Universitat Oberta de Catalunya

uoc.edu

Estudio de prevalencia y predictores de problemas de salud mental entre trabajadores tecnológicos

Byron Vinicio Lima Rojas
Máster Universitario en Ciencia de Datos
Área de Ciencia de datos aplicada a Salud

Jose Luis Iglesias Allones
Àngels Rius Gavidia

16/06/2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	Estudio de prevalencia y predictores de problemas de salud mental entre trabajadores tecnológicos.
Nombre del autor:	Byron Vinicio Lima Rojas
Nombre del consultor/a:	Jose Luis Iglesias Allones
Nombre del PRA:	Àngels Rius Gavidia
Fecha de entrega (mm/aaaa):	06/2019
Titulación::	Máster Universitario en Ciencia de Datos
Área del Trabajo Final:	Ciencia de datos aplicada a Salud
Idioma del trabajo:	Español
Palabras clave	Mental health, technology, data mining.

Resumen del Trabajo (máximo 250 palabras):

La salud mental es tan importante como el bienestar físico para las personas, las sociedades y los países. Sin embargo, solo una pequeña minoría de los 450 millones de personas sufre un trastorno mental o de comportamiento y recibe tratamiento [1].

En la actualidad, con el desarrollo e innovación constante de tecnologías obliga a los profesionales de TI a mantenerse a la vanguardia de mismas, tanto por presión social o presión laboral impuesta por las empresas para mantenerse adelante de la competencia. A nivel mundial, el costo global anual por salud mental es de $ 2.5 billones y se espera que esta cifra aumente a $ 6 mil millones en 2030. La depresión tiene efectos negativos significativos en el desempeño laboral de los empleados, contribuyendo al presentismo y el ausentismo [2].

Posterior al análisis de datos obtenidos mediante encuesta realizada a profesionales de distintas áreas, que se encuentra en Kaggle y fue obtenido por OSMI Salud Mental en el 2016, el estudio se centrará en la preparación de datos y posteriormente, mediante técnicas de aprendizaje automático, minería de datos y visualización de datos realizar el análisis y evaluar la información generada por profesionales vinculados a áreas de tecnología, con la finalidad de determinar la correlación entre el apoyo brindado por las compañías para profesionales con problemas de salud mental y las prestaciones de los seguros de vida.

Abstract (in English, 250 words or less):

Mental health is as important as physical well-being for people, societies and countries. However, only a small minority of the 450 million people have a mental or behavioral disorder and receive treatment.

At present, with the development and innovation, relationships are mandatory and the professionals to maintain the vanguard of the same, both social pressure and labor pressure. Globally, the annual global health cost is $ 2.5 billion and this figure is expected to increase to $ 6 million by 2030. Depression has negative effects on employees' job performance, contributing to presenteeism and absenteeism.

After the analysis of data obtained through a survey of professionals from different areas, which is in Kaggle and was obtained by OSMI Mental Health in 2016, the study will focus on the preparation of data and later, through techniques of machine learning, mining of data and data visualization perform the analysis and evaluate the information generated by professionals linked to technology areas, in order to determine the correlation between the support provided by the companies for professionals with mental health problems and the insurance benefits of lifetime.

Índice

Lista de figuras

Lista de tablas

Lista de códigos

1. Introducción

1.1 Contexto y justificación del Trabajo

En la actualidad, son frecuentes diagnósticos como anorexia, ansiedad, bulimia, depresión y otros tipos de trastornos en personas de cualquier edad, estos problemas de salud mental se detectan diariamente a nivel mundial y representan el 14% de las patologías conocidas [3].

Varios de estos problemas de salud mental han sido identificados, algunos ya cuentan con tratamientos exitosos, mientras que otros se encuentran en constante análisis, por lo cual es necesario realizar controles frecuentes en donde, el paciente diagnosticado cuente con el apoyo necesario para su recuperación; sin embargo, en muchas de las organizaciones no cuentan con planes de trabajo o beneficios de salud mental dentro de la cobertura provista, tampoco se ejecutan campañas de bienestar u otra comunicación oficial para brindar apoyo en este tipo de situaciones.

En los departamentos de TI de las organizaciones, este tipo de enfermedades vinculadas a la salud mental se encuentran en constante aparición, entre las múltiples causas se encuentra el estrés y presión laboral, antecedentes familiares, horarios de trabajo ajustados, problemas familiares entre otros, estos factores pueden llegar a desencadenar una serie de complicaciones de afectación personal, familiar y laboral.

El presente proyecto tiene como objetivo realizar el análisis de datos recolectados mediante encuesta a más de 1000 personas, centrándonos únicamente en profesionales de TI que permitan identificar factores, patrones y síntomas que puedan convertirse en problemas para el colaborador y la organización a futuro, estos datos serán analizados mediante técnicas de minería de datos y aprendizaje automático que permitan satisfacer los objetivos propuestos. Esta información fue obtenida por la OSMI Mental Health en el 2016 para el análisis de datos sobre prevalencia y actitudes hacia la salud mental en trabajadores de tecnología [4], con la finalidad de generar sensibilización y mejorar las condiciones de las personas con trastornos de salud mental en el lugar de trabajo de TI.

Mediante la implementación de técnicas y métodos propios de los modelos supervisados como árboles de decisión, los resultados obtenidos serán evaluados e interpretados con la finalidad de detectar patrones que permitan realizar sugerencias fundamentadas a organizaciones para controlar este tipo de enfermedades entre sus colaboradores y mejorar las prestaciones de seguro.

1.2. Motivación personal

De acuerdo a la experiencia adquirida en mis años de vida laboral, una de las principales novedades encontradas en cada una de las empresas donde he laborado es que todas cuentan con seguros públicos de acuerdo a la ley ecuatoriana, así mismo, cuentan con un seguro privado con varias prestaciones a nivel de salud general, intervenciones quirúrgicas, entre otras; sin embargo, no se cuentan con prestaciones para salud mental de sus trabajadores, ni planes de prevención en las diferentes áreas de trabajo.

La idea de analizar e interpretar los datos seleccionados es demostrar la importancia de la prestación de este tipo de seguro en las organizaciones y los factores que se deben analizar, así como, tener e prevenir inconvenientes de salud mental en colaboradores o compañeros de trabajo, sobre todo en áreas de TI donde la constante atención de

incidentes, requerimientos, cambios y trabajo bajo presión llegan a mantener al personal en altos grados de estrés que se pueden volver perjudiciales para la salud en un futuro.

1.3. Objetivos generales y específicos

1.3.1. Objetivo general

Analizar la prevalencia de problemas de salud mental entre trabajadores tecnológicos y definir predictores para detectar síntomas de forma temprana.

1.3.2. Objetivos específicos

- Identificar síntomas de prevalencia entre trabajadores tecnológicos y la afectación por género del trabajador.
- Evaluar la apertura de los trabajadores para buscar ayuda y las prestaciones de las organizaciones para tratar síntomas en sus colaboradores en base a factores de apertura/ayuda.
- Evaluar si el tipo de empresa (grande, mediana o pequeña) afecta a la proporción de los beneficios y prestaciones para la atención mental de sus colaboradores.

1.4. Enfoque y método seguido

El enfoque del presente trabajo se encuentra orientado a minería de datos, por tal motivo se realizará la ejecución de las siguientes fases:

1. **Selección del conjunto de datos:** en el conjunto de datos, se realizarán actividades de selección de atributos y datos que permitan responder a cada uno de los objetivos propuestos, así mismo la clasificación y estandarización de datos para comprensión lectora.
2. **Análisis de las propiedades de los datos:** revisión de la presencia de valores nulos, atípicos o inconsistentes para su posterior limpieza, construcción de diagramas de dispersión que permitan identificar patrones de datos anormales entre los encuestados.
3. **Transformación del conjunto de datos de entrada:** se ejecutará la revisión de los tipos de variables estadísticas y adaptaciones necesarias para cambiar entre tipos de variables, así mismo se asignarán las etiquetas adecuadas a cada variable. En caso de ser necesario, se aplicarán técnicas de estadística descriptiva o modelos de regresión lineal para comprender los datos a analizar.
4. **Selección y aplicación de técnicas de minería de datos:** se pretende revisar los tipos de variables y su adaptación a los modelos que nos permita generar conocimientos en base a cada objetivo específico.
5. **Extracción de conocimiento:** identificar patrones asociados a predictores de enfermedades mentales, validar hipótesis de empresas y prestaciones, pruebas mediante preprocesado de datos con distintas variables a través de la reducción de la funcionalidad. Así mismo, se utilizan métodos supervisados para la generación correcta de resultados.
6. **Interpretación y evaluación de datos:** Una vez obtenidos los modelos que nos permitan responder a los objetivos planteados, se realizará una comparativa de los mismos y las conclusiones de acuerdo a los resultados obtenidos.

Para el enfoque de este proyecto, se cuenta con la información completa para el análisis de datos y se han establecido las fases necesarias para cumplir con los objetivos planteados al inicio del mismo con estabilidad razonable, por tal motivo se utilizará el modelo en cascada/secuencial por cuanto al inicio de una fase se tiene que validar que

la fase anterior se encuentre concluida [5], así mismo se evaluará tras cada fase finalizada los resultados obtenidos y la claridad de la información para el cumplimiento de los objetivos.

1.5. Planificación del Trabajo

1.5.1. Calendario

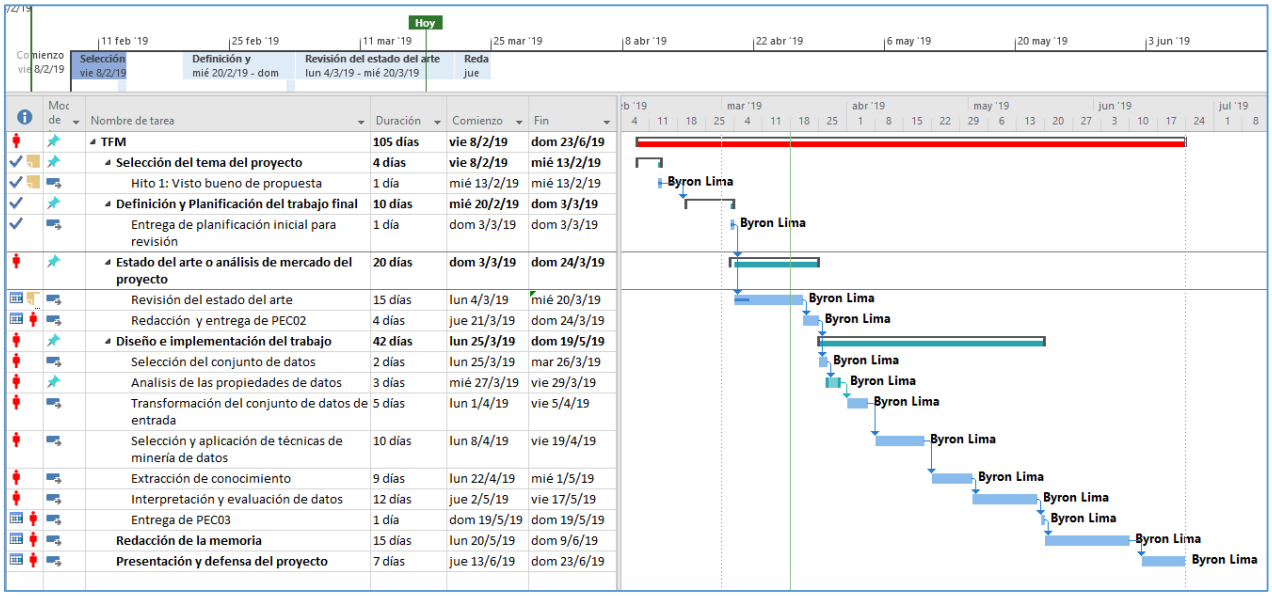

Figura 1. Calendario de elaboración del TFM

1.5.2. Tareas

Tiempo total para ejecución de trabajo: 150 días – 420 horas.

Actividades macro:

1. Selección del tema del proyecto.
2. Planificación de actividades.
3. Estado del arte.
4. Diseño e implementación del trabajo
   4.1. Selección de conjuntos.
   4.2. Análisis de las propiedades de los datos.
   4.3. Transformación del conjunto de datos de entrada.
   4.4. Selección y aplicación de técnicas de minería de datos.
   4.5. Extracción de conocimiento.
   4.6. Interpretación y evaluación de datos.
5. Elaboración de memoria.
6. Presentación y defensa del proyecto.

1.5.3. Herramientas

Para la ejecución del trabajo, es necesario la utilización de las siguientes herramientas de software:

- Jupyter notebook para la edición y ejecución de código.
- Microsoft Word para la edición de la memoria.

- Microsoft Project para la elaboración y seguimiento de la planificación.
- Microsoft Excel para la revisión y preparación de datos.
- Plataforma de Kaggle para la obtención de información.
- Python 3.7.X para el uso de librerías y comandos.
- Weka para Aprendizaje automático.

En cuanto a hardware, se utilizó un equipo Core i7 de octava generación y 16 gigas de memoria para el procesamiento de código.

1.5.4. Análisis de riesgos

1. Costos de equipamiento

En cuanto a costos de equipamiento para el desarrollo de este trabajo, la parte del hardware es propia, mientras que la parte de software se utilizará Python y sus diferentes librerías para el análisis de datos, en el caso del editor de texto se cuenta con Office 365 con licencia empresarial para uso personal.

2. Problemas técnicos

Para la parte técnica del trabajo, se han considerado el tema de respaldos sincronizando la información en OneDrive que permite manejar históricos de información, para el procesamiento de datos se utiliza un equipo repotenciado a 16 gigas de memoria RAM y un procesador Core i7 de octava generación, así mismo se mantendrá una fiel copia de los datos preparados para el análisis.

En caso de daños a nivel de hardware del equipo, únicamente es necesario la replicación de los ambientes de trabajo en un nuevo equipo, reduciendo de esta forma el impacto sobre la planificación inicial.

3. Cumplimiento del calendario

Conforme la planificación estimada, los tiempos dados se ajustaron de acuerdo a las actividades que se encuentran en cada entregable, en caso de no completar las expectativas o existir actividades no contempladas se realizaran los ajustes correspondientes para adaptarlos a los objetivos específicos sin perder de vista la fecha de los entregables.

1.6. Breve sumario de productos obtenidos

Los entregables del presente trabajo se detallan a continuación:

- Memoria de trabajo con los procedimiento, procesos y resultados obtenidos durante el desarrollo del TFM
- Mediante anexos, se incluye el código fuente utilizado en el análisis de la información, así como los datos preparados para facilitar la reproducción de los resultados.

1.7. Breve descripción de los otros capítulos de la memoria

A continuación, se procede con el detalle de los diferentes capítulos en los que se ha estructurado el trabajo:

- **Capítulo 1. Introducción:** se detalla y justifica el trabajo, indicando los objetivos y tareas asociadas a la investigación, así mismo, mediante calendario que detalla los hitos y fechas de entrega, también recursos y materiales utilizados.
- **Capítulo 2. Estado del Arte:** revisión de datos estadísticos disponibles sobre beneficios o ayuda prestados a colaborares de empresas, así mismo, revisión de contratación de seguros para salud mental. También, se realiza la revisión del análisis predictivo y sus diferentes modelos.
- **Capítulo 3. Diseño e implementación del trabajo:** se realiza la implementación de las diferentes fases de minería de datos sobre el conjunto de datos proveniente de la encuesta sobre *Mental Health in Tech* realizada en el 2016, se utilizará Jupyter Notebook y sus distintas librerías para la obtención de resultados orientados a los objetivos específicos. Así mismo, se agrega el análisis y revisión de resultados de acuerdo a la información extraída del estudio concluido.
- **Capítulo 4. Conclusiones y futuros proyectos:** valoración de los resultados del desarrollo del trabajo.
- **Capítulo 5. Glosario de términos.**
- **Capítulo 6. Bibliografía:** incluyendo artículos, libros, tesis doctorales y webs de referencia.
- **Capítulo 7. Anexos:** incluye detalle de la configuración del entorno de desarrollo, listado de librerías utilizadas y código implementado.

