Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 19 de julio de 2024

Autor: María de los Ángeles Rosero Montilla, Brayan Steven Ramírez Cortes, Juan Manuel Vivas Torres

Título del Trabajo de Grado: "Predicción De Variables En Salud Mental Para Colaboradores De Una Universidad Privada Ubicada En La Ciudad De Cali Por Medio De Aprendizaje Automático"

Director:

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

Firma del Director del Trabajo de Grado

Santiago de Cali, 7 de junio del 2024

Doctor
Diego Luis Linares Ospina
Director Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

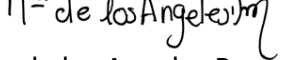
Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "Predicción De Variables En Salud Mental Para Colaboradores De Una Universidad Privada Ubicada En La Ciudad De Cali Por Medio De Aprendizaje Automático", el cual fue realizado por el (los) estudiante (s) María de los Angeles Rosero Montilla, Brayan Steven Ramírez Cortes, Juan Manuel Vivas Torres con código (s) 8980512, 8980266, 8980511 pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección de Julián Gil González.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,

Brayan Steven Ramírez Cortes
C.C. 1012406577 de Bogotá – Cundinamarca

Julián Gil González
C.C. 1088286439 de Pereira



María de los Angeles Rosero Montilla
C.C. 1089078168 de San Pedro de Cartago – Nariño

Jimena Botero Sarassa
C.C. 42123749 de Pereira

Juan Manuel Vivas Torres
C.C. 1061803782 de Popayán - Cauca

Documentación anexa:
Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

FICHA RESUMEN

TÍTULO: Predicción De Variables En Salud Mental Para Colaboradores De Una Universidad Privada Ubicada En La Ciudad De Cali Por Medio De Aprendizaje Automático

1. ÁREA DE TRABAJO: Salud mental, Machine Learning
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicado
3. ESTUDIANTE(S): María de los Ángeles Rosero Montilla, Brayan Steven Ramírez Cortes, Juan Manuel Vivas Torres
4. CORREO ELECTRÓNICO: mariaroseromontilla@javerianacali.edu.co , brayanrmz@javerianacali.edu.co , juanvivastorres@javerianacali.edu.co
5. DIRECCIÓN Y TELÉFONO:
   Juan Manuel Vivas: Calle 70 Norte #9-56 Popayán, Cauca, teléfono 3137014715
   María de los Ángeles Rosero: Carrera 8 calle 24 AN -Torres del rio, bloque H Apt 203 Popayán – Cauca, Teléfono 3215899819
   Brayan Ramírez: Diag 38 # 19-82 Bogotá, Teléfono 3046376838.
6. DIRECTOR: Julián Gil González
7. VINCULACIÓN DEL DIRECTOR: Profesor Catedrático
8. CORREO ELECTRÓNICO DEL DIRECTOR: julian.gil@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica): Jimena Botero
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica): No aplica
11. OTROS GRUPOS O EMPRESAS: No aplica
12. PALABRAS CLAVE (al menos 5): Salud mental, Machine Learning, Colaboradores, Predicción, Bases de datos
13. FECHA DE INICIO: 24 de julio de 2023
14. DURACIÓN ESTIMADA (En meses): 12 meses
15. RESUMEN: La salud mental se ha convertido en una de las principales preocupaciones en la actualidad. La Universidad consciente de ello, busca el bienestar de sus colaboradores y se esfuerza por ofrecerles las mejores condiciones de trabajo. Sin embargo, se detecta un preocupante deterioro de la salud mental dentro de la comunidad educativa, situación que se ve agravada por los efectos de la pandemia, por esta razón se desarrollaron modelos de predicción de variables de salud mental en colaboradores pertenecientes a la Universidad privada ubicada en la ciudad de Cali, por medio de técnicas de Machine Learning. Esta investigación ha dado como resultado el desarrollo de modelos predictivos y la creación de documentación detallada sobre el proceso, se espera que los modelos desarrollados en este proyecto puedan ser implementados en diversos sectores, con el objetivo de facilitar la detección temprana de problemas de salud mental en los trabajadores y contribuir a su bienestar integral [1], [2].

PREDICCIÓN DE VARIABLES EN SALUD MENTAL PARA COLABORADORES DE UNA UNIVERSIDAD PRIVADA UBICADA EN LA CIUDAD DE CALI POR MEDIO DE APRENDIZAJE AUTOMÁTICO

María de los Ángeles Rosero Montilla
Código 8980512
Brayan Steven Ramírez Cortes
Código 8980266
Juan Manuel Vivas Torres
Código 8980511

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director(a)
Julián Gil González

Codirector(a)
Jimena Botero

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO 7 DE 2024

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

GLOSARIO

OMS: Organización Mundial de la Salud
Covid-19: Coronavirus SARS-CoV-2
CRISP-DM: Cross-Industry Standard Process for Data Mining (Proceso Estándar de la Industria para la Minería de Datos)
Machine Learning: Aprendizaje Automático
Dataset: Conjunto de datos
IA: Inteligencia Artificial
Df: Dataframe (Estructura de datos)

INTRODUCCIÓN

La salud mental, según la Organización Mundial de la Salud (OMS), hace referencia a un estado de bienestar mental que le permite a las personas afrontar momentos difíciles en la vida. También se considera como base o soporte para las capacidades y habilidades de las personas, siendo un derecho humano fundamental que debe ser promovido para favorecer el desarrollo de la sociedad [3]. Sin embargo, a pesar de su relevancia, los problemas de salud mental están en aumento a nivel global. Entre 2007 y 2017, se registró un preocupante incremento del 13% en estos trastornos, lo que ha tenido amplias repercusiones negativas en el bienestar de las poblaciones [1]. Se estima que aproximadamente 970 millones de personas en todo el mundo vivían con algún trastorno mental en 2019, siendo los trastornos de ansiedad y depresión los más frecuentes [4].

Es así como la salud mental se convierte en una preocupación para la salud pública, ya que afecta significativamente la calidad de vida de las personas y ha sido identificada como una de las principales causas de discapacidad en el mundo [5]. Como se muestra en los informes de la OMS citados por [6] es fundamental atender la salud mental de las personas, pues trastornos de ansiedad y depresión pueden estar asociados con ideación suicida e intentos de suicidio, ya que cualquier diagnóstico psiquiátrico implica un factor de riesgo.

Sumado a esto, la salud y la enfermedad mental son ejes fundamentales de la comprensión del desarrollo humano, no se debe reducir la comprensión de este fenómeno a dar por hecho que la no existencia de enfermedades mentales signifique la presencia de salud mental y que la existencia de enfermedades mentales no significa necesariamente la ausencia de salud mental [7]. Por el contrario, se debe procurar por entenderlo desde diferentes ejes, como los determinantes de la salud mental, que, según la OMS son aquellos factores individuales, sociales y estructurales que pueden proteger o no la salud mental de las personas [1]. Estos determinantes son los que focalizarán la atención de este proyecto, ya que serán la base de la predicción a realizar sobre la salud mental de los individuos encuestados.

La salud mental es un estado de bienestar psicológico y emocional, que con el paso de los años se ha visto afectado por diferentes factores externos como es el caso del Covid-19, el cual ha generado un gran impacto sobre las condiciones de vida y la salud mental de la sociedad. Centrándonos en los colaboradores de la comunidad universitaria, los efectos de la pandemia se han sumado a una salud mental deteriorada, que ya se había identificado dentro de la Universidad, en particular los colaboradores están expuestos a diferentes presiones, ambientes y tareas que demandan gran esfuerzo y generan distintos trastornos mentales que deben ser identificados y tratados a tiempo [1].

