escolares
    - (F81.9) Trastorno de desarrollo de habilidades escolares sin especificar
- (F82) Trastornos de desarrollo específicos de funciones motoras
- (F83) Trastornos de desarrollo específicos mixtos
- (F84) Trastorno generalizado del desarrollo
    - (F84.0) Autismo en la niñez
    - (F84.1) Autismo atípico
    - (F84.2) Síndrome de Rett
    - (F84.4) Trastorno asociado a hiperactividad con retraso mental y movimientos estereotipados
    - (F84.5) Síndrome de Asperger
- (F88) Otros trastornos del desarrollo psicológico
- (F89) Trastornos del desarrollo psicológico sin especificar

(F90-F98) Trastornos emocionales y del comportamiento que aparecen habitualmente en la niñez o en la adolescencia

---

- (F90) Trastornos hipercinéticos
    - (F90.0) Trastorno de la actividad y la atención
    - (F90.1) Trastorno hiperquinético de la conducta
- (F91) Trastornos de conducta
    - (F91.0) Trastorno de conducta confinado al entorno familiar
    - (F91.1) Trastorno de conducta dessocializado
    - (F91.2) Trastorno de conducta socializado

        - (F91.3) Trastorno negativista desafiante
- (F92) Trastornos mixtos de conducta y emociones
    - (F92.0) Trastornos de conducta depresivos
- (F93) Trastornos emocionales específicos en el comienzo de la niñez
    - (F93.0) Trastorno de ansiedad por separación de la niñez
    - (F93.1) Trastorno de ansiedad fóbica de la niñez
    - (F93.2) Trastorno de ansiedad social de la niñez
    - (F93.3) Trastorno de rivalidad fraternal
- (F94) Trastornos de funciones sociales específicos del comienzo de la niñez y la adolescencia
    - (F94.0) Mutismo selectivo
    - (F94.1) Trastorno del vínculo reactivo de la niñez
    - (F94.2) Trastorno del vínculo desinhibido de la niñez
- (F95) Tics
    - (F95.0) Tic transitorio
    - (F95.1) Tic crónico motor o vocal
    - (F95.2) Tic combinado vocal y múltiple motor (Gilles de la Tourette)
- (F98) Otros trastornos emocionales y de comportamiento iniciados normalmente en la niñez y en la adolescencia
    - (F98.0) Enuresis nocturna
    - (F98.1) Encopresis
    - (F98.2) Trastorno de la alimentación de la infancia y la niñez
    - (F98.3) Pica de la infancia y la niñez
    - (F98.4) Trastornos del movimiento estereotipados
    - (F98.5) Tartamudez
    - (F98.6) Desorden lingüístico

(F99) Trastornos mentales sin especificar

- (F99) Trastorno mental no especificado en otra parte

Anexo III. Diccionario de las variables

Pregunta del cuestionario	Nombre de la variable
Are you self-employed?	self_employed
How many employees does your company or organization have?	n_employees
Is your employer primarily a tech company/organization?	tech_company
Is your primary role within your company related to tech/IT?	tech_it_role
Does your employer provide mental health benefits as part of healthcare coverage?	employer_mental_health_benefits
Do you know the options for mental health care available under your employer-provided coverage?	employer_mental_health_options
Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?	employer_discussed_mental_health
Does your employer offer resources to learn more about mental health concerns and options for seeking help?	employer_offer_resources
Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?	employer_anonymity_protected
Do you have medical coverage (private insurance or state-provided) which includes treatment of \xa0mental health issues?	medical_coverage
Do you know local or online resources to seek help for a mental health disorder?	know_local_or_online_resources
Do you have previous employers?	previous_employers

Have your previous employers provided mental health benefits?	previous_employers_mental_health_benefits
Were you aware of the options for mental health care provided by your previous employers?	previous_employers_aware_mental_health_options
Did your previous employers ever formally discuss mental health (as part of a wellness campaign or other official communication)?	previous_employer_discussed_mental_health
Did your previous employers provide resources to learn more about mental health issues and how to seek help?	previous_employers_provide_resources
Was your anonymity protected if you chose to take advantage of mental health or substance abuse treatment resources with previous employers?	previous_employer_anonymity_protected
Do you have a family history of mental illness?	family_history_mental_illness
Have you had a mental health disorder in the past?	past_mental_health_disorder
Have you been diagnosed with a mental health condition by a medical professional?	diagnosed_by_medical_professional
What is your age?	age
What is your gender?	gender
What country do you live in?	country
What US state or territory do you live in?	us_state
What country do you work in?	country_work
What US state or territory do you work in?	us_state_work
Do you work remotely?	work_remotely