2. Estado del arte

Los problemas de salud mental son muy frecuentes en varios ámbitos del sector laboral, en el año 2014 se dice que aproximadamente 1 de cada 5 personas se identificaron con algún tipo de trastorno mental común durante los últimos 12 meses previos a la evaluación médica y el 29.2% fue identificado por tras experimentar un trastorno mental durante su vida [6].

El estrés laboral es cada vez más frecuente y una de las principales causas de baja laboral. Cada año aumenta el número de trabajadores que sufre de estrés y ansiedad por motivos laborales, y se estima que en 2020 sea la primera causa de muerte a nivel mundial [7].

Las nuevas tecnologías han contribuido el incremento del estrés laboral, ya que gracias a la transformación digital, separar el horario laboral del horario personal se ha vuelto más difícil debido a la fácil accesibilidad de las herramientas tecnológicas desde cualquier dispositivos portátil que tenga una conexión a internet.

La sensación de cansancio, ansiedad y malestar general influye, de forma directa a la salud mental de las personas, por tal motivo en la actualidad, existen varias organizaciones enfocadas en combatir esta enfermedad mediante la investigación y la generación de apoyo a este grupo prioritario que va en constate aumento.

En el desarrollo del presente estudio, uno de los principales insumos para el análisis de datos será los resultados de las encuestas realizadas por OSMI. OSMI es una corporación sin fines de lucro que se dedica a crear conciencia, educar y brindar recursos para apoyar el bienestar mental en las comunidades de tecnología [8]. Los datos fueron obtenidos con la finalidad de ayudar a las empresas a crear entornos de apoyo para las personas afectadas por trastornos de salud mental y, estos datos serán analizados mediante técnicas de Minería de Datos y Aprendizaje Automático para el cumplimiento de los objetivos propuestos.

2.1. Conjunto de datos

Desde la parte de investigación de OSMI bajo el título de “Datos sobre prevalencia y actitudes hacia la salud mental en trabajadores de tecnología.”, en el año 2016 se realizó una encuesta obteniendo más de 1467 respuestas [4], la misma que tuvo como objetivo medir las actitudes hacia la salud mental en el lugar de trabajo de tecnología y examinar la frecuencia de los trastornos de salud mental entre trabajadores de tecnología.

Con la aplicación de la encuesta, se pretende obtener información importante mediante el análisis de variables como posibles influyentes en la salud mental de los trabajadores, entre ellas están la forma de trabajo, el tipo de empresa, la cobertura del seguro de vida y la apertura de las organizaciones para tratar temas delicados que requieran de especialistas e impliquen gastos adicionales a las mismas.

En la revisión preliminar de la información obtenida, tenemos 63 variables para análisis en las 1467 encuestas realizadas, existe la presencia de valores nulos debido a que varios trabajadores de tecnología trabajan de forma remota o independiente. Adicional, para cada uno de las variables que hace referencia a una pregunta, por tanto, es necesario por cada objetivo especificado en el primer capítulo analizar que variables aportan resultados y permiten generar conclusiones acertadas.

2.2. Salud Mental en ambientes laborales

A nivel mundial, las organizaciones tienen la obligación de prestar apoyo a las personas con trastornos mentales para realizar su trabajo o reincorporarse al mismo. De acuerdo a estudios realizados se ha constatado que el desempleo por mucho tiempo en una persona llega a ser perjudicial para la salud mental de la misma [9]. Muchas de las iniciativas descritas anteriormente pueden ayudar a las personas que padecen trastornos mentales.

En particular, la flexibilidad horaria, la adaptación de las tareas asignadas a estas personas, la lucha contra las dinámicas negativas en el lugar de trabajo, la confidencialidad y facilitación de la comunicación con los cuadros directivos aportan a continuar realizando su trabajo o reincorporarse al mismo.

Para estimar la prevalencia, la gravedad y el tratamiento del diagnóstico y estadístico de trastornos mentales, la OMS se encuentra constantemente analizando la gravedad de los trastornos mentales no tratados a nivel mundial.

De acuerdo al análisis realizado por la Organización Mundial de la Salud (OMS) en conjunto con World Mental Health (WMH), ejecutaron encuestas de hogares cara a cara de 60.463 adultos realizados entre 2001 y 2003 en 14 países de América, Europa, Medio Oriente, África y Asia [10]. Los trastornos graves se deben a discapacidades y la gravedad del trastorno se correlacionó con la probabilidad de tratamiento en casi todos los países, 35.5% a 50.3% de los casos graves en países desarrollados y 76.3% a 85.4% en países menos desarrollados no recibieron tratamiento en los 12 meses previos a la entrevista.

En los últimos años, la constante aparición de tecnologías disruptivas y la evolución de la transformación digital están realizando un importante incremento de la calidad de vida a nivel social y profesional [11]. Para la mayor parte de trabajadores, la actividad laboral empieza a volverse desafiante y exigente, los desafíos suelen ser aristas para el crecimiento e incremento de la capacidad de superación en cada organización provocando efectos adversos, en algunos escenarios irreversibles.

En la actividad laboral, uno de los principales factores implica que, desde el trabajador exista una predisposición de buscar ayuda, y en muchos de los casos el factor discriminación o despido injustificable se antepone a esta búsqueda. Sin embargo, ADA [12] prohíbe a los empleadores el acoso indebido por una condición de salud mental y terminen o tomen otras acciones adversas en su contra, en caso de realizarse se exige los beneficios establecidos en los derechos del trabajador.

Las personas con problemas de salud mental suelen desarrollar algunas de las limitaciones de la enfermedad diagnosticada, pero rara vez las desarrollan todas, además, el grado de limitación variará entre los individuos [13]. Las organizaciones deben emplear medidas eficaces para promover la salud mental en el lugar de trabajo y aumentar con ello la productividad.

2.2.1. Impacto de las tecnologías de la información en la salud mental

La estrecha vinculación del ser humano con las TICS, debido el rechazo frontal a las mismas, da lugar a pautas de conductas disfuncionales. El imparable desarrollo de nuevas tecnologías ha dado lugar al fenómeno del e-trabajo, que implica “trasladar las mentes, no los cuerpos”, dando paso a la creación de un espacio tácito de trabajo donde pueden coexistir todos los componentes de trabajo virtualmente reunidos (miembros de

un equipo, clientes, proveedores, entre otras), y que permite la aparición de nuevas formas de trabajo (trabajo en red) [14].

Para varias personas, el adaptarse a una nueva tecnología implica desafíos de acuerdo a su grado de actualización o falta de preparación, en muchos de los datos se ven obligadas a emplearlas y terminan por minimizar su uso, limitándolo a lo imprescindible y otras aceptan las TICS adaptándolas a sus necesidades.

Así mismo, existen personas que disfrutan los beneficios y prestaciones adaptándose de forma natural y otras que, por medio de las TICS generar procesos y trabajos acordes a su ambiente laboral, llegando al punto de la adicción impidiendo desconectarse del trabajo que termina afectando en su vida personal.

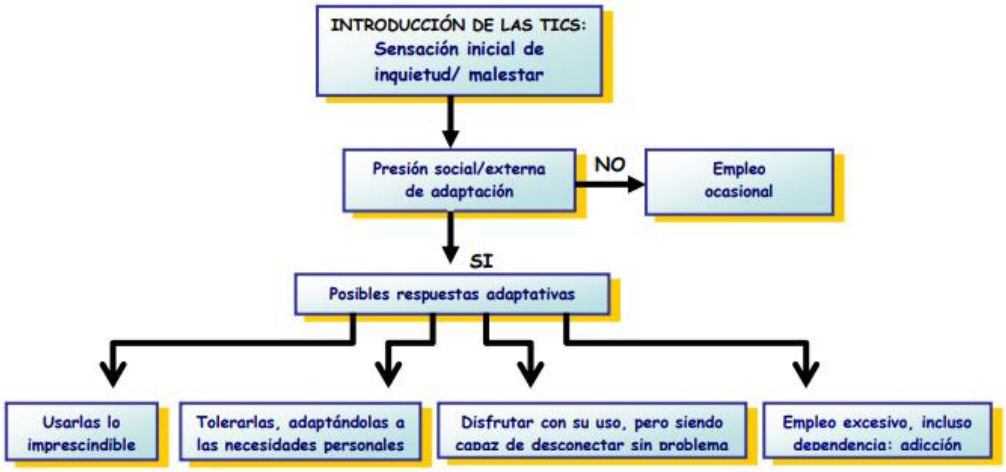

Figura 2. Progresión de la adaptación a la adicción a las TICS.
(Fuente: Porter y Kakabadse, 2006)

Según un estudio elaborado por Edenred e Ipsos [15], “el 37% de los trabajadores se ven sometidos a niveles de presión altos debido a la digitalización de sus empleos que no les permite separar la vida laboral de la personal y al estrés laboral que ello provoca, ya que la localización permanente hace que los empleados de las empresas se sientan requeridos fuera de su horario laboral”.

La facilidad que brinda las nuevas tecnologías en la conexión remota, implica que las empresas puedan requerir de sus trabajadores fuera de horarios laborales, así mismo, de acuerdo al informe de Edenred e Ipsos [15], de los 800 encuestados el 30% tienen que trabajar frecuentemente el fin de semana mientras que un 16 % tiene que dedicar tiempo de sus vacaciones al trabajo.

Desde los ámbitos de investigación de Psicología social [16], frecuentemente se está abordando la problemática de las consecuencias de la introducción de TIC en la salud de las personas en el trabajo como son los trastornos músculo esqueléticos (TME), dolores de cabeza, fatiga mental y física, ansiedad, temor, aburrimiento. Estos malestares se han identificado bajo el término de "tecnoestrés" que viene a ser el estrés derivado de la introducción de nuevas tecnologías en el trabajo y se considera como un daño psicosocial.

El término “Tecnoestrés” fue mencionado por el psiquiatra Craig Brod en 1984 [17], definiendo al mismo como “una enfermedad de adaptación causada por la falta de habilidad para tratar con las nuevas tecnologías del ordenador de manera saludable.

Hace referencia a los problemas de adaptación a las nuevas herramientas y sistemas tecnológicos."

En la actualidad, existen dos tipos de trastornos derivados por el tecnoestrés: la tecnofobia y la tecnoadicción, provocadas por el mal manejo de las tecnologías de la información y la comunicación.

La tecnofobia es considerada como una enfermedad en la que los trabajadores sienten un rechazo frontal al uso de las nuevas tecnologías en el trabajo, es decir, aquellas personas que presentan una actitud general en contra. Mientras que, con la tecnofilia, los trabajadores no han tenido que aprender la utilización de dichos medios en el trabajo, sino que forma parte de su vida cotidiana. Tal es el grado de utilización compulsiva de dichos medios que se les dificulta el desarrollo de trabajo sin la utilización de dichos medios electrónicos en el desempeño de su actividad [18].

De acuerdo a estudios de investigación realizados, se ha generado la siguiente matriz en donde se analizan la incertidumbre/abstracción contra el ritmo del uso de tecnología:

Incertidumbre / abstracción		**TRACCIÓN**	**PASIVIDAD**
	Baja	Baja tensión Pocas interrupciones Desarrollo del "flujo"	Alta Tensión Baja oportunidad de usar habilidades Efectos de subcarga / aburrimiento
		DISTRACIÓN	**RETO**
	Alta	Alta tensión Interrupciones frecuentes Desglose - evento negativo	Baja tensión Alivio del aburrimiento Descompostura - bienvenida diversión
		Alta	*Baja*

Ritmo de uso del computador/dispositivo

Figura 3. Modelo propuesto de la relación entre el uso de tecnología y las características tecnológicas. (Fuente y traducción: Mullarkey, Jackson, others, 1997 [19])

La figura 3, tomada de la investigación de "The impact of technology characteristics and job control on worker mental health" [19], procura resumir las relaciones entre las características de la tecnología, uso del computador y la tensión provocada. Esto propone que las demandas ambientales objetivas o subjetivas, solo se experimentarán como estresantes si son incongruentes con las preferencias o necesidades de un trabajador.

2.2.2. Prevalencia, gravedad por trastornos mentales en sector laboral

De acuerdo a datos oficiales publicados por el Workplace Mental Health [20], el estrés excesivo en el lugar de trabajo causa 120,000 muertes asombrosas y resulta en casi $190 mil millones en costos de atención médica anual. De tal forma, que representa del 5% al 8% del gasto nacional en salud, derivado principalmente de las altas demandas en el trabajo ($ 48 mil millones), la falta de seguro ($ 40 mil millones) y el conflicto entre el trabajo y la familia ($ 24 mil millones).

Estos son algunos de los efectos dañinos para la salud del estrés excesivo [20]:

- Daño a estructuras cerebrales y circuitos clave, capacidad reducida para enfrentar el estrés futuro y aumento de la ansiedad y depresión crónica.
- El inicio del trastorno por estrés postraumático (TEPT).

- Funcionamiento del sistema inmune reducido.
- Aumento de la inflamación y depresión.

De acuerdo al estudio publicado por Gray, P [21], “la mayor parte del costo de los problemas de salud mental se relaciona con los altos niveles de ausencia de los empleados. Se ha estimado que, en el Reino Unido, se pierden 91 millones de días hábiles cada año debido a dificultades de salud mental). En el mismo informe, se afirmó que la ausencia por enfermedad relacionada con el estrés tiene un costo estimado de £ 4 mil millones anuales.”

En la actualidad, el aumento de la ausencia no es el único impacto costoso de los problemas de salud mental en el trabajo, también afecta las relaciones interrumpidas, el trabajo ineficaz, el aumento de la rotación y, por último, la constante aparición de nuevas tecnologías disruptivas que, si bien son para mejorar la vida cotidiana y laboral, las empresas no se enfocan en preparar a sus empleados para nuevos retos que requerirán paciencia y entendimiento.

De acuerdo al artículo de “Mental Health in High-Tech System” 2005, se realizó análisis sobre el estrés laboral en países post-industrializados, en donde se concluyó que el 29% de los trabajadores experimentan diversos niveles de estrés en lugares de trabajo, y son los empleados y empleadores los que tienen que compartir la carga de los efectos perjudiciales del estrés laboral en los problemas de salud mental.

Los resultados obtenidos mediante la prueba exacta de Fisher se detallan a continuación:

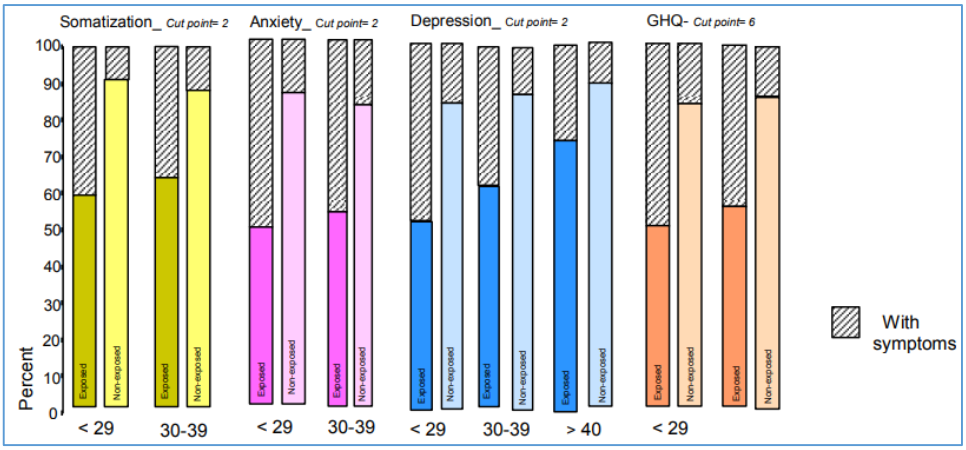

Figura 4. Comparación de salud mental en grupos expuestos y no expuestos.
(Fuente: Mental Health in High-Tech System, 2005 [22])

Los resultados de este análisis mostraron que para dos grupos de la primera edad (< 29 años y 30-39 años), los síntomas de dificultad mental fueron significativamente mayores en el grupo expuesto que del grupo no expuesto. Hubo efectos laborales significativos en la somatización, la ansiedad y la depresión. No se encontraron efectos significativos del trabajo en los síntomas de disfunción social en ningún grupo de edad.