Por esta razón se evidencia la necesidad de encontrar tecnologías basadas en datos, que apoyen la toma de decisiones dentro de la universidad y con esto lograr un aprovechamiento de los datos en pro de la comunidad universitaria. Tecnologías como el aprendizaje automático que permite crear diferentes modelos predictivos a través de técnicas y algoritmos que brinden soluciones

específicas.

Por lo tanto, este proyecto implementa modelos de machine Learning para predecir variables de salud mental (Ver tabla 1) usando los datos almacenados por la universidad, los cuales fueron obtenidos a través de una encuesta perteneciente al proyecto salud y bienestar realizada a los colaboradores de la comunidad educativa universitaria.

1. DEFINICIÓN DEL PROBLEMA

La salud mental se ha convertido en una preocupación a nivel mundial, debido a que muchas personas padecen trastornos que perjudican sus condiciones de vida, afectan estados emocionales, comportamientos y respuestas corporales. Estos trastornos disminuyen la productividad de los trabajadores, repercuten económicamente en las industrias y causan graves afecciones psicofísicas [8].

La gestión de la salud en el trabajo es fundamental para mejorar las condiciones psicosociales de los colaboradores. Las empresas más competitivas son aquellas que cuentan con trabajadores mental y físicamente saludables gracias a políticas que apoyan y protegen su salud [2], [9]. Es por esto por lo que las investigaciones que buscan el mejoramiento en esta área son parte fundamental en el logro efectivo de mejores ambientes de trabajo. Además, se ha encontrado que el Machine Learning puede aportar de forma significativa en áreas de diagnóstico, tratamiento, apoyo, investigación y administración clínica [9]. La predicción de la salud mental es una de las partes más esenciales para reducir la probabilidad de padecer enfermedades mentales graves. Al mismo tiempo, la predicción de la salud mental puede proporcionar una base teórica para que el departamento de salud pública elabore planes de intervención psicológica para los trabajadores médicos [10].

Considerando lo anterior, es importante abordar problemas de investigación para mejorar los procesos de conocimiento en una organización, como la universidad del estudio, que busca generar planes para promover la salud mental en su comunidad educativa. Es en este sentido que se ha creado el proyecto Salud y Bienestar en la comunidad educativa liderado por los investigadores María Teresa Varela, Iván Leonardo Cepeda y Ana Marcela Uribe del Grupo de investigación Salud y Calidad de Vida, junto con Natalia Cadavid y Jimena Botero del Grupo de investigación Bienestar, Trabajo, Cultura y Sociedad, proyecto que tiene como objetivo caracterizar la salud mental de la comunidad educativa (estudiantes, profesores, colaboradores, administrativos y directivos) e identificar sus determinantes sociales.

Como parte de este proyecto, se ha propuesto desarrollar una estrategia para caracterizar la población estudiada es la implementación de modelos de Machine Learning en la predicción de variables en la salud mental, en este caso, de los colaboradores de la Universidad. A partir de los anterior, se pretende proporcionar entre otras cosas, fundamentos para entender cómo es la salud mental en los colaboradores y cómo la afectan los determinantes sociales.

1.1 PLANTEAMIENTO DEL PROBLEMA

La salud de acuerdo con la OMS, “Es un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades” [11], el deterioro de la salud mental al igual que otras enfermedades puede atentar contra la vida e integridad de una persona. Por esta razón se evidencia la necesidad de promover su protección en la sociedad.

Además, la salud mental es un tema que ha venido tomando fuerza a lo largo de los años, un ejemplo claro de ello es la pandemia de covid-19, donde muchas personas experimentaron diferentes trastornos de salud mental [12], por esta razón y entendiendo los riesgos que pueden generar los problemas de salud mental, muchas organizaciones buscan implementar estrategias para solventarlos. Teniendo en cuenta lo anterior, la Universidad tiene como objetivo promover la salud mental en diferentes grupos, para ello cuenta con una base de datos, resultado de encuestas realizadas a colaboradores de la universidad, sin embargo, los datos por si solos no le dan información necesaria para predecir variables de salud mental, que le permitan tomar decisiones en pro del bienestar y salud de sus colaboradores.

Es por esta razón, que se hace importante el entender los factores que aumentan o reducen el riesgo de sufrir trastornos mentales, de tal manera que se puedan generar esfuerzos para comprender la realidad de los individuos y así identificar posibles problemas a tiempo, la salud mental no solo se trata del individuo, sino también de su entorno, sus condiciones de vida y lo que realice en el día a día [13], [9].

Por lo anterior se evidencia la necesidad de implementar una herramienta, basada en datos, que sirva como medio de apoyo para la toma de decisiones, de todo lo relacionado con las variables de salud mental en colaboradores de la universidad y que permita dar valor a los datos almacenados.

1.2 FORMULACIÓN DEL PROBLEMA

Teniendo en cuenta el desarrollo anterior se han planteado las siguientes preguntas para el proyecto aplicado: ¿Cómo predecir variables de salud mental en colaboradores utilizando bases de datos derivada del proyecto salud y bienestar en la comunidad educativa de la universidad y técnicas de machine Learning?, ¿Qué variables tener en cuenta a la hora de hacer la predicción?, ¿Cuáles técnicas de machine Learning se deben seleccionar, para darle solución al problema planteado?, ¿Cómo crear los modelos a través del uso de técnicas de Machine Learning?, ¿Cómo evaluar el nivel de desempeño de los modelos que se desarrollen para darle solución al problema? Estas preguntas responden con pertinencia a la necesidad identificada en el planteamiento del problema y su solución permitirá aportar al avance en el proyecto de salud y bienestar en la comunidad universitaria.

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar modelos de predicción de variables en salud mental para colaboradores de una Universidad privada ubicada en la ciudad de Cali, a partir de encuestas de bienestar universitario y técnicas de machine learning.

2.2 OBJETIVOS ESPECÍFICOS

- Desarrollar un módulo procesamiento y análisis de datos con el Dataset obtenido.
- Desarrollar modelos de predicción a partir de técnicas de aprendizaje automático.
- Validar los resultados estadísticos de los modelos.

3. MARCO TEÓRICO Y ANTECEDENTES

En este capítulo se presenta información referente a la salud mental y el enfoque que se le da a su análisis. De igual manera se conceptualiza el aprendizaje automático, focalizando sus principios básicos, métodos, tipos y fases. De esta forma se puede lograr una visión holística del campo que se usará en el desarrollo del proyecto. Así mismo se muestra un análisis de antecedentes en el que se recopilaron algunos documentos con proyectos e investigaciones encaminadas a resolver situaciones de salud mental, usando aprendizaje automático, se concluirá sobre lo hallado de tal manera que se cuente con un precedente para fortalecer la investigación.

3.1 MARCO TEÓRICO

En esta sección se presentan los conceptos fundamentales que soportan el presente proyecto aplicado en el campo de la ciencia de datos y la salud mental. Se explorarán conceptos relacionados con la salud mental y su abordaje en el entorno laboral, así como las variables que intervienen. Además, se estudiarán las técnicas de aprendizaje automático, sus tipos y el proceso estándar CRISP-DM.

3.1.1 SALUD MENTAL

La salud mental es un componente fundamental en la vida de las personas, se define como un estado de bienestar que permite hacer frente a las tensiones normales de la vida, es parte importante de la salud y el bienestar que sustenta nuestras capacidades individuales y colectivas para tomar decisiones, establecer relaciones y dar forma al mundo en el que vivimos [3].

Existen factores que afectan la manera en cómo se desarrolla la salud mental, algunos de riesgo y otros protectores, que definen la resiliencia con la que los individuos se enfrentarán a las dificultades que se encuentren en la vida [3]. La OMS hace hincapié en que dichos factores se pueden afectar de forma local si son amenazas para personas, familias o comunidades y mundial para poblaciones enteras como en el caso de las guerras o brotes de enfermedades, además, indican del carácter subjetivo y relativo de dichos factores al plantear que la mayoría de las personas no desarrollan afecciones de salud mental, aunque estén expuestas a factores de riesgo, y no todas las personas con afecciones mentales fueron expuestas a uno [3]. Sin embargo, conocer los factores protectores y de riesgo de la población facilita identificar el foco de aplicación de recursos para lograr intervenciones eficaces [14].

3.1.1.1 VARIABLES DE SALUD MENTAL

La salud mental se puede manifestar por medio de indicadores tanto positivos como negativos, que pueden dar muestra de la condición que presenta el individuo, estos indicadores se manifiestan en el siguiente proyecto como variables respuesta y serán aquellas condiciones que se pretenden predecir a partir de los determinantes sociales.

Las variables de salud mental son vitales para comprender y evaluar el bienestar psicológico en diferentes contextos como el laboral. Estas variables pueden ser negativas como el nivel de estrés, ansiedad y depresión o positivas como la resiliencia y la satisfacción con la vida. A continuación, se muestran las variables establecidas para el análisis en el presente proyecto.

Tabla 1.
Tabla Variables de resultado – Fuente [1].

Variables de resultado	**Indicadores**
Salud mental	Aspectos negativos: • Depresión • Ansiedad • Estrés • Soledad • Ideación suicida Aspectos positivos: • Resiliencia • Recursos psicológicos • Satisfacción con la vida

Para abordar dichas variables presentaremos definiciones y abordajes para cada una.

Aspectos Negativos

Depresión: Es un trastorno del estado de ánimo caracterizado por sentimientos persistentes de tristeza, desesperanza y falta de interés en las actividades diarias. Puede interferir significativamente con el funcionamiento personal y laboral, afectando la calidad de vida y el rendimiento en el trabajo [15].

Ansiedad: Se refiere a un estado de preocupación excesiva y miedo frente a situaciones futuras percibidas como amenazantes. En el entorno laboral, puede manifestarse como nerviosismo, dificultad para concentrarse y problemas de sueño, lo que puede afectar el desempeño laboral y la satisfacción en el trabajo [15].

Estrés: Es una respuesta fisiológica y emocional a las demandas del trabajo que superan la capacidad de afrontamiento de un individuo. El estrés crónico en el trabajo puede contribuir al desarrollo de problemas de salud mental, como la depresión y la ansiedad, así como a problemas físicos como enfermedades cardiovasculares y trastornos musculoesqueléticos [15].

Soledad: La soledad en el trabajo se refiere a la falta de conexión social y apoyo interpersonal en el entorno laboral. Puede surgir cuando los empleados se sienten aislados, excluidos o marginados en el lugar de trabajo, lo que puede aumentar el riesgo de depresión, ansiedad y estrés laboral [15].

Ideación Suicida: La ideación suicida es la experiencia de pensamientos, planes o deseos de causarse daño o terminar con la propia vida. En el contexto laboral, la ideación suicida puede ser un síntoma de angustia emocional grave y puede requerir intervención y apoyo profesional inmediato [15].

Aspectos Positivos:

Resiliencia: La resiliencia en el trabajo se refiere a la capacidad de los individuos para adaptarse y recuperarse frente a desafíos y adversidades laborales. Los empleados resilientes pueden mantener un rendimiento efectivo incluso bajo presión, superar obstáculos y aprender de experiencias difíciles, lo que contribuye al bienestar emocional y al éxito laboral [16].

Recursos Psicológicos: Son habilidades, atributos personales y estrategias de afrontamiento que ayudan a los individuos a manejar el estrés y promover el bienestar en el trabajo. Estos recursos pueden incluir la autoeficacia, el optimismo, la autoestima, la capacidad de regulación emocional y el apoyo social, y pueden proteger contra los efectos negativos del estrés laboral [16].

Satisfacción con la Vida: Se refiere a la evaluación subjetiva de los empleados sobre su bienestar y felicidad en el ámbito laboral. Los empleados que reportan altos niveles de satisfacción con la vida en el trabajo tienden a estar más comprometidos, productivos y satisfechos con su carrera profesional [16].

3.1.1.2 DETERMINANTES SOCIALES

La salud mental puede verse influenciada por diferentes factores individuales, sociales y estructurales que pueden proteger o afectar el bienestar psicológico de las personas, estos factores son llamados determinantes de salud mental y pueden estar presentes en diferentes etapas de la vida. Los determinantes individuales comprenden factores genéticos, habilidades emocionales, estado físico y otros aspectos personales. Por otro lado, los determinantes sociales están relacionados con el entorno familiar, el nivel socioeconómico y el empleo, entre otros. Por último, los determinantes estructurales incluyen aspectos políticos y sistemas sociales que también influyen en la salud mental [3], [13].

Un determinante social se refiere a las condiciones en las que una persona nace, crece, vive, trabaja y envejece, y cómo estas condiciones impactan su vida, salud y bienestar. Estos determinantes sociales incluyen aspectos como el nivel socioeconómico, el nivel educativo, el empleo u ocupación, su contexto físico y social, entre otros [13].

Estos determinantes adquieren importancia en el estudio de la salud mental, en tanto se ha investigado esta relación desde diferentes estudios [17], [18], [19]. Encontrando que una mirada a estos determinantes y factores psicológicos pueden mejorar los procesos de promoción en salud.

En la encuesta asociada a este proyecto se recolectó información frente a determinantes sociales en trabajadores de la universidad, las cuales se denominarán para el proceso de análisis como variables de exposición y se encuentran explícitas en la tabla 2.

Tabla 2.
Tabla Variables de exposición – Fuente [1].

Variables de exposición		**Indicadores**
Determinantes intermedios	Factores individuales	Sociodemográficos: sexo, edad, nivel educativo, estrato socioeconómico, procedencia, residencia rural/urbana, composición familiar. Psicosociales: Antecedentes de violencia y de abuso sexual, apoyo social, funcionamiento familiar, afrontamiento
	Aspectos del contexto laboral universitario	Cargo, tiempo de vinculación, trabajar y estudiar, demandas cuantitativas, carga mental, jornada de trabajo, relaciones sociales en el trabajo, influencia del trabajo sobre el entorno extralaboral, demandas emocionales, tecnoestrés, satisfacción laboral, conocimiento y acceso a programas de salud y bienestar en la universidad, ambiente alimentario universitario, vacaciones. <u>Solo para profesores:</u> Tipo de contrato (planta y cátedra), número de asignaturas a cargo (pre y posgrado), número de estudiantes, dedicación del tiempo docente, funciones adicionales <u>Solo para directivos:</u> Funciones adicionales, aspectos del rol directivo (cooperación, autonomía, mediación, demandas, toma de decisiones), aspectos del rol docente (solo para directivos académicos)
	Condiciones de vida	Condiciones de la vivienda y del barrio, convivencia (con quien vive), situación económica del grupo familiar y dependencia del núcleo familiar, seguridad alimentaria, transporte hogar universidad, seguridad social en salud
Determinantes estructurales		Género, etnia, posición socioeconómica (nivel educativo de los padres, nivel de ingresos del hogar) Conocimiento y acceso a programas de salud y bienestar en la universidad

3.1.1.3 SALUD MENTAL EN EL ENTORNO LABORAL

En el entorno laboral, la salud mental puede influir de forma considerable en nuestro bienestar

psicológico, ya que gran parte de nuestra vida adulta transcurre en un lugar de trabajo. Las condiciones laborales, las relaciones interpersonales, el nivel de estrés y la satisfacción en el trabajo son factores clave que pueden afectar nuestra salud mental [20].