En el tercer grupo de edad (más de 40 años) se observó una depresión significativamente mayor en el grupo expuesto que en el grupo no expuesto. Las otras variables dependientes no mostraron una diferencia significativa entre el grupo expuesto y el grupo no expuesto en este grupo de edad [22].

2.3. Minería de datos y Aprendizaje Automático

La minería de datos o data mining existe desde hace más de un siglo, partiendo desde la máquina universal de Alan Turing para realizar cálculos similares a los computadores actuales, dando paso a la “Prueba de Turing” [23].

Data Mining se aplica a una variedad de propósitos, incluida la investigación financiera [23]. Los inversores pueden utilizar la extracción de datos y el rastreo web para ver las finanzas de una nueva empresa y ayudar a determinar si quieren ofrecer financiación

El aprendizaje automático incorpora bases de la minería de datos [23], pero también puede crear correlaciones automáticas y aprender de ellas para aplicarlas a nuevos algoritmos. Es la tecnología que se encuentra detrás de los autos autónomos que pueden adaptarse rápidamente a las nuevas condiciones mientras se conduce.

Las técnicas de aprendizaje automático permiten mejorar el rendimiento comercial a través de la segmentación de clientes, la generación de material publicitario de marca, la extracción y clasificación de contenido relevante, la comunicación con el cliente y la productividad y el rendimiento general [24]. El éxito para la obtención de resultados implica conocimiento de datos, técnicas y modelos de minería de datos, clasificación de datos, análisis de resultados e interpretación.

Así mismo, para la interpretación de datos se deben combinar conceptos y métodos procedentes de diferentes disciplinas como bases de datos, estadística, aprendizaje automático, la integración bajo una perspectiva puntual beneficia al cumplimiento de objetivos prácticos finales [25].

En la actualidad, grandes empresas de redes sociales como Google, Facebook, NBA y servicios como Uber Eats utilizan Data Mining y Machine Learning para la atraer nuevos clientes y retenerlos con constante innovación, así mismo, les permite mejorar ingresos y generar productos de calidad que diferencien de la competencia. Adicional, la minería de datos, realiza análisis de datos y permite detectar cuándo es necesario realizar un cambio o emprender en un nuevo proyecto.

Los algoritmos de aprendizaje automático se clasifican en dos grandes categorías: supervisados (predictivos) y no supervisados (descubrimiento del conocimiento). Los algoritmos supervisados predicen el valor de un atributo (etiqueta) de un conjunto de datos que se conocen como otros atributos (atributos descriptivos). Las relaciones de estos atributos ayudan a la predicción en datos conociéndose como aprendizaje supervisado, desarrollándose en dos fases: entrenamiento y pruebas [26]

Los algoritmos no supervisados, a diferencia de los algoritmos de clasificación supervisada no disponen de un reconocimiento de patrones o conjunto de entrenamiento, y valiéndose de algoritmos de agrupamiento intentan construirlo [27], estos algoritmos parten de un conjunto de datos del que no se tiene un conocimiento a priori, siendo el objetivo en este tipo de análisis la comprensión de los datos o la transformación automática de los datos [28].

2.3.1. Aprendizaje supervisado

El aprendizaje supervisado intenta extraer aquellas propiedades que permiten discriminar mejor la clase de cada ejemplo, y como consecuencia requieren de una clasificación previa (supervisión) del conjunto de entrenamiento.

Las principales caracteristicas del aprendizaje supervisado se detallan a continuación [29]:

- El aprendizaje se realiza por contraste entre conceptos (¿Qué características distinguen a los ejemplos de un concepto de otros?).
- Un conjunto de heurísticas (Función de preferencia que guía en el proceso) permitirán generar diferentes hipótesis.
- Existirá un criterio de preferencia (sesgo) que permitirá escoger la hipótesis más adecuada a los ejemplos.
- Como resultado da el concepto o conceptos que mejor describen a los ejemplos.

Este tipo de aprendizaje se aplica en problemas de clasificación, como identificación de dígitos, diagnósticos, o detección de fraude de identidad, así mismo, se aplica en problemas de regresión como predicciones meteorológicas, de expectativa de vida, de crecimiento, etc.

El aprendizaje supervisado se divide en dos tipos como clasificación y regresión, las mismas que se distingen por el tipo de variable objetivo. En los casos de clasificación, es de tipo categórico, mientras que, en los casos de regresión, la variable objetivo es de tipo numérico [31].

Los diferentes algoritmos más usados en el aprendizaje supervisado se detallan a continuación [32]:

- Árboles de decisión.
- Clasificación de Naïve Bayes.
- Regresión por mínimos cuadrados.
- Regresión Logística.
- Support Vector Machines (SVM).
- Métodos “Ensemble” (Conjuntos de clasificadores).

Para la aplicación del algoritmo supervisado y con la finalidad de que el modelo implementado aprenda, se dividen los datos en dos conjuntos: conjunto de entrenamiento y conjunto de validación. Con esta división, evitamos el sobreajuste (overfitting) y el infrajuste (underfitting), ambos ocurren cuando el modelo no aprende de forma óptima [33].

- El infrajuste se presenta cuando el modelo que obtenemos es demasiado simplista a la hora de hacer predicciones sobre datos que no ha visto aún.
- El sobreajuste se presenta cuando el modelo se aprende tan bien las características de los datos de entrenamiento, que no es capaz de generalizar bien cuando le presentamos nuevos datos.

Otro término para el aprendizaje supervisado es la clasificación, una amplia gama de clasificadores están disponibles, cada uno con sus fortalezas y debilidades. El clasificador de rendimiento depende en gran medida de las características de los datos que deben clasificarse [34].

2.3.2. Aprendizaje no supervisado

Los algoritmos no supervisados consisten en encontrar la partición más adecuada del conjunto de entrada a partir de similitudes entre sus ejemplos. Las principales caracteristicas del aprendizaje no supervisado se detallan a continuación [29]:

- No existe una clasificación de los ejemplos.
- Se busca descubrir la manera más adecuada de particionar los ejemplos (buscamos su estructura).
- El aprendizaje se guía por la similaridad/disimilaridad entre ejemplos.
- Existirán criterios heurísticos de preferencia que guíen la búsqueda.
- Como resultado se tiene una partición de los ejemplos y una descripción de la partición

Con la aplicación de este aprendizaje, un problema surge cuando el sistema toma la decisión de elegir un determinado patrón de entre todos los proporcionados del conjunto de datos, siendo el sistema el que debe determinar la clase de la cual obtendrá ese patrón. Otro inconveniente es la toma de decisiones, correctas o no, aquí entra en juego la lógica difusa, en la utilización de técnicas de clustering difuso [30].

Este aprendizaje se utiliza comúnmente en problemas de clustering, agrupamientos de co-ocurrencia y profiling; sin embargo, los problemas que implican tareas de encontrar similitud, predicción de enlaces o reducción de datos, pueden ser supervisados o no.

Los algoritmos más habituales son [32]:

- Algoritmos de clustering.
- Análisis de componentes principales.
- Descomposición en valores singulares (singular value decomposition).
- Análisis de componentes independientes (Independent Component Analysis).

Otra forma de aprendizaje no supervisado es la agrupación (en inglés, clustering), el cual a veces no es probabilístico [36]. Las tareas de clustering buscan agrupamientos basados en similitudes, pero esto no garantiza que tengan algún significado o utilidad. En ocasiones, explorar los datos sin un objetivo definido se pueden encontrar correlaciones espurias curiosas, pero muy pocas prácticas a la hora de realizar un estudio más detallado [32].

2.3.3. Selección de algoritmos para resolución de problemas

La parte más difícil de resolver un problema de aprendizaje automático puede ser encontrar el estimador adecuado para el trabajo, existen diferentes estimadores que son adecuados para diferentes tipos de datos y diferentes problemas.

Cuando la información que deseamos analizar, ha pasado por las etapas de limpieza de datos y estandarización de información, en conjunto con los objetivos específicos a cumplir, es necesario seleccionar uno o varios algoritmos que nos permitan crear modelos de conocimiento o descubrir patrones interesantes que cumplan con las expectativas requeridas.

El siguiente diagrama de flujo está diseñado para ofrecer a los usuarios una guía aproximada sobre cómo abordar los problemas con respecto a qué estimadores deben usar con su información.

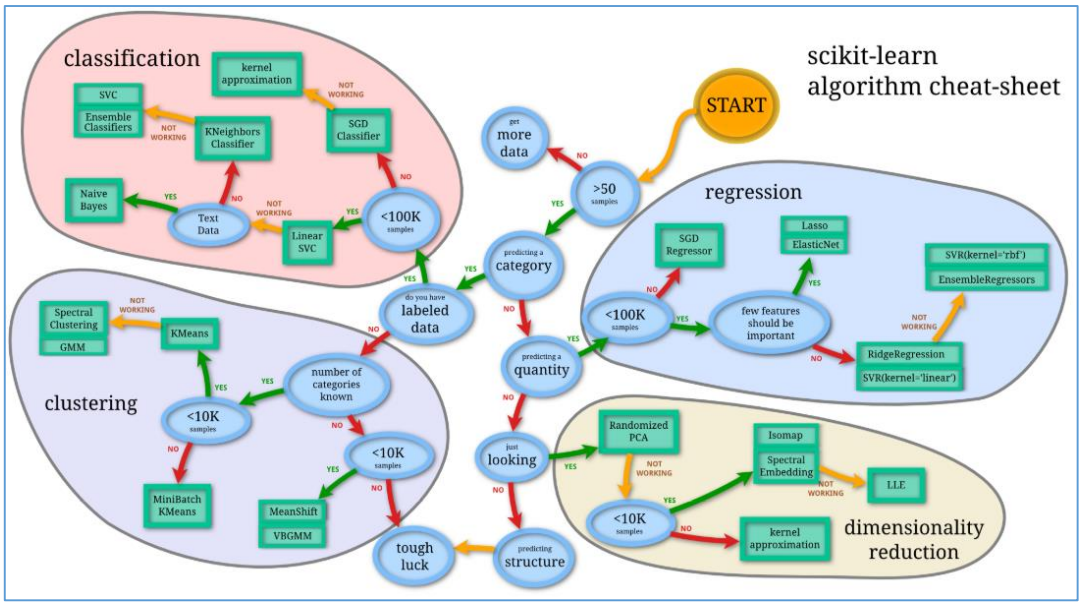

Figura 5. Cheat-sheet de selección de algoritmo de Scikit-learn.
Fuente: Scikit Learn [35].

Mediante el uso de hojas de ruta, se puede seleccionar de una forma más fácil el algoritmo que mejor se adapte al estudio a realizarse. La hoja de ruta de scikit-learn es la más conocida e interactiva para la consulta de información y búsqueda de ejemplos, existe otra correspondiente a Microsoft Azure Machine Learning Algorithm cheat sheet.

2.4. Estudios de Salud Mental con técnicas de Minería de datos y Aprendizaje Automático.

En internet, se encuentran varios estudios y artículos que nos hablan de la salud mental y los síntomas o patrones que presenta una persona o un grupo de individuos de acuerdo a estudios realizados, esta información en la actualidad mediante técnicas de minería de datos o aprendizaje automático puede ayudar a predecir o identificar patrones comunes en personas que sufren problemas de salud mental o personas que presentan características comunes de un potencial desorden en el futuro.

El aprendizaje automático mejora su capacidad predictiva conforme se analizan las diferentes variables disponibles de personas con antecedentes de algún tipo de desorden mental, predecir o identificar estos factores potenciales permitirá tomar acciones a futuro en los distintos factores a los que se encuentran expuestos los seres humanos.

Las consecuencias de no tratar a tiempo desórdenes mentales genera varias consecuencias negativas, tanto para la sociedad que rodea a un individuo como al mismo, uno de los resultados negativos puede llegar el suicidio que causa entre 800.000 y un millón de muertes al año según la OMS. En los últimos años, “el aprendizaje automático mejoró su capacidad predictiva a medida que el riesgo de suicidio se acercaba en el tiempo, dando más valor a aquellas variables que mostraban mayor relevancia en los días previos a un suicidio, aunque parecieran menos importantes semanas atrás, como por ejemplo el índice de masa corporal en adolescentes con depresión: las máquinas dieron más significado al sobrepeso en este grupo a corto plazo y le quitaron valor a largo plazo” [37].

En este último estudio con respecto a los suicidios, se utilizó una variante del aprendizaje automático basada en los llamados “bosques aleatorios”, una combinación de árboles de decisión y como fuente de datos se utilizó registros anonimizados de ingresos hospitalarios de adolescentes con autolesiones entre 1998 y 2015. Así mismo, mediante técnicas de aprendizaje automático se puede ayudar en la detección de la esquizofrenia, el algoritmo identificó con éxito a pacientes con esquizofrenia con un 78% de precisión. También predijo con un 82% de precisión si un paciente respondería positivamente a un tratamiento antipsicótico específico llamado risperidona [38].

En IBM, sus científicos están utilizando transcripciones y audios de entrevistas psiquiátricas, junto a técnicas de aprendizaje automático para identificar patrones en el habla, para ayudar a los médicos clínicos a predecir y monitorear con precisión enfermedades como psicosis, esquizofrenia, manía y depresión. Hoy, toma solamente son necesarias 300 palabras para ayudar a los médicos clínicos a predecir la probabilidad de que una persona padezca psicosis [39].

A nivel de estudios e investigación, existe una revisión sistemática mediante algoritmos y técnicas de minería de datos en salud mental sobre todo a enfermedades más prevalentes como: demencia, Alzheimer, esquizofrenia y depresión. Se encontraron un total de 211 artículos relacionados con técnicas y algoritmos de Data Mining aplicados a los principales problemas de salud mental, se han identificado 72 artículos como trabajos relevantes, de los cuales el 32% son Alzheimer, 22% demencia, 24% depresión, 14% esquizofrenia y 8% trastornos bipolares. Muchos de los artículos muestran la predicción de los factores de riesgo en estas enfermedades [40].

Así mismo, existe un estudio para la generación de reglas y predicción del trastorno de ansiedad utilizando árboles de modelo logístico, cuya finalidad es estudiar el estilo de vida del paciente, la extracción de datos es una solución a través de la cual los datos contextuales se puede obtener del paciente y se puede generar estrés con respecto a las reglas. El objetivo principal de esta investigación es proporcionar un nuevo enfoque para generar el estrés relevante en relación con las reglas y predecir el nivel de estrés de un paciente que utiliza árboles de modelos logísticos para ayudar a la sociedad médica [41].

Para la clasificación de los diferentes tipos de estrés, se realizó la gestión del tecnostress a través de la minería de datos, enfocado a empleados y usuarios que se enfrentan a un uso excesivo de las Tecnologías de la Información y la Comunicación (TIC) de los Sistemas de Información (SI) en muchos aspectos de su vida, se detecta y evalúa mediante el uso de Data Mining y se clasifica a los pacientes utilizando algunos algoritmos de aprendizaje según su tipo de estrés [42].

Desde luego, existen algunos estudios e investigaciones que analizan la salud mental través de comentarios publicados en redes sociales a fin de preparar una cantidad suficiente de datos supervisados para el aprendizaje automático, lo que se espera es que la capacidad de clasificar las afirmaciones diarias conduzca a la detección temprana de trastornos mentales [43], así mismo, la detección temprana de riesgo de depresión a partir del contenido generado por el usuario en las redes sociales [44].

Por último, con la finalidad de mejorar la salud mental en estudiantes universitarios e identificar patrones negativos que permitan tomar acciones para una ayuda personalizada, se ha realizado la aplicación de reglas de asociación negativas y positivas en el análisis de salud mental [45].

3. Diseño e implementación

Para el estudio sobre la prevalencia y predictores de problemas de salud mental entre trabajadores tecnológicos, se ejecutaron las siguientes fases para la preparación y análisis de datos:

1. Carga del conjunto de datos.
2. Análisis de las propiedades de los datos.
3. Transformación del conjunto de datos de entrada.
4. Selección y aplicación de técnicas de minería de datos.
5. Extracción de conocimiento.
6. Interpretación y evaluación de datos.