Es así como las condiciones laborales pueden afectar tanto positiva como negativamente la salud de los trabajadores, cuando el trabajo es gratificante y contribuye a la autorrealización personal tiene impactos positivos en la salud mental. Pero si el trabajo genera situaciones de estrés, horarios inadecuados y experiencias de abuso y acoso puede tener efectos negativos [21]. Es vital que las organizaciones implementen estrategias de promoción de salud mental y prevención de enfermedades en los lugares de trabajo [21], pues como se muestra en [cita] las personas menos felices en su trabajo son menos productivas, lo que revela un riesgo para las empresas que no atiendan la salud mental de sus colaboradores [22], [23].

Teniendo en cuenta lo presentado anteriormente, se aplicarán métodos de Machine Learning que permitan predecir la clasificar individuos en las variables resultado, a partir de las condiciones que presenten los individuos en sus variables de exposición. En ese sentido se presenta a continuación un abordaje conceptual del Machine Learning, sus métodos y características fundamentales, dando base al proceso posterior.

3.1.2 MACHINE LEARNING

El aprendizaje automático o Machine Learning es un campo de la inteligencia artificial que se centra en el estudio de sistemas capaces de aprender a partir del análisis de datos. El término "aprendizaje automático" se refiere a la capacidad de las máquinas para aprender a través de diversas técnicas y algoritmos para generalizar y automatizar comportamientos basados en datos de entrada mediante la detección de sus patrones [24].

En lugar de programar, paso a paso, soluciones específicas para cada necesidad, como se hace desde la programación clásica, el aprendizaje automático se dedica al desarrollo de algoritmos que generalicen la extracción de patrones de una fuente de datos [25].

Se han desarrollado procesos que permiten al aprendizaje automático adquirir la capacidad de pensar de manera similar a un ser humano y, por lo tanto, llevar a cabo tareas específicas sin necesidad de una programación explícita para cada función asignada. En esta búsqueda, se ha observado que las máquinas pueden aprender a potenciar su inteligencia artificial usando modelos de aprendizaje automático, lo que les permite ser autónomas al desarrollar nuevas tareas y funciones [26].

Lo anterior se ve reflejado en el hecho de que, en la inteligencia artificial, el aprendizaje automático se ha convertido en el método preferido para el desarrollo práctico de software en visión por computadora, reconocimiento de voz, procesamiento de lenguaje natural, control robótico y otras aplicaciones [27].

3.1.2.1 EL ABORDAJE DE UN PROBLEMA CON INTELIGENCIA ARTIFICIAL

Como ya se ha visto, el Aprendizaje profundo es un campo de la inteligencia artificial que ha cobrado bastante importancia en el desarrollo actual de diferentes sectores, que lo usan como medio para obtener información fidedigna y así poder tomar mejores decisiones.

Se menciona que hay 6 conceptos clave para poder abordar problemas que se puedan resolver haciendo uso de Inteligencia artificial [26]:

Tabla 3.
Tabla Conceptos de abordaje Inteligencia artificial – Fuente [26].

Concepto 1	La IA necesita problemas bien definidos, con límites específicos.
Concepto 2	Inteligencia significa la presencia de al menos una de las ocho características. Estos incluyen el razonamiento, la percepción, el lenguaje natural, el movimiento, el aprendizaje, la representación del conocimiento, la planificación y la conciencia social
Concepto 3	Las actividades inteligentes generalmente se componen de varias más pequeñas, solo algunas de las cuales pueden ser inteligentes.
Concepto 4	Los datos son el combustible de la inteligencia artificial y el aprendizaje automático.
Concepto 5	Las actividades inteligentes se pueden representar usando el lenguaje de las matemáticas.
Concepto 6	La inteligencia artificial repite pequeñas tareas muchas veces con diferentes datos para encontrar el resultado correcto, lo que generalmente alimenta una actividad mayor.

Con estos conceptos se pretende dar base a la comprensión de los problemas que se pueden resolver con IA y como la convergencia de ellos permite mejorar el abordaje de estos.

3.1.2.2 PASOS PARA RESOLVER UN PROBLEMA DE APRENDIZAJE AUTOMÁTICO

En este apartado se va a mostrar una generalidad propuesta por [28] para resolver un problema por medio del aprendizaje automático.

Definición del problema: Determinar cuál es exactamente el problema antes comenzar a resolverlo. Se necesita descubrir cuál es el problema, si es factible usar el aprendizaje automático para resolverlo, y plantear una hipótesis de acuerdo con la comprensión del contexto.

Recopilación de datos: Se recopilan los datos en función de la definición del problema. Se debe garantizar una recopilación sistemática de los datos, asegurando contar con los factores necesarios para el análisis.

Preprocesamiento de los datos: Limpiar los datos para mejorar su uso. Esto incluye la eliminación de valores atípicos, el manejo de la información faltante, etc. Esto se hace para disminuir la posibilidad de obtener errores de análisis.

Desarrollo del modelo: Crear el modelo de aprendizaje automático que se utilizará para resolver el problema. Este modelo toma los datos como entrada, realiza cálculos sobre ellos y luego produce algún resultado a partir de ellos.

Evaluación del modelo: El modelo debe evaluarse para verificar su precisión y asegurarse de que pueda funcionar con cualquier dato nuevo que se le proporcione.

Estos pasos pueden variar en estructura, orden o abordaje y dar pie a la creación de diferentes metodologías para proyectos con aprendizaje autónomo. En un apartado posterior se abordará la metodología elegida para el desarrollo del proyecto y se podrá ver la similitud que tiene con este modelo de trabajo.

3.1.2.3 TIPOS DE APRENDIZAJE AUTOMÁTICO

Para desarrollar proyectos en ciencia de datos, es esencial conocer algunos conceptos clave de aprendizaje automático, así como las métricas más utilizadas. Entre los conceptos básicos se encuentra la comprensión de los distintos tipos de aprendizaje automático [24]. A continuación, se presenta un mapa conceptual que ilustra estos tipos. Posteriormente, se hablará de cada uno de ellos, con un enfoque especial en el aprendizaje supervisado.

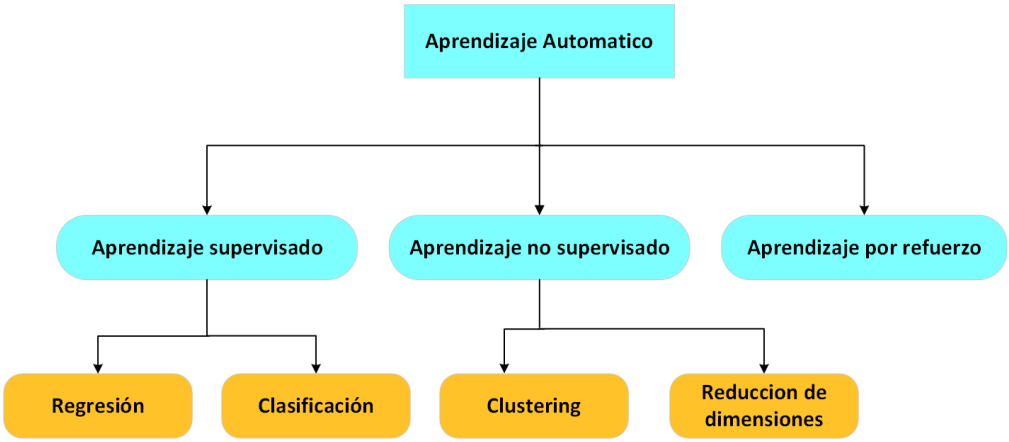


Ilustración 1. Tipos de aprendizaje automático. Fuente[25]

El objetivo principal del aprendizaje automático es analizar y construir algoritmos que puedan en base a datos históricos, aprender y realizar predicciones sobre nuevos datos de entrada y que sean capaces de definir que tan bien el modelo realizado está aprendiendo [24]. En ese sentido los autores definen lo siguiente:

Aprendizaje no supervisado

Consiste en que un algoritmo es entrenado en un conjunto de datos sin etiquetar, sin información previa sobre resultados esperados. El objetivo principal del aprendizaje no supervisado es descubrir patrones, estructuras o relaciones esenciales en los datos, como agrupamientos naturales, tendencias o correlaciones [29].

Aprendizaje supervisado

Consiste en entrenar un modelo utilizando un conjunto de datos que incluye ejemplos de entrada y la salida esperada correspondiente. En otras palabras, el modelo aprende a partir de ejemplos etiquetados, donde se conoce la respuesta correcta para cada instancia de datos [29]. Hay dos tipos de aprendizaje supervisado:

Clasificación

La clasificación se basa en variables de salida categóricas. Los algoritmos de clasificación aprenden a asignar una etiqueta o categoría específica a un conjunto de variables de entrada. Los objetos clasificados pueden ser registros en una base de datos y tener formatos de texto, imágenes o señales de comunicación. El proceso de clasificación consiste en dividir los datos del conjunto de datos (dataset) en dos conjuntos: uno de entrenamiento, conformado por elementos previamente clasificados, y un segundo conjunto cuya clase es desconocida. Así se busca que con el primer conjunto se pueda crear un modelo que permita clasificar el segundo conjunto [29].

Regresión

La regresión, en el ámbito del aprendizaje supervisado, se define como la predicción de un valor continúo utilizando un conjunto de variables de entrada. Por ejemplo, se puede emplear un algoritmo de regresión para estimar el precio de un automóvil basándose en características como su tamaño, marca y antigüedad [29].

Aprendizaje semi-supervisado:

Un enfoque de aprendizaje que utiliza una combinación de datos etiquetados y no etiquetados para el entrenamiento del modelo [25].

Aprendizaje por refuerzo

Consiste en que un agente aprende a través de la interacción con un entorno, recibiendo retroalimentación en forma de recompensas o penalizaciones en función de las acciones que realiza [25]

3.1.2.4 ALGORITMOS O TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Para abordar un problema que requiera el uso de Machine Learning, se han desarrollado diferentes algoritmos que permiten encontrar patrones, comportamientos y garantizar validez, usando modelos matemáticos como base. Estos modelos responden a diferentes campos de las matemáticas y pueden ser usados de acuerdo con las necesidades particulares de cada proyecto.

Se incluye una tabla de algoritmos o técnicas de aprendizaje automático populares en la industria [30]:

Tabla 4.
Técnicas de aprendizaje automático – Fuente [30].

TIPOS DE APRENDIZAJE	ALGORITMOS DE APRENDIZAJE
Aprendizaje supervisado	Regresión, Regresión logística, Árboles de decisión (Decision Trees), Bosques aleatorios (Random Forest), Gradient boosting, Redes neuronales artificiales (ANN: Artificial Neural Networks), Máquinas de vectores de soporte (SVM: Support Vector Machines), Redes neuronales convolucionales (CNN: Convolution Neural Networks), Redes neuronales recurrentes (RNN: Recurrent Neural Networks), Redes bayesianas (Bayesian network).
Aprendizaje no supervisado	Redes antagonistas generativas (GAN: Generative Adversarial Networks)
Aprendizaje por Refuerzo	Enfoques evolutivos Redes lógicas de Markov (Markov Logic Networks)
Modelos Probabilísticos y Algoritmos Basados en Modelos Gráficos	Modelos ocultos de Markov (HMM: Hidden Markov Model)

3.1.2.4.1 DECISION TREES

Los *Decision trees* o árboles de decisión son algoritmos de Machine Learning que permiten desarrollar tareas de clasificación y regresión para bases de datos complejas [31]. Pertenecen a los modelos de aprendizaje supervisado y su objetivo fundamental es el aprendizaje inductivo a partir de observaciones y construcciones lógicas, permitiendo así clasificar un dato según sus atributos o mediciones [32], [33].

El funcionamiento del algoritmo se da al dividir repetidamente el conjunto de datos en

subconjuntos usando sus atributos o medidas, de manera que se maximice la homogeneidad de las muestras en cada subconjunto y se maximice la heterogeneidad entre los distintos subconjuntos creados [34].

El Decision Tree se construye tomando como raíz las características de los datos y conforme se van creando particiones en subconjuntos se van formando nodos de decisión, en cada nodo se evalúan las características de los datos generando nuevas particiones [33].

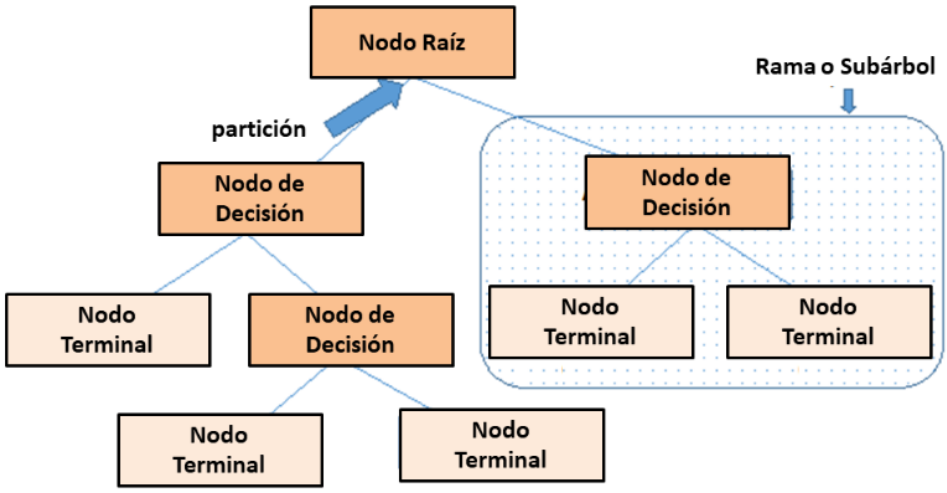


Ilustración 2. Estructura de un árbol. Fuente [35]

Para determinar aquella característica que permita una mejor clasificación el algoritmo verifica dos medidas, la entropía o la medida de impureza Gini. La entropía mide el *desorden, si* su valor es cercano a cero se dice que el conjunto es homogéneo, si su valor es uno se dice que las muestras están divididas de forma equitativa [34].

Para calcular la entropía H se usa la fórmula:

$$H_i = -\sum_{k=1}^{n} P_{i,k} log_2(P_{i,k}), P_{i,k} \neq 0$$

Donde P_i es la proporción de instancias de la clase k entre las instancias de entrenamiento en el i-ésimo nodo [31], n indica que puede haber más de dos características y se dice que $P_{i,k} \neq 0$ debido a que si es cero significa que ninguna muestra posee la característica indicada [34].

La medida de impureza Gini se calcula de la siguiente manera:

$$G_i = 1 - \sum_{k=1}^{n} (P_{i,k})^2$$

Ambas medidas ofrecen información importante sobre la impureza de los nodos, y como se expresa en [31]: "La impureza de Gini es ligeramente más rápida de calcular, por lo que es una buena opción por defecto. Sin embargo, cuando difieren, la impureza de Gini tiende a aislar la clase más frecuente en su propia rama del árbol, mientras que la entropía tiende a producir árboles ligeramente más equilibrados".

Posterior a la construcción del árbol con mejores métricas, el algoritmo provee información para clasificar una nueva instancia, siguiendo el camino desde la raíz hasta llegar a un nodo terminal, el cuál es tomado como la clase a la que pertenece [33].

Dentro de los algoritmos de clasificación existe la posibilidad de que se dé un fenómeno llamado sobreajuste, el cual indica que el modelo clasifica bien con los datos de entrenamiento, pero no generaliza correctamente cuando se presentan datos de prueba distintos a los que entrenó [31].

El desarrollo del modelo en python se genera a partir de la librería sklearn.tree de scikitLearn con el modelo DecisionTreeClassifier.