El enfoque para la revisión, análisis y aplicación de técnicas de aprendizaje automático se realizará de acuerdo a los objetivos específicos del estudio en curso, los mismos que se detallan a continuación:

1. Identificar síntomas de prevalencia entre trabajadores tecnológicos y la afectación por género del trabajador.
2. Evaluar la apertura de los trabajadores para buscar ayuda y las prestaciones de las organizaciones para tratar síntomas en sus colaboradores en base a factores de apertura/ayuda.
3. Evaluar si el tipo de empresa (grande, mediana o pequeña) afecta a la proporción de los beneficios y prestaciones para la atención mental de sus colaboradores.

En cada una de las fases propuestas se detalla las actividades ejecutadas y los lineamientos que se tomaron en cuenta para el cumplimiento de las mismas, así mismo, únicamente se mencionan los resultados relevantes de la investigación, el código fuente relevante se encuentra en el anexo 3.

Para este estudio se utilizó las siguientes herramientas de software:

- Jupyter Notebook 5.7.8: se utiliza para carga de datos, transformación, análisis y aplicación de métodos supervisados.
- Microsoft Excel 2016: revisión de la información disponible de la encuesta.
- Weka 3.6: para la realización de validaciones adicionales mediante la aplicación de estrategias/algoritmos luego de la fase de extracción del conocimiento.

3.1. Carga del conjunto de datos

En primer lugar, se realizó la descarga del conjunto de datos OSMI Mental Health in Tech Survey 2016 (disponible en https://www.kaggle.com/osmi/mental-health-in-tech-2016/home).

El conjunto de datos se encuentra disponible en Kaggle y existen versiones desde el 2015 al 2018, con la finalidad de cumplir con los objetivos planteados en el TFM y por la cantidad de datos disponibles, se utilizó la versión del año 2016 que cuenta con 1433 registros finales para análisis.

Previo a la carga del archivo .csv, se realizó una exploración de datos en Excel con la finalidad de conocer la estructura de los datos, la tabulación, los tipos de datos y los nombres de las columnas, así mismo, se exploró los tipos de datos que se manejan para tener una idea clara de los tipos de variables que dispondremos para el análisis de datos.

En este análisis se encontraron registros que no cumplían con el objetivo de las encuestas, por tal motivo se procedió con la eliminación de las mismas de forma manual, así mismo, existían comas entre las separaciones de las respuestas y se creaba un conflicto en el análisis de datos, por lo cual se procedió con la estandarización de las mismas para la carga en Jupyter Notebook.

```
mentalhealthbd=pd.read_csv(r'mh2016.csv', delimiter=",")
```

Código 1. Carga del conjunto de datos

Una vez que tenemos el conjunto de datos, tenemos todos los registros de la encuesta disponibles para trabajar. Cada fila corresponde a la respuesta de una persona en particular, cada columna a una pregunta de la encuesta, y cada celda puede tener un valor nulo o no, esto debido a la relación de las preguntas entre sí.

3.2. Análisis de las propiedades de los datos

Con la finalidad de conocer un poco más a detalle la información recolectada mediante la encuesta realizada por OSMI, es necesario la revisión de la presencia de valores nulos, valores atípicos o inconsistentes para la posterior limpieza y estandarización de datos, se dispone en el conjunto de datos de 1433 encuestas válidas y 63 preguntas o atributos.

En esta fase, se analizó los primeros 5 registros del conjunto de datos con la finalidad de tener una idea de las posibles respuestas brindadas a cada pregunta por parte de los encuestados y, se generan estadísticas descriptivas y la revisión de los tipos de variables asignados a cada pregunta (columna).

	Are you self-employed?	How many employees does your company or organization have?	Is your employer primarily a tech company/organization?	Is your primary role within your company related to tech/IT?	Does your employer provide mental health benefits as part of healthcare coverage?	Do you know the options for mental health care available under your employer-provided coverage?	Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?	Does your employer offer resources to learn more about mental health concerns and options for seeking help?	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?	If a mental health issue prompted you to request a medical leave from work, asking for that leave would be:	Do you think that discussing a mental health disorder with your employer would have negative consequences?
0	0	26-100	1.0	NaN	Not eligible for coverage / N/A	NaN	No	No	I don't know	Very easy	No
1	0	6-25	1.0	NaN	No	Yes	Yes	Yes	Yes	Somewhat easy	No
2	0	6-25	1.0	NaN	No	NaN	No	No	I don't know	Neither easy nor difficult	Maybe
3	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	0	6-25	0.0	1.0	Yes	Yes	No	No	No	Neither easy nor difficult	Yes

Figura 6. Análisis de tipos de datos en los primeros 5 registros

De acuerdo a una breve inspección de los datos, se puede visualizar varios campos con valores nulos, algunos campos tienen información que se puede estandarizar con la asignación de un valor numérico de acuerdo a cada escenario. A nivel de enfermedades mentales diagnosticadas a cada uno de los entrevistados, se puede visualizar nombres extensos y en muchos casos, pacientes identificados con más de un problema mental por lo cual es necesario realizar una revisión por separado de cada enfermedad y realizar una estandarización con las siglas médicas que se dan a cada desorden.

Del conjunto de datos, se generaron estadísticas descriptivas que resumen la tendencia central, la dispersión y la forma de la distribución de los datos excluyendo los valores nulos. También, se revisó las series de objetos y números, la salida de información varía dependiendo de la información proporcionada.

	Are you self-employed?	Is your employer primarily a tech company/organization?	Is your primary role within your company related to tech/IT?	Do you have medical coverage (private insurance or state-provided) which includes treatment of mental health issues?	Do you have previous employers?	Have you ever sought treatment for a mental health issue from a mental health professional?	What is your age?
count	1433.000000	1146.000000	263.000000	287.000000	1433.000000	1433.000000	1433.000000
mean	0.200279	0.770506	0.942966	0.644599	0.882066	0.585485	34.286113
std	0.400349	0.420691	0.232350	0.479471	0.322643	0.492810	11.290931
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	3.000000
25%	0.000000	1.000000	1.000000	0.000000	1.000000	0.000000	28.000000
50%	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	33.000000
75%	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	39.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	323.000000

Figura 7. Estadísticas descriptivas del conjunto de datos.

En esta revisión se visualiza únicamente variables categóricas y cuantitativas y se puede determinar que existen variables que no se están correctamente clasificadas, por lo cual será necesario la actualización del tipo de variable o corrección de valores atípicos. Se aprecia 6 preguntas que tienen contestación de 'Si' o 'No', así mismo, en la edad de los encuestados se identificó un valor máximo de 323 considerado como un valor atípico.

A continuación, se revisó los tipos de variables que asigno Python a cada pregunta del conjunto de datos, esto permite tener una idea de los datos que se clasificación correctamente y aquellas variables que requieren una estandarización en la siguiente fase.

```
1433 non-null object
What is your age?
1433 non-null int64
What is your gender?
1430 non-null object
What country do you live in?
1433 non-null object
What US state or territory do you live in?
840 non-null object
What country do you work in?
1433 non-null object
What US state or territory do you work in?
851 non-null object
Which of the following best describes your work position?
1433 non-null object
Do you work remotely?
1433 non-null object
dtypes: float64(3), int64(4), object(56)
memory usage: 705.4+ KB
```

Figura 8. Revisión de tipos de variables.

De acuerdo a los tipos de datos, se registran 3 variables como dato flotante, 4 variables como tipo de dato entero y 56 variables con tipo de dato objeto. En la siguiente fase, se realizará una normalización de las variables con la finalidad que los datos de lectura se lean correctamente para el posterior análisis.

3.3. Transformación del conjunto de datos de entrada

Esta fase también conocida como preprocesado de los datos, se realiza el proceso de análisis de datos que sigue a la recolección de datos y es previa al modelado posterior, que constituirá el núcleo de la extracción de información de los datos.

En este parte del proyecto, se procede a normalizar/estandarizar variables cualitativas, revisar posibles inconsistencias entre variables, buscar valores atípicos en las variables cuantitativas e imputar valores perdidos.

Es necesario normalizar valores de las variables que se consideren como datos no correctos. En esta fase, se procede a imputar los valores de edades atípicas identificadas en el ítem anterior a partir de los k-vecinos más próximos.

Primeramente, para este estudio no se toma en cuenta a empleados que trabajen de forma independiente, ya que tienen la posibilidad de adaptar su tiempo y sus actividades personales a su forma de trabajo, reduciendo de esta manera las posibilidades de desarrollar algún tipo de enfermedad mental o sufrir algún tipo de accidente laboral.

```
features=(list(mentalhealthbd))
count=0
for index,col in enumerate(features):
    idx=mentalhealthbd.index[mentalhealthbd[col].isnull()]
    if(len(idx)==287):
        k=idx
        count+=1
mentalhealthbd.drop(k,inplace=True)
mentalhealthbd.shape
```

Código 2. Eliminación de encuestas de trabajadores independientes.

Se descartan de 287 registros de encuentras de trabajadores independientes, del conjunto original quedan 1146 registros. Posteriormente, se descartan aquellas preguntas (columnas) que no superan el 50% de respuestas del total de encuestados, debido que no aportarán mayor información o detalle para el cumplimiento de los objetivos establecidos.

```
real_features=(list(mentalhealthbd))
count=0
less_answers=[]
for col in real_features:
    if(sum(pd.isnull(mentalhealthbd[col]))>721):
        count=count+1
        less answers.append(col)
mentalhealthbd.drop([i for i in less_answers],axis=1,inplace=True)
mentalhealthbd.shape
```

Código 3. Eliminación de columnas con más del 50% de preguntas no respondidas.

En este punto, ya se han eliminado registros que no aportan significado o un valor agregado al conjunto de datos ideal; sin embargo, también es importante descartar datos de columnas que no tienen respuestas iguales y pueden afectar al modelo final que estamos trabajando, quedando únicamente con 1015 registros y 53 columnas.

```
real_features=(list(mentalhealthbd))
count=0
for index,col in enumerate(real_features):
    idx=mentalhealthbd.index[mentalhealthbd[col].isnull()]
    if(len(idx)==131):
        k=idx
        count+=1
mentalhealthbd.drop(k,inplace=True)
mentalhealthbd.shape
```

Código 4. Eliminación de respuestas ambiguas.

Cada una de las columnas, representa una pregunta de la encuesta, en la revisión de las mismas se determina que existen columnas que no aportan valor o dirección al análisis de datos final.

Aquellas preguntas se detallan a continuación:

- Would you be willing to bring up a physical health issue with a potential employer in an interview?: Why Why or why not?'
- Would you bring up a mental health issue with a potential employer in an interview?: Why or why not?
- What US state or territory do you live in?
- What US state or territory do you work in?
- What country do you live in?
- Which of the following best describes your work position?
- Do you have previous employers?
- If yes, what condition(s) have you been diagnosed with?

Esta filtración de columnas se debe a que no es necesario conocer de cada encuestado donde vive o trabaja, así como el rol que desempeña en su empresa. Así mismo, se descarta el diagnostico auto-detectado por el encuestado, ya que no fue identificado por un profesional de la salud, el resto de variables no se encuentran vinculadas entre sí.

Para la estandarización de la información, primeramente, se identificaron valores atípicos en la edad de los encuestados, por cuanto se reemplazará esta información con la estadística media de las edades.

```
mentalhealthbd.loc[(mentalhealthbd['What is your age?'] > 90), 'What
is your age?'] = 34
mentalhealthbd.loc[(mentalhealthbd['What is your age?'] < 10), 'What
is your age?'] = 34
```

Código 5. Estandarización de edades atípicas.

Con respecto al género de cada encuestado, se estandarizó esta información en 3 categorías como Male, Female y Genderqueer/Other. En el caso de valores nulos se incluyeron dentro de la categoría Genderqueer/Other debido a que no tenemos un patrón específico para asignar una categoría como Male o Female, bajo esta definición se estandarizó 3 registros.

```
mentalhealthbd['What is your gender?'] = mentalhealthbd['What is your
gender?'].replace([
    'male', 'Male ', 'M', 'm', 'man', 'Cis male', 'Male.',
    'Male (cis)', 'Man', 'Sex is male','cis male', 'Malr', 'Dude',
    "I'm a man why didn't you make this a drop down question. You shou
ld of asked sex? And I would of answered yes please. Seriously how muc
h text can this take? ",'mail', 'M|', 'male ', 'Cis Male',
    'Male (trans, FtM)', 'cisdude', 'cis man', 'MALE'], 'Male')
mentalhealthbd['What is your gender?'] = mentalhealthbd['What is your
gender?'].replace([
    'female', 'I identify as female.', 'female ',
    'Female assigned at birth ', 'F', 'Woman', 'fm', 'f',
    'Cis female', 'Transitioned, M2F', 'Female or Multi-Gender Femme',
    'Female ', 'woman', 'female/woman', 'Cisgender Female', 'mtf',
    'fem', 'Female (props for making this a freeform field, though)',
    ' Female', 'Cis-woman', 'AFAB', 'Transgender woman',
    'Cis female '], 'Female')
mentalhealthbd['What is your gender?'] = mentalhealthbd['What is your
gender?'].replace([
    'Bigender', 'non-binary,', 'Genderfluid (born female)', 'Other',
    'Other/Transfeminine', 'Androgynous', 'male 9:1 female, roughly',
    'nb masculine', 'genderqueer', 'Human', 'Genderfluid', 'Enby',
    'genderqueer woman', 'Queer', 'Agender', 'Fluid', 'Nonbinary',
    'Genderflux demi-girl', 'female-bodied; no feelings about gender',
    'non-binary', 'Male/genderqueer', 'none of your business',
    'Unicorn', 'human', 'Genderqueer'], 'Genderqueer/Other')
mentalhealthbd['What is your gender?']  = mentalhealthbd['What is your
gender?'].replace(np.NaN, 'Genderqueer/Other')
mentalhealthbd['What is your gender?'].value_counts()
```

Código 6. Estandarización de género de las encuestas.

De esta forma, tenemos 746 personas de sexo masculino, 245 personas de sexo femenino y 24 personas identificadas con otro género.

En la revisión de los primeros 5 valores del conjunto de datos, en la pregunta que se realizó a los encuestados con respecto a **"Las condiciones o trastornos que fueron identificados por profesionales de la salud"** existen uno o varios trastornos detallados en las respuestas, por tal motivo se realizó la estandarización de la información con los acrónimos médicos respectivos.

Los desórdenes de salud mental diagnosticados y los acrónimos médicos que se utilizaron para la estandarización se detallan a continuación:

Item	Enfermedad diagnosticada	Acrónimo Médico
1	Addictive Disorder	AD
2	Anxiety Disorder (Generalized, Social, Phobia, etc)	GAD
3	Attention Deficit Hyperactivity Disorder	ADHD
4	Autism	ASD
5	Eating Disorder (Anorexia, Bulimia, etc)	ED
6	Mood Disorder (Depression, Bipolar Disorder, etc)	MD
7	Obsessive-Compulsive Disorder ADD (w/o Hyperactivity)	OCD
8	Personality Disorder (Borderline, Antisocial, Paranoid, etc)	PD
9	Post-traumatic Stress Disorder	PTSD

10	Psychotic Disorder (Schizophrenia, Schizoaffective, etc)	PSYCHOSIS
11	Seasonal Affective Disorder	SAD
12	Sexual addiction	SA
13	Stress Response Syndromes	SRS
14	Substance Use Disorder	SUD
15	Traumatic Brain Injury	TBI

Tabla 1. Enfermedades diagnosticadas por profesionales de la salud.

Con respecto al número de empleados que tienen las empresas donde laboran los encuestados, se estandarizó la información convirtiendo esta variable en categórica, la información de la misma es clave para el cumplimiento de los objetivos específicos. La estandarización se la realizó de la siguiente forma:

Item	Empleados por empresa	Estandarización
1	1 - 5	5
2	6 - 25	25
3	26 - 100	100
4	100 - 500	500
5	500 - 1000	1000
6	More than 1000	5000

Tabla 2. Estandarización de rangos de empleados por empresa.