3.1.2.4.2 RANDOM FOREST

Random Forest o Bosques Aleatorios, es un algoritmo de Machine Learning que agrupa varios Decision trees con subconjuntos aleatorios, para hacer predicciones y luego obtener el mejor predictor entre todos ellos [31].

El algoritmo Random Forest comienza seleccionando una muestra aleatoria de tamaño n en el conjunto de datos, a partir de una muestra inicial el árbol crece y se selecciona para cada nodo de características. Luego se divide el nodo con la función que maximice el objetivo de Information Gain o ganancia de información [36]:

$$IG = (D_P, f) = l(D_P) - \frac{N_l}{N_p} l(D_l) - \frac{N_r}{N_p} l(D_r)$$

En donde IG es la función Information Gain, f es la característica para realizar la división, N_p, N_l y N_r corresponden al número de muestras en el nodo padre, nodo hijo izquierdo y nodo hijo derecho respectivamente, l es la medida de impureza (gini, entropía o error de clasificación) y D_P, D_l y D_r son el conjunto de datos en el nodo padre, en el nodo hijo izquierdo y nodo hijo derecho, respectivamente.

El algoritmo de Random Forest conlleva un proceso aleatorio adicional al crecer los árboles, en lugar de buscar la mejor característica cuando divide un nodo como se hace en Decision Tree, busca la mejor característica entre un subconjunto aleatorio de características, lo que resulta en una mayor cantidad de árboles (resultando así su nombre) y generando un modelo globalmente mejor que Decision Tree [31].

Para ejecutar el modelo en python se usó el modelo RandomForestClassifier del módulo sklearn.ensemble de Scikit-Learn.

Parámetros

Para definir el modelo se cuentan con diferentes parámetros que ajustan el clasificador en diferentes opciones, algunos de ellos son [31]:

n_estimators: El número de árboles en el bosque.
Criterion(gini o entropy): Es la función para medir la calidad de una división. “gini” para la impureza de Gini y “entropy” para la ganancia de información
max_depth: Es la profundidad máxima del árbol. Si su valor es “None”, los nodos se extienden hasta que todas las hojas sean puras o hasta que todas las hojas contengan el mínimo de muestras definido.
min_samples_split: Número mínimo de muestras necesarias para dividir un nodo interno:
min_samples_leaf: Número mínimo de muestras requerido para estar en un nodo hoja.
max_features(“auto”, “sqrt”, “log2”): Número de características a considerar a la hora de buscar la mejor división, se puede establecer por medio de la raíz cuadrada da la cantidad de muestras con “auto” y “sqrt” o con el logaritmo en base 2 del número de muestras.

3.1.2.4.3 XGBOOST

XGBoost es un algoritmo de machine learning que fue desarrollado en el marco de las gradient boosting y usa un método de ensamble sobre Decision Trees [37]. Su diseño se enfoca en la eficiencia, la versatilidad y portabilidad [38]. Para entender su funcionamiento se abordarán los tres últimos estadios de evolución de los algoritmos basados en Decision Trees planteados por [39]: Boosting, Gradint Boosting y XGBoost.

Boosting

El boosting es un método de aprendizaje en el que se divide el conjunto de datos en N partes, donde los Decision Trees aprenden de los errores cometidos por los árboles anteriores, obteniendo como resultado un último árbol optimizado, El modelo Boosting pertenece al tipo de modelos llamados modelos de ensamblaje, debido a que se construyen conjuntos con árboles de decisión individuales entrenados de forma secuencial [40].

Gradient Boosting

El algoritmo Gradient Boosting busca reducir el valor de la función de pérdida en el conjunto de datos de entrenamiento. Esto se logra haciendo la suma ponderada de los Decision Trees definidos para el modelo [40].

Extreme Gradient Boosting (XGBoost)

XGBoost surge como una mejora al modelo Gradient Boosting, incluyendo penalizaciones para evitar sobreajuste, reducción proporcional de las hojas de los árboles, inclusión del método de Newton-Raphson para la función de pérdida y entrenamiento eficiente usando multiprocesadores [37], [40].

Como se plantea en [39] el modelo funciona de la siguiente manera:

Primero se obtiene un árbol inicial F_0 para predecir la variable y, el resultado se asocia con un residuo $y - F_0$, luego se define un árbol h_1 que ajusta el residuo anterior. Se combinan los árboles de la siguiente manera:

$$F_1(x) <- F_0(x) + h_1(x)$$

De tal manera que el error cuadrático medio de F_1 sea menor que el de F_0. Este proceso se itera de la siguiente manera:

$$F_m <- F_{(m-1)}(x) + h_m(x)$$

Donde F_m es el árbol final que minimiza el error.

Para ejecutar el proceso en python se utilizó el modelo proveniente de la librería denominada xgboost.

Parámetros
Los parámetros definidos para este modelo al ser basados en árboles tienen similitudes con los de otros modelos como el Random forest, algunos de ellos son [31], [39]:

booster: Determina el tipo de modelo a usar. En este caso son árboles y se definen con "gbtree"
n_estimators: Número de árboles de decisión que se van a entrenar.
max_depth: Profundidad máxima del árbol.
subsample: Proporción de muestras utilizadas para cada árbol. Ayuda a prevenir el sobreajuste.
learning_rate (también "eta"): Controla cuánto contribuye cada árbol al modelo final.
colsample_bytree: Proporción de características seleccionadas para cada árbol.

3.1.2.5 LENGUAJES DE PROGRAMACIÓN EN APRENDIZAJE AUTOMÁTICO

Debido a la amplia gama de avances tecnológicos que surgieron en años anteriores y con la posibilidad de crear entornos de programación diversos, el aprendizaje automático se puede desarrollar en diferentes lenguajes, [30] condensa una lista con de lenguajes de programación ampliamente utilizados por científicos de datos de todo el mundo, entre los cuales se encuentran: Python, R, SQL, Java, JavaScript, MATLAB

Según la demanda del proyecto, es posible que un experto en aprendizaje automático necesite usar una variedad de herramientas y lenguajes de programación.

3.1.3 METODOLOGÍA CRISP-DM

El modelo CRISP-DM cubre todas las fases necesarias para que un proyecto se lleve a cabo, este se encuentra conformado por seis fases que son flexibles, puesto que no hay una estructura definida y se adapta fácilmente a las distintas investigaciones, es así como dicha metodología se posiciona como una de las más usadas al momento de estructurar proyectos en ciencia de datos. A continuación, se puede observar la metodología de manera gráfica.

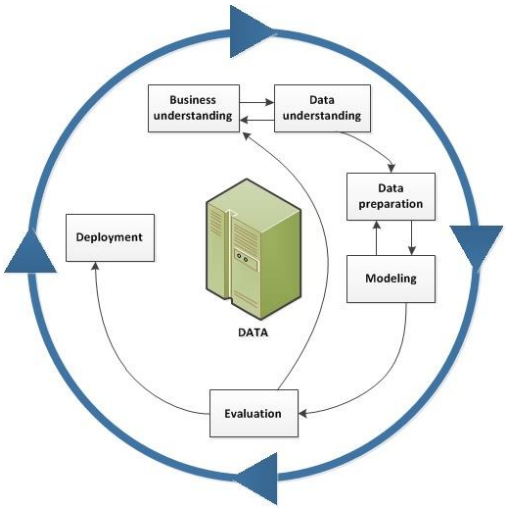


Ilustración 3. Metodología CRISP-DM [41]

3.2 ANTECEDENTES

En este apartado se presentan algunos proyectos que han usado técnicas de machine Learning para abordar problemáticas de salud mental, también se exponen estudios previos que muestran las tendencias en uso de modelos de aprendizaje automático, de tal manera que se cuente con una base para la elección de los modelos a desarrollar en el presente proyecto.