A continuación, se procede a estandarizar todas las respuestas que se encuentran en cada columna bajo las siguientes definiciones:

- Se asigna un valor numérico a cada tipo de respuesta.
- Para respuestas binarias como afirmativo y negativo, se ha considerado el uso de 0 para "No" y 1 para "Yes".
- Otras preguntas que tienen respuestas escalables de tipo nominal, las mismas se transforman a una escala numérica de acuerdo al valor que aporte cada una de las respuestas al estudio que estamos ejecutando.
- En este caso se mantienen valores nulos en algunas columnas, para lo cual las respuestas nulas se regulan como una respuesta negativa.
- En las respuestas de desconocimiento o que no se encuentran seguros de la respuesta se consideraron como respuestas negativas. Esta última definición se da debido a que, varias de las preguntas se enfocan a la comunicación que tiene el empleador con sus empleados y si existe desconocimiento de parte de los mismos se pueden deducir dos razones como que, la empresa no presta dicho beneficio/apertura o la empresa brinda dicho servicio, pero su comunicación no es efectiva a sus empleados.

La decisión de estandarizar todas las variables a numéricas se debe a que, en varios de los algoritmos o bibliotecas de minería de datos y Aprendizaje Automático este tipo de variables producen mejores resultados, cada valor estandarizado a numérico se realiza de acuerdo al conocimiento laboral adquirido personalmente y de acuerdo a las reglas impuestas con el ministerio de relaciones laborales vigente en Ecuador.

Una vez estandarizadas las respuestas de todas las preguntas, se estandarizó los nombres de las variables con identificadores cortos que pueden ser utilizados en las gráficas para un análisis visualmente atractivo.

En la revisión de datos se elimina la variable “Are you self-employed” debido a que, hasta este punto de la limpieza y estandarización de datos únicamente se mantiene información de personas que cuentan con un empleo fijo y esta variable presenta un valor único.

3.3.1. Análisis estadístico de datos

Posterior a la transformación de datos del conjunto inicial, se realizó un análisis estadístico de datos que permiten tener una visión más amplia de ciertas variables que aportan un valor significativo al estudio.

Revisión estadística de empresas tecnológicas

En el conjunto de datos final, no se realizó un filtro respecto a las personas que están vinculadas directamente a empresas de tecnología, por tal motivo es necesario conocer de los encuestados cuales pertenecen y cuales no a empresas de tecnología.

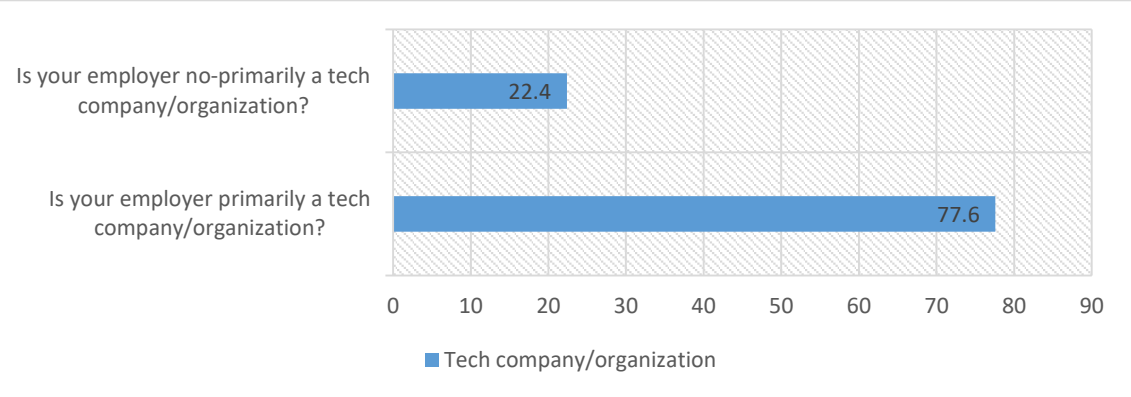

Figura 9. Trabajadores encuestados de empresas tecnológicas.

Del total de encuestados, se analizan datos únicamente de las personas que mantienen un trabajo estable, el 77,6% de las personas trabajan en empresas que se encuentran asociadas a la parte tecnológica. Como el objetivo de la investigación es el estudio de prevalencia y predictores, únicamente se selecciona la información del 77,6% de los encuestados que trabajan en empresas de tecnología o relacionadas a la parte tecnológica, quedando para análisis final 788 registros y 43 variables.

Desordenes de salud mental diagnosticados por profesionales de la salud en trabajadores tecnológicos

En la siguiente gráfica, se muestran los desórdenes identificados por profesionales de salud mental, se han segmentado en caso de que alguien respondiera a más de un desorden, con la finalidad de identificar la mayor cantidad de problemas entre los encuestados.

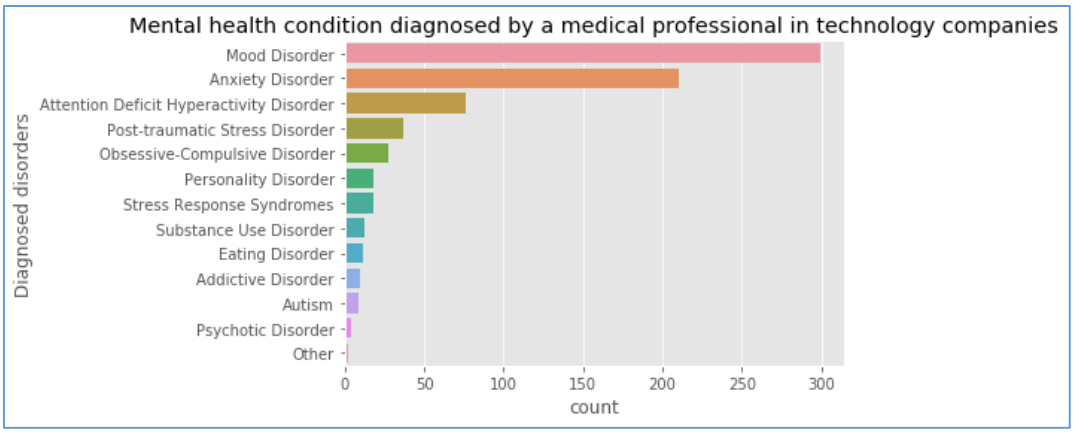

Figura 10. Condición de salud mental diagnosticada por un profesional en empresas tecnológicas.

Entre los desórdenes más frecuentes se encuentra el trastorno del estado de ánimo (depresión, trastorno bipolar, etc.), trastorno de ansiedad (generalizada, social, fobia, etc.), desorden hiperactivo y déficit de atención.

Problemas de salud mental por edad y género en encuestados

En el caso de las empresas tecnológicas, el modelo de negocio y las nuevas tecnologías están en constante cambio, siendo este un factor causante del estrés es necesario evaluar por edad y género la afectación que puede tener sobre un trabajador, por tal motivo se revisara los datos de las personas encuestas en relación al diagnóstico médico.

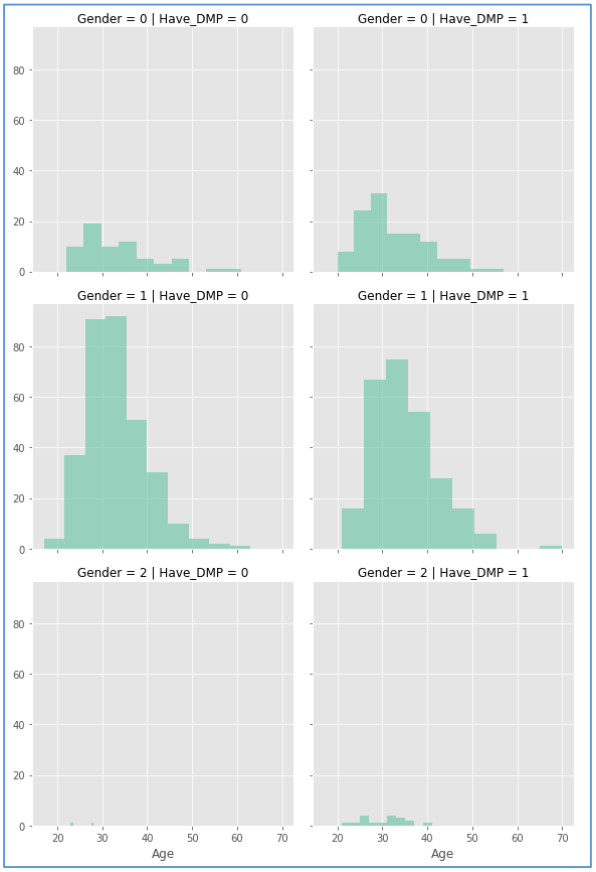

Figura 11. Identificación de problemas de salud mental por edad y género.

En las gráficas de la izquierda se muestran los encuestados que no presentan problemas de salud mental y en las gráficas de la derecha se muestran los encuestados diagnosticados con algún problema de salud mental. En la estandarización, el género se encuentra con Female (0), Male (1) y Genderqueer/Other (2). En el eje horizontal se presentan las edades de los encuestados y el eje vertical el número de encuestados.

De acuerdo esta información que se visualiza en las gráficas, se llega a las siguientes deducciones:

- La mayor parte de los encuestados oscilan entre los 20 años y 40 años de edad, y un patrón muy común es que, entre 25 años y 40 años de edad la mayor parte han sido diagnosticados con alguna enfermedad relacionada con problemas de salud mental.
- Los empleados hombres son menos propensos a ser diagnosticados con enfermedades de salud mental que las mujeres; sin embargo, de los encuestados con género diferente, el 90% ha sido diagnosticados con algún problema de salud mental, por tal motivo es importante ampliar más las preguntas de la encuesta para identificar la causa, debido a que su condición sexual podría desencadenar conflictos entre colaboradores afectando el rendimiento laboral.
- De los encuestados, se visualiza que las mujeres han sido identificadas con problemas de salud mental desde los 20 años, mientras que los hombres presentan desordenes a partir de los 23 años en casi igual volumen que las mujeres.

Salud Física vs Salud Mental: Predisposición para plantear problemas de salud en entrevista de trabajo

Entre los principales puntos del estudio realizado, es conocer la apertura de brindar información en una entrevista de trabajo por parte de un posible empleado con respecto a temas de salud, tanto física como mental.

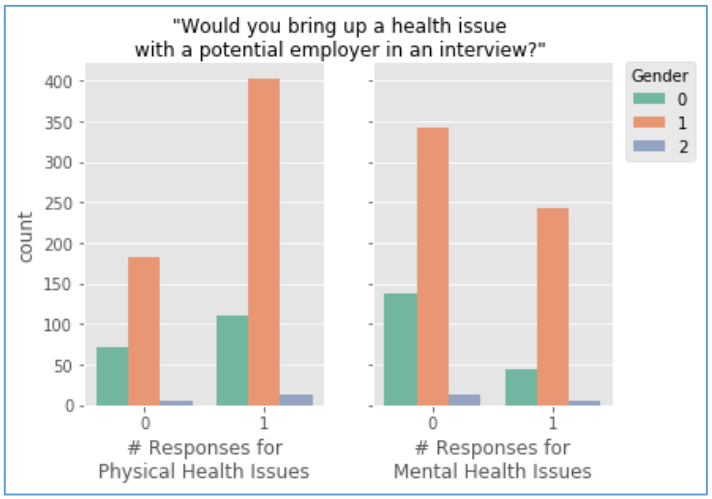

Figura 12. Salud física vs salud mental: entrevista de trabajo.

En el eje horizontal de las gráficas se encuentra la respuesta a la pregunta con No (0) y Yes (1), en el eje vertical se encuentran el total de respuestas y cada barra representa el género del encuestado. De acuerdo a las gráficas obtenidas, se puede evidenciar lo siguiente:

- Con respecto a problemas de salud físicos, existe mayor apertura de los hombres para hablar sobre algún problema físico, en el caso de las mujeres y personas de otro sexo la apertura es más reducida.
- Con respecto a problemas de salud mental, la apertura de las personas (indistintamente del género) para conversarlo con un posible empleador es reducido, en muchas de las empresas la salud mental de las personas a contratar es un factor importante para la contratación y cada empresa busca un rendimiento efectivo en las actividades asignadas por parte del nuevo recurso. En las entrevistas de trabajo una de las principales preguntas es ¿**Está acostumbrado al trabajo bajo presión?**, y el presentar algún tipo de desorden mental podría influir mucho en la respuesta.

Matriz de correlación de datos del conjunto de datos

Mediante la matriz de correlación se muestra los valores de correlación de Pearson, se exploró la relación de algunas preguntas tanto gráficamente como cuantitativamente, la finalidad es observar la relación que se tiene entre variables, un valor de correlación alto y positivo (color violeta) indica que los elementos miden una misma destreza o característica en común.

Si los elementos no están altamente correlacionados (color blanco a rosa), entonces pueden medir diferentes características o no estar claramente definidos, a partir de 0.75 se consideran variables altamente correlacionadas, esta información nos permite tener una visión general de las variables que más influyen en la clase principal **Have_DMP**.

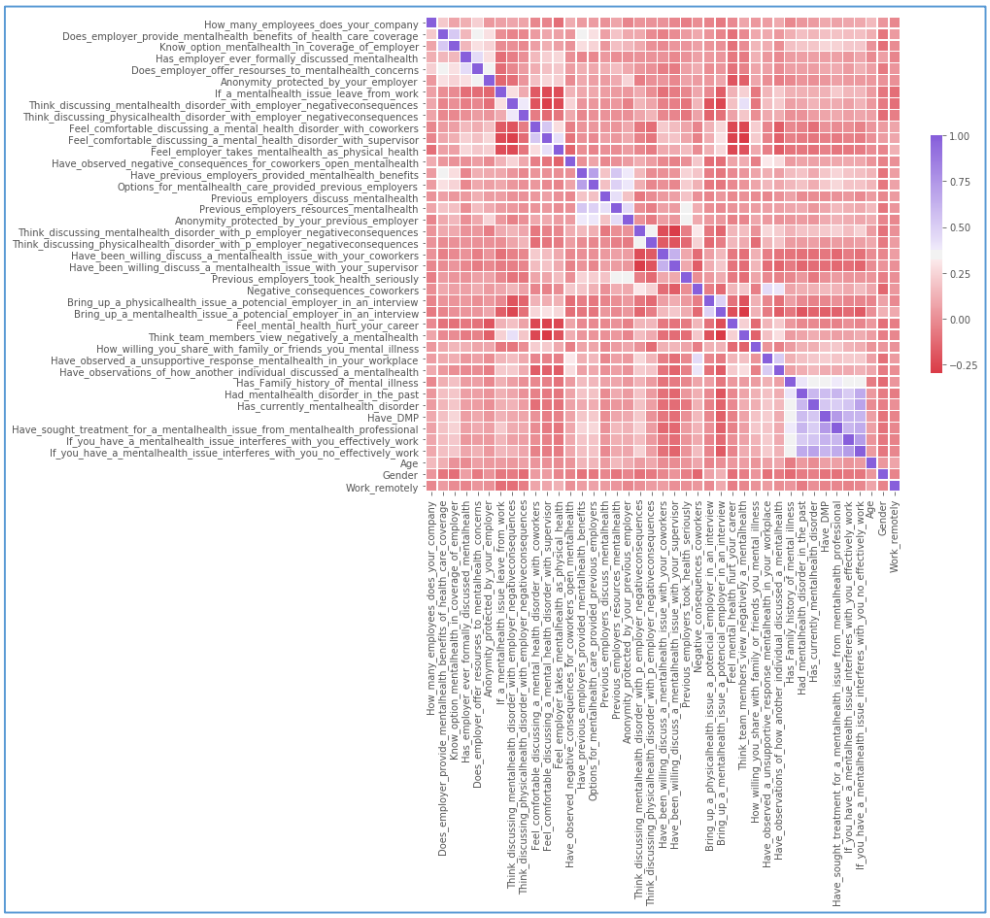

Figura 13. Matriz de correlación entre variables del conjunto de datos.

De las 43 variables del conjunto de datos, la relación entre preguntas es muy pobre, únicamente 16 variables superan la correlación de 0.5 y solo 4 variables tienen una relación que supera el 0.75, la relación más fuerte es plantearse la posibilidad de hablar de problemas de salud mental con el supervisor y los compañeros de trabajo, así mismo, sobre los beneficios de salud mental provistos por los anteriores empleadores y las diferentes opciones que cubría el seguro médico.

La revisión de estas relaciones entre variables brinda una pauta sobre aquellas variables que aportan mayor impacto al estudio en desarrollo.

Revisión de información personal y empresarial de los encuestados

Mediante la generación de un gráfico de tipo histograma de valores, se selecciona tres atributos de la encuesta con la finalidad de observar cómo se distribuyen las respuestas de estas preguntas en función del valor de la clase principal de **Have_DMP**, los atributos evaluados son:

- Número de empleados de la compañía.
- Antecedentes familiares de problemas mentales,
- Desordenes de salud mental identificados en el pasado a cada encuestado mediante un diagnóstico médico.

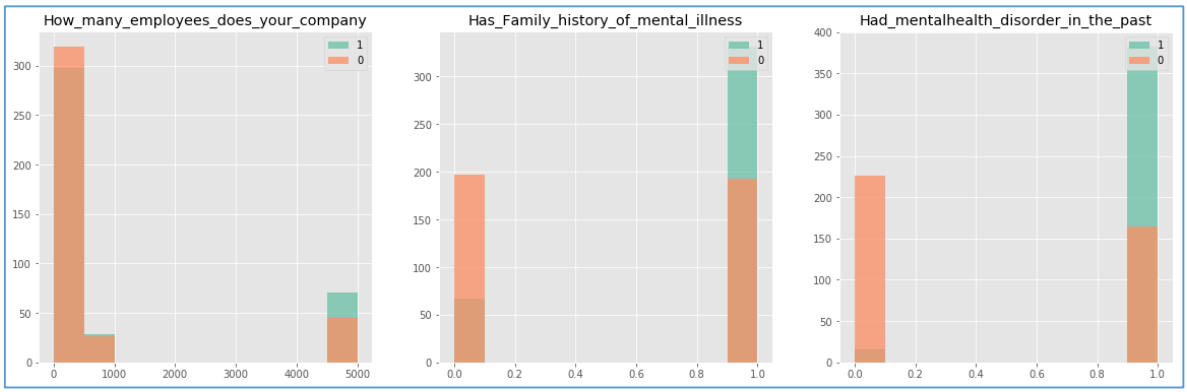

Figura 14. información personal y empresarial de los encuestados.

Para la generación de los histogramas, los 3 atributos analizados en base a un diagnóstico médico que se detalla a continuación:

- "How many employees does your company or organization have?": los empleados de la compañía se segmentan en proporciones de 5, 25, 100, 500, 1000, 5000.
- "Do you have a family history of mental illness?": esta pregunta tiene dos respuestas cerradas como Yes (1) y No (0).
- "Have you had a mental health disorder in the past?": esta pregunta tiene dos respuestas cerradas como Yes (1) y No (0).

Todas las respuestas se encuentran identificadas en el eje horizontal de las tres gráficas, en los ejes verticales se encuentran las estadísticas de las respuestas generadas por los encuestados, las personas diagnosticadas con alguna enfermedad mental se encuentran representadas por el color azul y las personas sin diagnostico están representadas con el color naranja. De los análisis realizados en cada gráfica se obtienen las siguientes conclusiones:

- Con respecto al número de empleados de cada empresa u organización, de 1 a 1000 empleados no hay mucha diferencia entre pacientes diagnosticados y pacientes no diagnosticados; sin embargo, en empresas de más de 1000 empleados si hay una diferencia de un 30% más de empleados que fueron diagnosticados con problemas de salud mental.
- Las personas que tienen antecedentes familiares con desórdenes mentales son en la gran mayoría quienes actualmente tienen algún desorden identificado. Las personas con problemas mentales y que actualmente no tienen ningún problema de salud mental presentan cifras muy bajas.
- Con respecto a la tercera gráfica, se visualiza la información de las personas que sufrieron algún desorden mental en el pasado y nuevamente fueron diagnosticadas con un desorden igual o similar. De las personas que no presentan desórdenes mentales, las respuestas con respecto a problemas de salud identificados en el pasado se encuentran equilibradas.

Estas estadísticas permiten tener una mayor visibilidad de la información que se está manipulando dentro del conjunto de datos, una vez aplicadas las técnicas de minería de datos conocemos cuales de estas variables influyen más como tendencia para el diagnóstico de enfermedades mentales.

3.4. Selección y aplicación de técnicas de minería de datos.

En esta fase, se revisan los tipos de variables y su adaptación a los modelos que nos permite generar conocimientos en base a cada objetivo específico. En este caso, se construyen modelos propios de la minería de datos.

Inicialmente, es necesario conocer los valores y porcentajes que contiene cada pregunta para tener una idea de los datos a utilizar para la selección de variables a utilizar en cada una de las técnicas. En esta parte, se realizó la selección de características que permiten reducir el sobreajuste, mejorar la precisión y reducir el tiempo de entrenamiento.

Del conjunto de datos final no se tomaron en cuenta 4 variables:

- Conditions_diagnosed.
- Country_work.
- Age.
- La cuarta variable Has_currently_mentalhealth_disorder redunda con la clase Have DMP y este diagnóstico no cuenta con el aval médico.

Todas estas características no presentan información que brinde valor a los predictores para la construcción de modelos.

A continuación, se procedió con la selección de atributos que tienen mejores características en sus datos y que contribuyeron más a la variable de predicción como **Have DMP (Have you been diagnosed with a mental health condition by a medical professional?)**, en este caso se utilizó RFE con el algoritmo de regresión logística para seleccionar los atributos principales mediante la eliminación de la característica recursiva.

```
# Feature Extraction with RFE
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
# load data

array = mentalhealthbd.values
X = array [:,0:39]
Y = array [:,38]

# feature extraction
model = LogisticRegression(solver='lbfgs')
rfe = RFE (model, 16)
fit = rfe.fit(X, Y)
print (("Num Features: %d") % (fit.n_features_))
print (("Selected Features: %s") % (fit.support_))
print(("Feature Ranking: %s") % (fit.ranking_))
```

Código 7. Regresión logística para selección de atributos.

Dando como resultado las 16 mejores variables:

```
Num Features: 16
Selected Features: [False False False  True False False False  True  True  True False  True
 False False  True False False  True False  True  True False False False
  True  True False  True False  True False  True  True False False False
 False False  True]
Feature Ranking: [24  5  6  1 23 13  2  1  1  1 15  1  4  3  1 11 19  1 22  1  1 12 14 16
  1  1  7  1 21  1 20  1  1 17 18  9 10  8  1]
```

Figura 15. Resultados de la selección de atributos.

Mediante la ejecución de RFE con el algoritmo de regresión logística, se realizó varias ejecuciones del algoritmo para la selección de variables en un rango de 3 a 18 principales, con la elección de 16 características del conjunto de datos que se acoplaron a las variables que superan el 0.5 de relación de acuerdo a la matriz de correlación, estas variables se muestran a continuación:

- Has_employer_ever_formally_discussed_mentalhealth
- Think_discussing_mentalhealth_disorder_with_employer_negativeconsequences
- Think_discussing_physicalhealth_disorder_with_employer_negativeconsequences
- Feel_comfortable_discussing_a_mental_health_disorder_with_coworkers
- Feel_employer_takes_mentalhealth_as_physical_health
- Options_for_mentalhealth_care_provided_previous_employers
- Anonymity_protected_by_your_previous_employer
- Think_discussing_physicalhealth_disorder_with_p_employer_negativeconsequences
- Have_been_willing_discuss_a_mentalhealth_issue_with_your_coworkers
- Bring_up_a_physicalhealth_issue_a_potencial_employer_in_an_interview
- Bring_up_a_mentalhealth_issue_a_potencial_employer_in_an_interview
- Think_team_members_view_negatively_a_mentalhealth
- Have_observed_a_unsupportive_rensponse_mentalhealth_in_your_workplace
- Has_Family_history_of_mental_illness
- Had_mentalhealth_disorder_in_the_past
- Work_remotely

En la revisión del conjunto de datos original, estas características presentan la información más relevante de la encuesta y permiten tomar las pautas para realizar la construcción de los modelos para la generación de conocimiento, en este caso es necesario validar los síntomas de prevalencia entre los encuestados y se construye un nuevo conjunto de datos con estas características.

3.5. Extracción de conocimiento

El objetivo es identificar síntomas de prevalencia en enfermedades mentales, validar hipótesis de empresas y prestaciones, pruebas mediante pre-procesado de datos con distintas variables a través de la reducción de la funcionalidad. En el presente estudio se implementarán las reglas de clasificación del aprendizaje supervisado.

De acuerdo a la selección de variables, se crea el nuevo conjunto de datos para la extracción de conocimiento con las 16 características más la clase principal (Have_DMP) y se analizan 788 registros.

Aplicación de técnica de clasificación supervisada: Árboles de Decisión para Clasificación

De las diferentes técnicas de clasificación supervisada, los arboles de decisión para la clasificación son los más utilizados en el mundo del aprendizaje automático debido a su sencillez del modelo, accesibilidad, explicación aportada por cada clasificación, representación gráfica y clasificación con nuevos patrones.

Para graficar un árbol de decisión es necesario tener una variable dependiente o clase, es nuestro conjunto de datos tenemos la variable **Have DMP (Have you been diagnosed with a mental health condition by a medical professional?)** y el objetivo del clasificador será con dicha clase averiguar patrones e identificar síntomas de prevalencia entre los trabajadores tecnológicos.

Se ha optado en el presente estudio por esta técnica de clasificación supervisada debido a que los árboles son capaces de manejar problemas de múltiples salidas, utiliza un modelo de caja blanca y permite la posibilidad de validar un modelo mediante pruebas estadísticas.

Para la extracción de conocimiento, se generó tres árboles con respecto a empresas de tecnología, los lineamientos de cada árbol se detallan a continuación:

- Árbol de clasificación: conjunto final de datos con 788 registros, 16 variables y una clase.
- Árbol de clasificación: subconjunto de datos con registros de empleados con trabajo remoto (en línea).
- Árbol de clasificación: subconjunto de datos con registros de empleados con trabajo en sitio.

Árbol de decisión: conjunto final de datos con 788 registros, 16 variables y una clase.

Para el primer árbol de decisión se utilizó el 80% de datos para entrenamiento y 20% de datos para prueba, el detalle de la cantidad de los datos se muestra a continuación:

- Entrenamiento X: (630, 16)
- Entrenamiento Y: (630,)

- Validación X: (158, 16)
- Validación Y: (158,)

La construcción del árbol con los datos de entrenamiento tiene las siguientes características por defecto:

```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=6,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best')
```

Figura 16. Características del árbol de decisión.

La revisión del aprendizaje del algoritmo con datos de entrenamiento y test brinda los siguientes resultados de precisión con el modelo inicial:

- Train: 0.806
- Test: 0.766

La matriz de confusión de las predicciones del grupo Test. La diagonal de esta matriz se lee: arriba a la izquierda **True Negatives** y abajo a la derecha **True Positives**.

53	27
10	**68**

En donde:

- 53 son los verdaderos positivos: número de encuestados con diagnóstico de salud mental en los que la prueba dio resultado acertado (diagnóstico correcto).
- 27 son los falsos positivos: número de encuestados sin problemas de salud mental en los que la prueba dio resultado acertado (diagnóstico incorrecto).
- 10 son los falsos negativos: número de encuestados con diagnóstico de salud mental en los que la prueba dio resultado no acertado (diagnóstico incorrecto).
- 68 son los verdaderos negativos: número de encuestados sin problemas de salud mental en los que la prueba dio resultado no acertado (diagnóstico correcto).

A continuación, se procede con la creación del árbol de decisión:

```
with open (r"tree1.dot", 'w') as f:
     f = tree.export_graphviz(classifier,
                              impurity = False,
                              feature_names = list(mentalhealthbd
Clean.drop(['Have_DMP'], axis=1)),
                              class_names = ['NO','YES'],
                              filled= False)
# Convertir el archivo .dot a png para poder visualizarlo
check_call(['dot','-Tpng',r'tree1.dot','-o',r'tree1.png'])
PImage("tree1.png")
```

Código 8. Generación del árbol de decisión del conjunto de datos principal.

El árbol generado del conjunto de datos final tiene una profundidad de 5 ramas, el árbol se ha divido en dos a partir de su nodo raíz para una mejor visualización de las ramas:

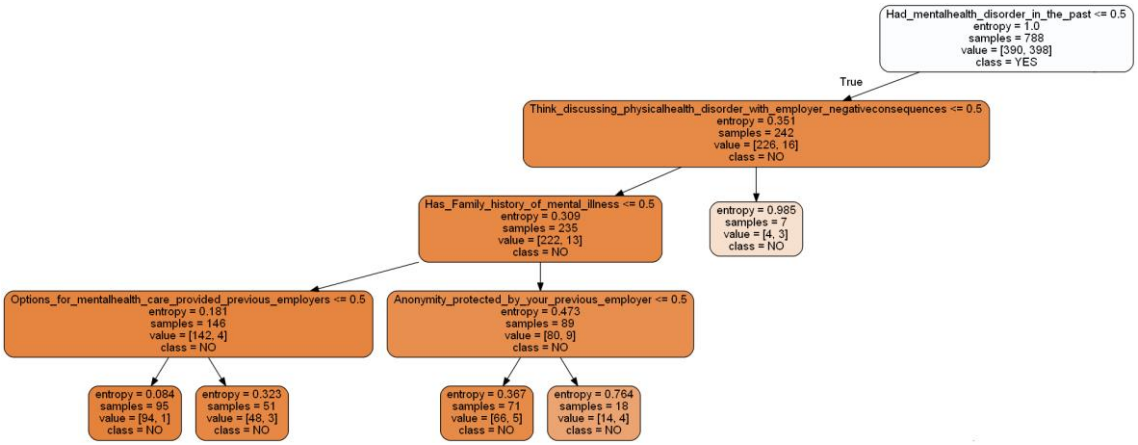

Figura 17. Parte 1 - Árbol de decisión para clasificación del conjunto de datos original.

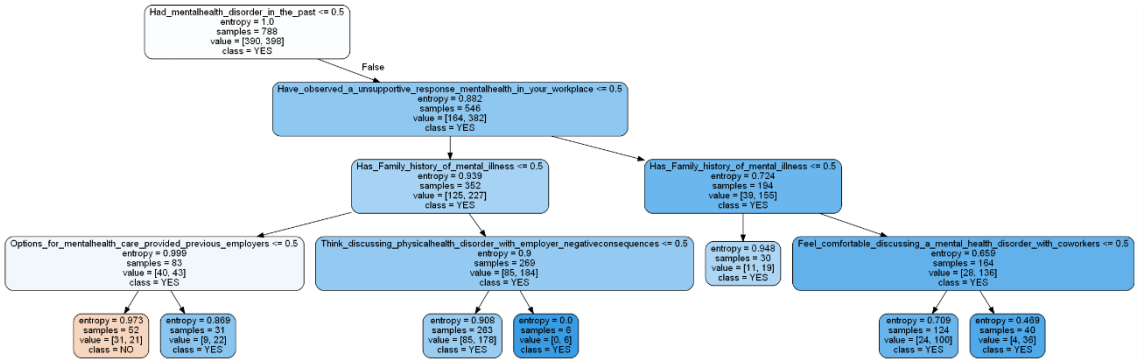

Figura 18. Parte 2 - Árbol de decisión para clasificación del conjunto de datos original.

En este primer árbol de clasificación se analizaron 788 encuestas, podemos apreciar que el nodo raíz **Desórdenes mentales identificados en el pasado** hace una primera división hacía la izquierda con todas las respuestas negativas y hacia la derecha con todas las respuestas afirmativas. El árbol de decisión asigna el color azul a los diagnósticos médicos positivos mientas que de color naranja a los diagnósticos médicos negativos.