Técnicas de aprendizaje automático para la predicción de la salud mental [42]: Este articulo tiene como objetivo identificar la posibilidad futura de salud mental para reducir el número de casos de suicidio, para este estudio se tuvieron en cuenta 76 atributos de cada una de las personas encuestadas, por otro lado las técnicas usadas para la detección fueron las siguientes: (Árbol de decisión (DT), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression (LR), XGBoost (XGB), Gradient Boost Classifier (GBC) y Artificial Neural Network (ANN). El resultado obtenido fue de 87,38 % con el modelo Support Vector Machine (SVM).

El anterior estudio nos puede aportar mayor conocimiento acerca de las técnicas de Machine Learning usadas para la predicción, de igual forma nos permite conocer los parámetros que existen para evaluar la precisión de un modelo.

Se desarrolló una revisión sistemática en bases de datos sobre las principales metodologías de aprendizaje automático que para hacer predicción de la depresión. El estudio encontró que es fundamental la integración de diferentes modelos de tal manera que se puedan elegir los que mejor generen las predicciones [43].

En cuanto a la frecuencia con la que aparece un modelo u otro, los autores encontraron lo siguiente:

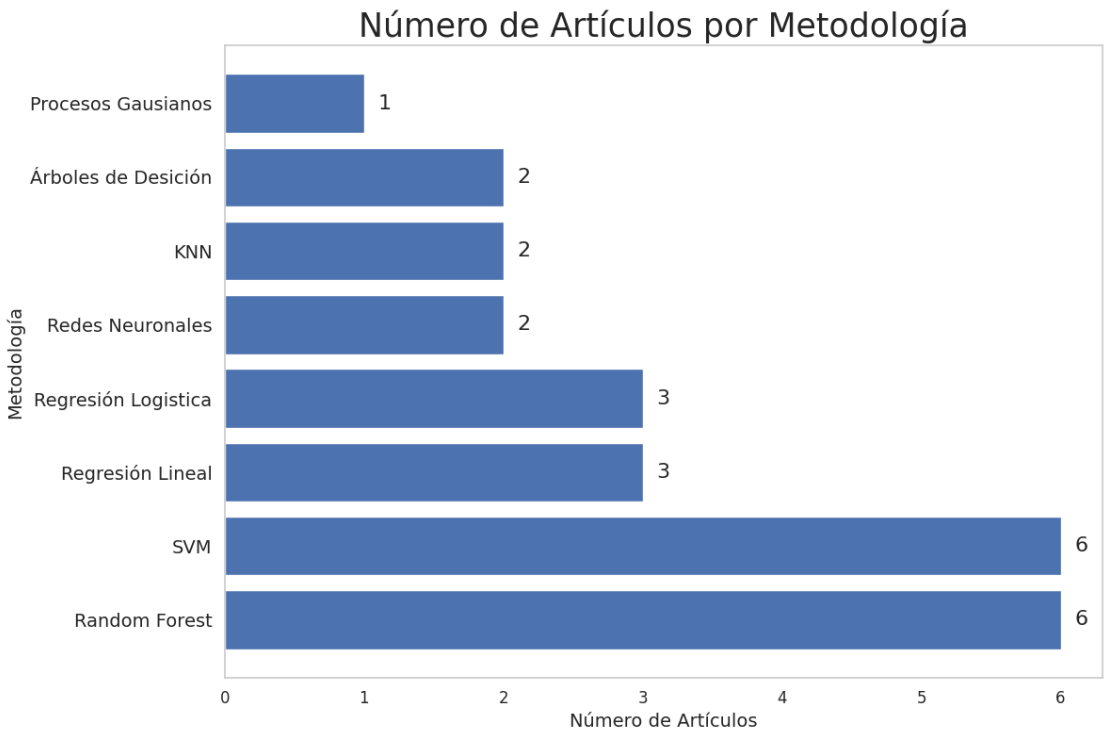


Ilustración 4. Frecuencia de uso de modelos de predicción [43].

Y respecto a la precisión que lograron los modelos dentro de los estudios revisados, los autores encontraron que el SVM y las redes neuronales convolucionales obtuvieron los mejores resultados.

Como se puede observar, los modelos predilectos en los estudios analizados son RF, SVM y las redes neuronales convolucionales, por lo tanto, deben ser objeto de análisis y revisión para el presente proyecto.

Este análisis se ve reflejado en que hay proyectos como el desarrollado por [44] en el cual los autores muestran que hay aspectos de la vida personal y laboral que afectan la salud mental de los trabajadores de un sector en particular.

Por medio de la metodología CRISP DM los autores aplicaron el modelo RF para desarrollar su proyecto, detectando factores de riesgo para el padecimiento de enfermedades mentales en los trabajadores.

Otra investigación revisada fue la desarrollada en [45], la cual quiere mostrar cómo se puede detectar de forma temprana los problemas de salud mental mediante algoritmos de Machine Learning, aquí se menciona que la ansiedad y la depresión son problemas de salud pública que afecta a cualquier individuo y esto puede verse reflejado en diversos síntomas somáticos, en este trabajo se abordan diferentes modelos como logistic regression, KNN, Decision Trees, Random Forest, apilamiento, el cual se llegó a la conclusión que este último tiene la mejor predicción con un 81.75%, todo esto fue gracias a un conjunto de datos que estaba compuesto por 27 columnas y 1259 entradas.

Tabla 5.
Resultados de precisión. Fuente [45].

Methods	**Accuracy (%)**
Logistic Regression	79.63
KNeighbors Classifier	80.42
Decision Tree classifier	80.69
Random Forests	81.22
Stacking	81.75

Este artículo [46] se centra en la detección de ansiedad, depresión y apatía en personas mayores con demencia preclínica y deterioro cognitivo leve, este estudio tiene como objetivo desarrollar un modelo capaz de identificar estos síntomas basándose en el análisis del habla y las expresiones faciales. La muestra incluyó a 319 adultos mayores diagnosticados con deterioro cognitivo leve. Para la creación del modelo de clasificación de emociones multicategoría, se emplearon bosques aleatorios, los cuales obtuvieron una puntuación F1 promedio ponderada del 96,6 %. Además, el modelo demostró una alta precisión, exactitud y recuperación, con valores del 87,4 %, 86,6 % y 87,6 % respectivamente.

En resumen, estos estudios han demostrado el potencial del aprendizaje automático en el campo de la salud mental, generando predicciones precisas, lo que resulta de gran utilidad para la presente investigación, ya que confirman la viabilidad de utilizar técnicas de aprendizaje automático para predecir variables de salud mental.

Por otro lado, lo que hace diferente a este proyecto de las investigaciones vistas en este capítulo es su enfoque en predecir tanto variables positivas como negativas de la salud mental. A diferencia de otras investigaciones que se centran en la identificación de problemas de salud mental específicos como la depresión o la ansiedad, este proyecto aborda un espectro más amplio de variables de salud mental. Esto incluye no solo la detección de problemas, sino también la identificación de factores positivos que contribuyen al bienestar mental. Además, busca desarrollar modelos que no solo predigan variables de salud mental con alta precisión, sino que también ofrezcan interpretaciones claras y útiles.

4. METODOLOGÍA

En este capítulo se expone la metodología utilizada en el desarrollo de la investigación, detallando cada una de las fases del proyecto, como es el caso de la limpieza y procesamiento de los datos, así como la creación de los modelos. La elección la metodología CRISP-DM se fundamenta en su capacidad para garantizar que cada etapa del proceso contribuya de manera efectiva al logro de los objetivos planteados para el proyecto, estas fases se presentan en detalle en el Anexo 2.

A continuación, se describen los procesos abordados en cada fase.

4.1 COMPRENSIÓN DE LOS DATOS

En este apartado se revisó la estructura de la base de datos proporcionada por el equipo de investigadores de la universidad, así mismo se describió la cantidad de datos faltantes y por último la naturaleza de las variables.