La importancia de las 17 variables del conjunto de datos en el árbol de decisión se muestra a continuación:

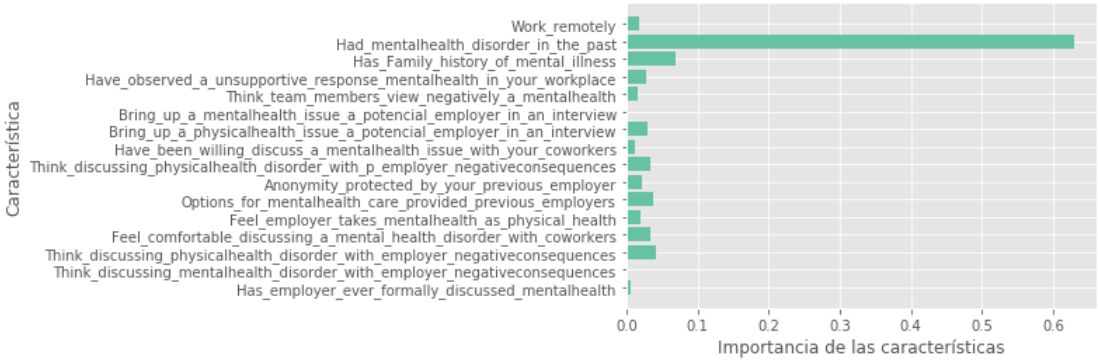

Figura 19. Importancia de variables en árbol de decisión del conjunto de datos original.

La variable con respecto a los antecedentes de desorden mental tiene mayor importancia en la construcción del árbol, mientras que el resto de variables aportan muy poca información y algunas de estas variables no son contempladas en el mismo, con una profundidad de 5 ramas y resultados binarios se generaron los mejores resultados de entrenamiento y validación.

En este caso se tienen las siguientes observaciones:

- Con respecto a los encuestados que tuvieron desórdenes mentales en el pasado se contabilizan 546 encuestados en donde el 90,5% sufre de algún problema de salud mental en la actualidad, de los cuales se tienen las siguientes observaciones:
    - De las personas que sufrieron desórdenes mentales en el pasado y no cuentan con antecedentes familiares, el 9.5% no presenta problemas de salud mental.
    - Únicamente 31 personas (5.7%) conocían de los beneficios de salud mental de sus anteriores empleadores.
    - El 35.5% de las personas han observado o experimentado una respuesta poco favorable o mal manejada a un problema de salud mental en su lugar de trabajo.
    - El 79.3% de los encuestados tienen antecedentes familiares con desorden mental.
    - El 48.2% piensa que discutir sobre problemas de salud física no traerá consecuencias negativas con sus empleadores.
    - El 7.32% se sentiría cómodo discutiendo un trastorno de salud mental con sus compañeros de trabajo.
- Se contabilizan 242 encuestados que no presentan desórdenes mentales en el pasado y de los mismos el 36.8% cuenta con antecedentes familiares de desórdenes mentales. Así mismo, solo el 21.1% de los encuestados conocía los beneficios de salud mental de sus anteriores empleadores.

Del conjunto final se realizaron dos subconjuntos, uno para trabajadores con soporte remoto y otro para trabajadores en sitio, de los cuales se tiene los siguientes resultados:

	Datos del árbol con datos de encuestados con trabajo remoto	Datos del árbol con datos de encuestados con trabajo en sitio
Valores de train	X(156,16), Y (156)	X(474,16), Y (474)
Valores de test	X(39,16), Y (39)	X(119,16), Y (119)
Precisión train	0.853	0.823
Precisión test	0.769	0.815
Verdaderos positivos	17	40
Falsos positivos	2	15
Falsos negativos	7	7
Verdaderos negativos	13	57

Tabla 3. Detalle de árboles de decisión por forma de trabajo.

El árbol generado a partir del subconjunto de datos del conjunto final, se muestra una mejor precisión a nivel de entrenamiento y prueba con los datos de las personas que trabajan en sitio, superando el 80% de efectividad, y con mejores datos a nivel de verdaderos positivos y verdaderos negativos.

Este árbol tiene una profundidad de 5 ramas, el árbol se ha divido en dos a partir de su nodo raíz para una mejor visualización:

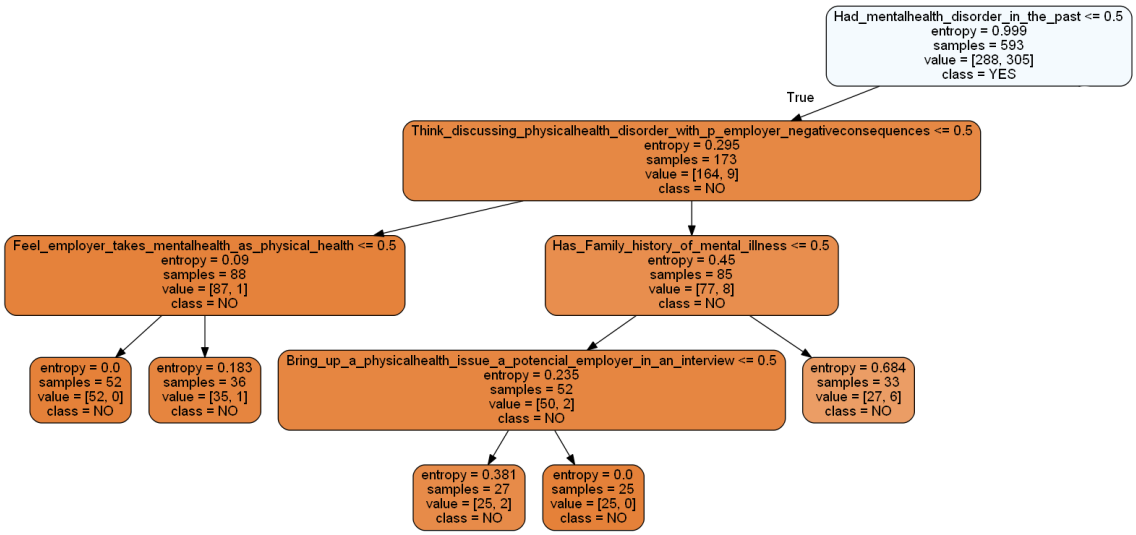

Figura 20. *Parte 1 - Árbol de decisión para clasificación del subconjunto de trabajadores en sitio.*

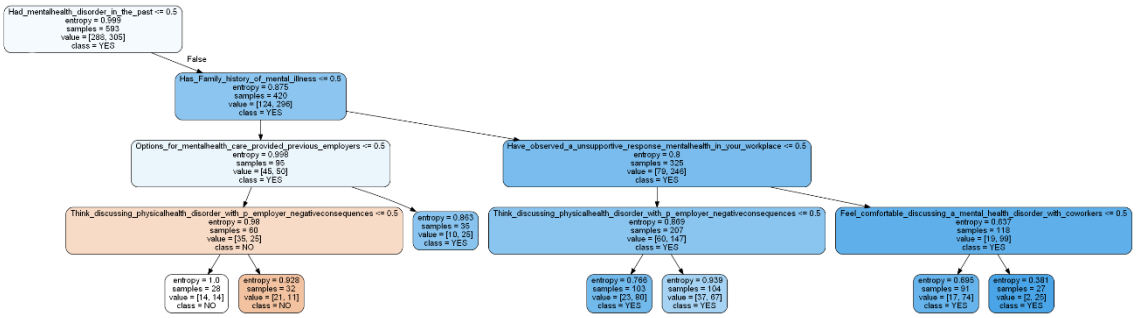

Figura 21. Parte 2 - Árbol de decisión para clasificación del subconjunto de trabajadores en sitio.

La variable con respecto a los antecedentes de desorden mental tiene mayor importancia en la construcción del árbol al igual que el anterior, mientras que el resto de variables son complementarias, con igual profundidad de ramas y resultados binarios se generaron mejores resultados de entrenamiento y validación que el árbol inicial que trabajo con el conjunto de datos final.

En este caso, se tienen las siguientes observaciones:

- Con respecto a los encuestados que tuvieron desórdenes mentales en el pasado se contabilizan 420 encuestados en donde el 85.7% sufre de algún problema de salud mental, algunas observaciones complementarias se detallan a continuación:
    - De las personas que sufrieron desórdenes mentales en el pasado y no cuentan con antecedentes familiares, el 14.3% no presenta problemas de salud mental en la actualidad
    - El 8.3% de encuestados conocían de los beneficios de salud mental de sus anteriores empleadores.
    - El 24.8% de los encuestados considera que hablar de sus problemas físicos traería consecuencias negativas con sus anteriores empleadores.

- Existen 173 encuestados que no presentan desórdenes mentales en el pasado y de los mismos el 19% cuenta con antecedentes familiares de desórdenes mentales.

3.6. Interpretación y evaluación de datos.

Una vez aplicada la técnica de clasificación supervisada mediante los árboles de decisión para la clasificación se plantean las conclusiones finales de acuerdo a los objetivos planteados para este estudio, en este caso se realizó también un análisis estadístico que permite fundamentar los resultados de cada uno de los árboles obtenidos.

Del total de encuestas, son 788 encuestas que cumplen con los parámetros requeridos para el estudio, como principal característica se analizaron las encuestas de empleados que están vinculados a empresas de tecnología sin importar el perfil o cargo, para la revisión de los síntomas de prevalencia entre los encuestados se analiza de manera independiente de acuerdo a la forma de trabajo (remoto o en sitio) en la primera fase.

La interpretación de los resultados se detalla a continuación y se segmenta en varios puntos de estudio:

- Resultados generales.
    - Del total de encuestados, el 77.6% corresponden a empresas u organizaciones de tecnología.

- Desordenes de salud mental con mayor diagnóstico.
    - Entre los síntomas identificados en los encuestados se encuentra el trastorno de ansiedad (generalizado, social, fobia, etc.), trastorno del estado de ánimo (depresión, trastorno bipolar, etc.) y trastorno por déficit de atención e hiperactividad.
    - En varios de los escenarios evaluados, estos desordenes pueden desencadenar una serie de trastornos, por lo cual al primer síntoma es necesario evaluarlos y tratarlos de forma adecuada.

- Problemas de salud mental por edad y género.
    - Los empleados hombres son menos propensos a ser diagnosticados con enfermedades de salud mental que las mujeres; sin embargo, de los pacientes que se identifican con un género diferente el 90% de los encuestados han sido diagnosticados con algún problema de salud mental, por tal motivo es importante ampliar más las preguntas de la encuesta puesto que su condición sexual podría desencadenar conflictos internos entre colaboradores y afectar a su rendimiento.

- Antecedentes de salud mental y apertura para la búsqueda de ayuda.
    - Existe un volumen muy elevado de las personas que sufrieron algún tipo de desorden en el pasado y actualmente lo tienen, así mismo, los antecedentes familiares son factores principales para desarrollar algún tipo de desorden.
    - La apertura de todos los encuestados para discutir o buscar ayuda cuando tienen algún tipo de problema de salud es muy baja, siendo un reto para cada una de las empresas en saber tratar cada uno de estos factores.
    - Aquellas personas que tuvieron un desorden mental en el pasado y sintieron que esto interfería en su trabajo, fue menos del 25% de

personas que en su anterior empresa se habló formalmente sobre la salud mental (como parte de una campaña de bienestar u otra comunicación oficial).

Para el estudio de las diferentes variables se utilizaron modelos supervisados, mediante el uso de árboles de decisión para la clasificación en el conjunto de datos final se tienen los siguientes resultados:

- En la revisión del aprendizaje del algoritmo con datos de entrenamiento y validación se obtuvo un entrenamiento de 0.806 de precisión y 0.766 de precisión en la validación.
- Las variables que mayor influencia tienen en los árboles de decisión son:
  - Desordenes de salud mental en el pasado.
  - Antecedentes familiares de problemas de salud mental.
  - Apertura del trabajador para hablar de problemas de salud mental con sus empleadores, compañeros de trabajo y entrevistas de trabajo.
  - Beneficios de salud mental por empleados anteriores.
- Con respecto a los encuestados que tuvieron desórdenes mentales en el pasado se contabilizan 546 encuestados en donde el 90,5% sufre de algún problema de salud mental en la actualidad.
- De las personas que sufrieron desórdenes mentales en el pasado y no cuentan con antecedentes familiares, el 9.5% no presenta problemas de salud mental.
- El 79.3% de los encuestados tienen antecedentes familiares con desorden mental.

Posterior a la interpretación de resultados, evaluaremos el conjunto de datos final en la plataforma de software para el aprendizaje automático WEKA [46], en el mismo contamos con 788 registros y 7 variables.

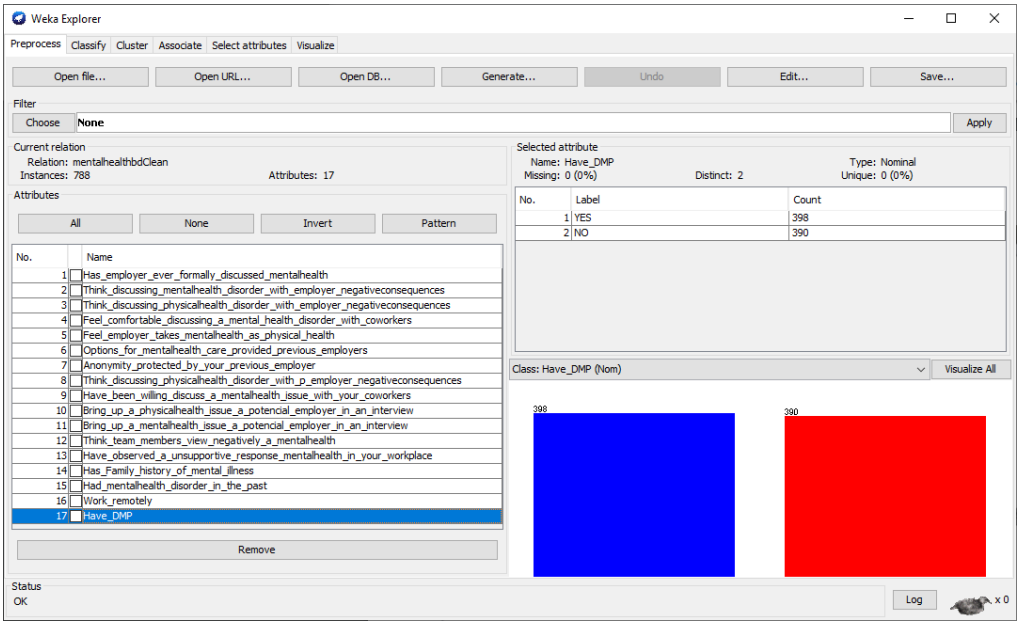

Figura 22. Carga del conjunto de datos final en WEKA.

De acuerdo a los datos reales, tenemos 398 personas (50.5%) con problemas de salud mental de los 788 encuestados, dos variables que influyen en los desórdenes mentales son los antecedentes familiares y los antecedes de desórdenes de salud mental.

Para validar los resultados del primer árbol de decisión se analizarán otros árboles de decisión para la clasificación en WEKA con un 80% de entrenamiento y un 20% de validación, los resultados se detallan a continuación:

	Decision Tree Classifier in Python	Naive Bayers Tree	J48
Instancias correctamente clasificadas	0.7658%	73.4177%	70.8861%
Instancias incorrectamente clasificadas	0.2342%	26.5823%	29.1139%
Verdaderos positivos	45	57	55
Falsos negativos	33	14	16
Falsos Positivos	12	28	39
Verdaderos negativos	68	59	57

Tabla 4. Porcentajes de clasificación con árboles de decisión.

Con estos resultados, se tiene la probabilidad de realizar una validación de un nuevo encuestado con al menos un 70% de seguridad que el valor resultante es el correcto.

Por cada uno de los objetivos planteados, se procede a realizar un análisis sobre las conclusiones o respuestas obtenidos por cada uno, esto previo a la definición de conclusiones y futuros proyectos.

1. Identificación de síntomas de prevalencia entre trabajadores tecnológicos y la afectación por género del trabajador.
    a. Con respecto a los síntomas de prevalencia, los factores que más influyen son antecedentes mentales identificados anteriormente y antecedentes familiares de desorden mental.
    b. Entre las afectaciones por género, aquellas personas con género diferente son las más afectadas siendo un 90% al menos las que sufren algún tipo de desorden mental. En hombres se cuenta con una elevada presencia de desórdenes entre los 25 años y 40 años mientras que, en las mujeres existe mayor presencia entre los 25 años y 30 años
2. Evaluar la apertura de los trabajadores para buscar ayuda y las prestaciones de las organizaciones para tratar síntomas en sus colaboradores en base a factores de apertura/ayuda.
    a. Existe poca apertura para hablar de temas de salud física y mental frente a empleados y compañeros de trabajo.
    b. La mayor parte de las personas no se han planteado la posibilidad de plantear estos inconvenientes de salud mental a sus empleadores, la búsqueda de ayuda es reducida.
    c. Respecto a las prestaciones, no se tiene mayor detalle del tipo de seguro y las prestaciones del mismo con respecto a salud mental.
3. Evaluar si el tipo de empresa (grande, mediana o pequeña) afecta a la proporción de los beneficios y prestaciones para la atención mental de sus colaboradores.
    a. En las empresas de hasta 1000 trabajadores, las estadísticas presentan diagnósticos equitativos entre personas que padecen algún desorden mental o no, en casos de empresas con más de 1000 empleados los diagnósticos de salud mental incrementan en un 50%.
    b. Cerca del 95% de las personas que han sido diagnosticas con algún problema de salud mental desconocen los beneficios que brindan las empresas, ya sea por falta de comunicación de las mismas o por falta de prestaciones de los seguros contratados.

Con el estudio realizado para el cumplimiento de los objetivos específicos, se han llegado a resultados satisfactorios que permiten plantear conclusiones y, permiten dar un enfoque a trabajos futuros con respecto al manejo de la recolección de la información y recomendaciones orientadas a la empresa en temas de contratación de nuevo personal y el alcance que deben tener los seguros que brindan en la actualidad.

4. Conclusiones y futuros proyectos

De acuerdo al estudio realizado con los resultados de las encuestas realizadas por OSMI en el año 2016, se han llegado a las siguientes conclusiones:

a. Entre los síntomas de prevalencia de salud mental entre trabajadores tecnológicos o trabajadores que se encuentran vinculados a áreas de tecnología se encuentran los antecedentes de problemas de salud mental y también, influyen factores de antecedentes familiares. Este tipo de desórdenes tiene mayor impacto entre los 25 años y 40 años, en el caso de las personas con género diferente existen un 90% de probabilidad de presencia de algún tipo de desorden.
b. Las estadísticas de apertura de los trabajadores para búsqueda de ayuda son muy reducidas, en muchos de los casos no es una opción considerada por las personas para manifestarla en sus ambientes de trabajo o en entrevistas de trabajo. Esta falta de apertura llega a tener una relación directa con respecto al alcance de las prestaciones que brindan las organizaciones en temas de salud.
c. A nivel de empresas, se tiene información de prestaciones de los empleadores anteriores en donde no se conoce el tipo de empresa. De acuerdo a los datos estadísticos, existe una mayor presencia de personas diagnosticadas con algún tipo de desorden en empresas con más de 1000 empleados; sin embargo, a nivel de beneficios que brinden las empresas existe desconocimiento de los mismos por parte de los encuestados.

Los objetivos específicos fueron planteados acuerdo a una breve revisión de las variables o preguntas de las encuestas realizadas en la primera fase, durante la limpieza de datos, estandarización de información y análisis de la importancia de las 63 variables, para el trabajo final únicamente se revisaron 17 variables que aportan un valor significativo al estudio realizado. Con esta información se ha complicado 2/3 objetivos de forma significativa.

Para la revisión del conjunto de datos, las fases de la minería de datos aportaron de forma acertada al cumplimiento de la planificación y, sobre todo a reforzar conocimientos sobre otras técnicas para el análisis de datos. En este caso, se realizó una validación adicional mediante WEKA para los árboles de decisión con la finalidad de fundamentar los resultados de entrenamiento y validación.

De acuerdo a las conclusiones, se procede a detallar las líneas de trabajo futuro:

- Estudio aplicado a diferentes empresas de acuerdo al modelo de negocio, con la finalidad de mejorar los resultados de predicción de los árboles de decisión.
- Presentación de estudios al Ministerio de Relaciones Laborales para gestionar cambios a las normativas vigentes de cumplimiento de las organizaciones, en donde se permita ampliar los beneficios al trabajador en temas de salud mental.

Estas actividades a futuro permitirán mejorar las prestaciones a los empleados indistintamente del tipo de organización, con la finalidad de garantizar un ambiente de trabajo agradable y con las prestaciones necesarias que garanticen la menor afectación o cambio en caso de diagnósticos de desórdenes de salud mental en colaboradores.

5. Glosario

- **ADA:** Ley sobre estadounidenses con Discapacidades
- **Ausentismo:** alude a la inasistencia de una persona al sitio donde debe cumplir una obligación o desarrollar una función.
- **IBM:** International Business Machines Corporation
- **Kaggle:** comunidad en línea de científicos de datos y aprendices de máquinas, propiedad de Google LLC
- **OMS:** Organización mundial de la salud.
- **Presentismo:** Sensación o sospecha que algo va a ocurrir.
- **Prueba exacta de Fisher:** se basa en la distribución hipergeométrica, es una prueba de significación estadística utilizada en el análisis de tablas de contingencia.
- **Psicosis o sicosis:** Enfermedad mental grave que se caracteriza por una alteración global de la personalidad acompañada de un trastorno grave del sentido de la realidad.
- **Tecnología disruptiva:** aquella tecnología o innovación que conduce a la aparición de productos y servicios frente a una estrategia sostenible a fin de competir contra una tecnología dominante, buscando una progresiva consolidación en un mercado.
- **TIC's:** Tecnologías de la Información y la Comunicación. Se considera al conjunto de tecnologías, tales como: software, soportes, canales, herramientas, etc., que permiten al usuario acceder, almacenar, transmitir y procesar la información.

6. Bibliografía

[1]. World Health Organization. World Health Report 2001: Mental Health – New Understanding, new hope. Francia, 2002.
[2]. https://www.corporatewellnessmagazine.com/article/tech-manage-mental-health-workplace. (18/03/2019).
[3]. https://www.paho.org/hq/index.php?option=com_content&view=article&id=7305:2012-dia-mundial-salud-mental-depresion-trastorno-mental-mas-frecuente&Itemid=1926&lang=es. (02/03/2019).
[4]. https://www.kaggle.com/osmi/mental-health-in-tech-2016/home (02/03/2019)
[5]. Pressman, Roger. Ingeniería de software, un enfoque práctico. Séptima Edición. McGrawHill, México, 2010.
[6]. Zachary Steel, Marnane, Claire, Changiz Iranpour, Tien Chey, John W Jackson, Vikram Patel and Derrick Silove. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013, International Journal of Epidemiology, 476–493, 2014.
[7]. https://www.obs-edu.com/int/blog-project-management/gestion-del-tiempo/las-nuevas-tecnologias-actual-origen-del-estres-laboral (19/03/2019)
[8]. https://osmihelp.org/about/about-osmi (19/03/2019)
[9]. https://www.who.int/mental_health/in_the_workplace/es/ (18/03/2019).
[10]. OMS, autores varios. Prevalencia, severidad y necesidad insatisfecha de tratamiento de trastornos mentales en la Organización Mundial de la Salud Encuestas mundiales de salud mental. JAMA, 2581-2590, 29121, 2004
[11]. BRILLHART, Technostress in the Workplace. Managing stress in the electronic workplace. Journal of American Academy of Business, Vol. 5, Cambridge, 2004
[12]. OSMI, Mental Health in Tech: Guidelines for Employees, Open Sourcing Mental Illness, Ltd, 2018.
[13]. Job Accommodation Network. Accommodation and compliance series: Employees with mental health impairments, 2015
[14]. De prado, Ana. Nuevas tecnologías y nuevos riesgos laborales: estrés y tecnoestrés, Universidad de Sevilla, 2016.
[15]. https://www.lavanguardia.com/economia/20150609/54432714429/el-37-de-empleados-muestra-estres-y-presion-por-la-digitalizacion-laboral.html (18/03/2019)
[16]. Salanova, Marisa. Nuevas tecnologías y nuevos riesgos psicosociales en el Trabajo. Universitat Jaume I (Castellón), 2007.
[17] Brod, Craig. TECHNOSTRESS: The Human Cost of the Computer Revolution. Basic Books, Addison-Wesley, 1984.
[18]. Aragüez, Lucía. El impacto de las tecnologías de la información y de la comunicación en la salud de los trabajadores: El Tecnoestrés. Universidad de Málaga, Málaga, 2017.
[19]. Mullarkey, Sean. Jackson, Paul. Wall, Toby. Wilson, John and Grey-Taylor, Susan. The impact of technology characteristics and job control on worker mental health. Journal Of Organizational Behavior, Vol. 18, 471-489, 1997.
[20]. http://workplacementalhealth.org/Mental-Health-Topics/Workplace-Stress (19/03/2019).
[21]. GRAY, P. Mental health in the workplace: Tackling the effects of stress. London: Mental Health Foundation, 2003.
[22]. Arghami, Sh, Seraji, J Nasl. Mohammad, K. Zamani, Gh. Farhangi, A. Vuuren, Van. Mental Health in High-Tech System, Iranian J Publ Health, vol. 34, pp. 31-37, 2005.
[23]. https://www.import.io/post/data-mining-machine-learning-difference/ (20/03/2019)
[24]. http://www.7puentes.com/blog/2018/05/14/machine-learning-en-marketing-7-aplicaciones-para-impulsarlo/ (20/03/2019)
[25]. Mor, Enric. Sangüesa, Ramon. Molina, Luis. Data Mining. Editorial UOC.

[26]. Moreno, María. Miguel, Luis. García, Francisco. Polo, José. Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. Universidad de Salamanca. 2001.
[27].https://www.ecured.cu/Algoritmos_de_clasificaci%C3%B3n_no_supervisada (20/03/2019)
[28]. https://www.interactivechaos.com/manual/tutorial-de-machine-learning/algoritmos-no-supervisados (20/03/2019).
[29]. Garcia, Cristina. Gómez, Cristina. Algoritmos De Aprendizaje: KNN & KMEANS. Universidad Carlos III de Madrid. 2009.
[30]. Cáceres, Jesús. Reconocimiento de patrones y el aprendizaje no supervisado. Universidad de Alcalá. 2006.
[31]. http://www.cs.us.es/~fsancho/?e=77 (20/03/2019).
[32]. https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/ (20/03/2019).
[33]. https://medium.com/@juanzambrano/aprendizaje-supervisado-o-no-supervisado-39ccf1fd6e7b (20/03/2019).
[34]. https://es.wikipedia.org/wiki/Aprendizaje_supervisado (20/03/2019).
[35]. https://scikit-learn.org/stable/tutorial/machine_learning_map/ (20/03/2019).
[36]. https://es.wikipedia.org/wiki/Aprendizaje_no_supervisado (20/03/2019).
[37]. https://www.elsaltodiario.com/salud-mental/conseguiran-las-maquinas-predecir-el-suicidio (20/03/2019).
[38]. https://medicalxpress.com/news/2018-07-machine-treatment-outcomes-schizophrenia.html (20/03/2019).
[39]. https://www-03.ibm.com/press/ar/es/pressrelease/51343.wss (20/03/2019).
[40]. Alonso, S. G., de la Torre-Díez, I., Hamrioui, S., López-Coronado, M., Barreno, D. C., Nozaleda, L. M., & Franco, M. (2018). Data mining algorithms and techniques in mental health: A systematic review. Journal of Medical Systems, 42(9) doi:10.1007/s10916-018-1018-2.
[41]. Dmonte, S., Tuscano, G., Raut, L., & Sherkhane, S. (2018). Rule generation and prediction of anxiety disorder using logistic model trees. Paper presented at the 2018 International Conference on Smart City and Emerging Technology, ICSCET 2018, doi:10.1109/ICSCET.2018.8537258.
[42]. Salah-Eddine, M., Belaissaoui, M., Alami, A. E., & Salah-Eddine, K. (2019). Technostress management through data mining. Journal of Management Information and Decision Science
[43]. Baba, T., Baba, K., & Ikeda, D. (2020). Detecting mental health illness using short comments doi:10.1007/978-3-030-15032-7_23
[44]. Almeida, H., Briand, A., & Meurs, M. -. (2017). Detecting early risk of depression from social media user-generated content. Paper presented at the CEUR Workshop Proceedings, 186.
[45]. Hao, F., Zhao, L., Xu, T., & Dong, X. (2019). Application of negative and positive association rules in mental health analysis of college students. Paper presented at the ICNC-FSKD 2018 - 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, 717-723. doi:10.1109/FSKD.2018.8687181
[46] https://www.cs.waikato.ac.nz/ml/weka/ (05/06/2019)

7. Anexos

Anexo 1: Configuración del entorno de desarrollo

El entorno de desarrollo utilizado para la investigación se basa en la plataforma de ciencia de datos Anaconda para Python 3.7:

https://repo.anaconda.com/archive/Anaconda3-2019.03-MacOSX-x86_64.pkg

Anexo 2: Detalle de librerías utilizadas

```
import numpy as np
import pandas as pd
import seaborn as sns
import sklearn.metrics
import matplotlib.pyplot as plt
import graphviz

from sklearn.model_selection import KFold, cross_val_score, train_te
st_split
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn.metrics import classification report, accuracy score, c
onfusion_matrix
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from IPython.display import Image as PImage
from subprocess import check_call
from sklearn.tree import export graphviz
from PIL import Image, ImageDraw, ImageFont
```

Código 9. Librerías de Python

Anexo 3: Código implementado.

```
pearson = mentalhealthbd.corr(method = 'pearson')
f, ax = plt.subplots(figsize=(11, 9))
cmap = sns.diverging_palette(10, 275, as_cmap=True)
sns.heatmap(pearson, cmap=cmap,square=True,
            xticklabels=True, yticklabels=True,
            linewidths=.5, cbar_kws={"shrink": .5}, ax=ax)
```

Código 10. Matriz de correlación

```
array = mentalhealthbd.values
X = array[:,0:39]
Y = array[:,38]
# feature extraction
model = LogisticRegression(solver='lbfgs')
rfe = RFE(model, 16)
fit = rfe.fit(X, Y)
print(("Num Features: %d") % (fit.n_features_))
print(("Selected Features: %s") % (fit.support_))
print(("Feature Ranking: %s") % (fit.ranking_))
```

Código 11. Selección de variables con mejores características.

```
Classifier = DecisionTreeClassifier(max_depth = 5, criterion='entrop
y')
classifier.fit(pred_train,tar_train)
```

Código 12. Árbol de entrenamiento.

```
sklearn.metrics.confusion_matrix(tar_test,predictions)
```

Código 13. Matriz de confusión.

```
with open(r"tree1.dot", 'w') as f:
     f = tree.export_graphviz(classifier,
                              impurity = False,
                              feature_names = list(mentalhealthbdCle
an.drop(['Have_DMP'], axis=1)),
                              class_names = ['NO','YES'], #SI [1] y
NO [0]
                              filled= False )

# Convertir el archivo .dot a png para poder visualizarlo
check_call(['dot','-Tpng',r'tree1.dot','-o',r'tree1.png'])
PImage("tree1.png")
```

Código 14. Generación de la gráfica del árbol de decisión.

```
caract = predictors.shape[1]
plt.barh(range(caract),classifier.feature_importances_)
plt.yticks(np.arange(caract),list(mentalhealthbdClean.drop(['Have_DM
P'], axis=1)))
plt.xlabel('Importancia de las características')
plt.ylabel('Característica')
plt.show()
```

Código 15. Importancia de variables en el árbol de decisión.