4.1.1 DESCRIPCIÓN DE LOS DATOS

La base de datos denominada "BD Colaboradores", fue suministrada por un equipo de investigadores de la universidad, esta información se recolectó por medio de encuestas a colaboradores , a través de una batería de cuestionarios que evalúan las variables de resultado y los determinantes sociales. Para este estudio los cuestionarios estuvieron conformados por 292 preguntas, realizadas a 1441 colaboradores (incluyendo profesores, administrativos y directivos de la Universidad).

Una vez conocida la procedencia del dataset se procedió a explicar su estructura haciendo uso del lenguaje de programación Python en la herramienta Google Colab. Para esta tarea se emplearon bibliotecas como pandas, lo que permitió determinar del tamaño del conjunto de datos y los tipos de datos que contiene.

4.1.2 EXPLORACIÓN DE LOS DATOS

En esta sección se buscó conocer la estructura del dataset y las características de los datos, haciendo uso de medidas de tendencia central, gráficos y diferentes funciones para entender las fortalezas y debilidades en el dataset, así como diagnosticar los procesos que se deben realizar para construir una base de datos limpia.

A continuación, se muestra un resumen de las librerías utilizadas tanto en esta sección como a lo largo del proyecto:

Pandas

Es una librería de Python que ofrece un objeto llamado DataFrame (en adelante df) el cuál es muy eficiente para manipular datos, junto con herramientas para leer y escribir datos en diferentes formatos como CSV, Excel y bases de datos SQL. Ofrece herramientas para limpieza, análisis y reestructuración de bases de datos. Pandas es ampliamente utilizado en diversos campos como finanzas, neurociencia, economía y análisis web [47].

NumPy

Es la librería principal de programación de matrices para Python, es de código abierto y desarrollada por la comunidad. Ofrece un objeto de matriz Python multidimensional y funciones compatibles con matrices. La matriz NumPy es ampliamente utilizada debido a su simplicidad y se ha convertido en el formato estándar para el intercambio de datos de matriz en Python. Su importancia radica en su papel crucial en los procesos de análisis e investigación en una variedad de campos, desde la física y la química hasta la economía y la ingeniería [48].

Matplotlib

Es un paquete gráfico 2D utilizado en Python para el desarrollo de aplicaciones, scripts interactivos y generación de imágenes de calidad de publicación en diferentes interfaces de usuario y sistemas operativos [35].

Seaborn

Es una librería de Python para visualización estadística que simplifica el proceso de creación de gráficos complejos mediante una interfaz compatible con matplotlib y pandas. Sus funciones ofrecen una API orientada a conjuntos de datos, permitiendo la asignación de valores de datos a atributos visuales y así aplicar transformaciones estadísticas internas para decorar gráficos con etiquetas informativas y leyendas [49].

Cada una de estas librerías cuenta con una documentación online que permite identificar las funciones dentro de ellas, así como su manual de usuario para la aplicación de las mismas.

Inicialmente su usó la función `read_excel()` de pandas para leer el archivo con la base de datos y luego las funciones de python `Shape()` e `info ()` de pandas para conocer su estructura, en cuanto a cantidad de filas, columnas y tipos de datos de cada variable.

4.1.3 VERIFICACIÓN DE LA CALIDAD DE LOS DATOS

En la sección de verificación de la calidad de los datos, se usaron diferentes comandos para encontrar problemas dentro del dataset, como la presencia de valores faltantes, de igual manera se verificaron los tipos de datos en cada uno de los atributos. Para llevar a cabo este proceso, se

incorporó el comando `info()` para reconocer cuantos datos faltantes había por columna, adicional se usaron métodos como `isnull().sum()` de pandas para determinar el número total de datos faltantes por cada uno de los atributos [50].

Tomando como referencia la cantidad de valores nulos encontrados, se quiso revisar la existencia de valores duplicados usando la función `duplicated()` y únicos usando la función `unique()`, ambas de pandas, de tal manera que se pudiera hacer tratamiento posterior a dichos datos.

Usando gráficas de barras de seaborn se analizó la distribución de las variables respuesta verificando si existían datos atípicos o errores perceptibles.

Usando la función `describe()` de pandas se obtuvo la cantidad, media, desviación estándar, mínimo, máximo y cuartiles de las variables de exposición de tipo cuantitativo. Información importante para entender si hay datos atípicos. En cuanto a las variables de exposición cualitativas se utilizó la función de pandas `mode()` que permitió encontrar la moda de cada uno de los atributos para revisar la existencia de datos atípicos.

4.2. ANÁLISIS DE DATOS

Una vez realizada la fase de comprensión de los datos, se procede a seleccionar aquellos que cumplen con criterios de calidad, relevancia, integridad y precisión. Posteriormente, se lleva a cabo la limpieza de la base de datos para tratar los datos faltantes, y por último la construcción de los datos seleccionados. Este enfoque garantiza una base de datos sólida que permita un análisis correcto y efectivo

4.2.1 SELECCIÓN DE DATOS

En esta fase se verificó que el conjunto de datos proporcionado por la universidad incluyera todas las variables necesarias para el desarrollo del proyecto. De igual manera se tuvieron en cuenta factores como el objetivo del proyecto, ya que se deben elegir las variables con mayor peso para la investigación, igualmente la calidad de los datos es muy importante ya que si existen variables con una gran cantidad de valores nulos deben ser descartadas, debido a que no aportarían en el análisis.

Para asegurar la relevancia de las variables seleccionadas, se consultó a una experta en psicología. Esta consulta permitió validar que las variables elegidas eran las más adecuadas para abordar los objetivos del estudio desde una perspectiva psicológica. Además, se documentaron todas las decisiones tomadas durante el proceso de selección (ver anexo 6).

4.2.2 LIMPIEZA DE DATOS

En esta sección se narran los pasos desarrollados para el proceso de limpieza de la base de datos.

Eliminación de Variables Redundantes

Para descartar atributos, se realizó un análisis de correlación entre las variables cuantitativas de exposición. Este análisis permitió identificar las variables que mostraban una alta correlación entre sí, es decir, aquellas que tenían una relación muy fuerte y similar en sus valores. Al detectar estas variables altamente correlacionadas, se evaluó su redundancia para determinar cuáles podrían ser eliminadas.

Revisión de Datos Incompletos

Se realizó un análisis detallado de cada registro individual para identificar encuestados con un elevado número de respuestas incompletas. Utilizando la función isnull().sum() de Pandas, se contaron los valores faltantes en cada fila de la base de datos. Los registros con una cantidad significativa de datos faltantes fueron revisados y, en función de su impacto en el análisis, eliminados o imputados.

Con base en el análisis previo, se procedió a la limpieza de datos, aplicando las siguientes técnicas:

A continuación, se llevó a cabo la limpieza de datos, aplicando las siguientes técnicas:

Creación de nuevas columnas:

Hay atributos que sólo aplican para cierto tipo de colaboradores, por ejemplo, docentes, administrativos, etc. Se crearon nuevas columnas con la función apply de pandas para identificar a cada individuo según su rol en la institución. Se aplicó esta misma técnica para identificar individuos con características como condición psicológica, enfermedad, tipo de contrato de profesores y también para saber si los individuos contestaron o no a preguntas sobre NIVDEMCUANT (Demandas Cuantitativas), NIVJORLAB (Demanda de la jornada laboral), NIVINFLTRA (Influencia del trabajo sobre el entorno extralaboral), SEXPRESERV (Uso de preservativo).

Tratamiento para atributos con valores Nulos

Para manejar los valores nulos, se utilizaron distintas estrategias según la cantidad y tipo de datos faltantes. Las técnicas principales fueron la imputación y eliminacion de datos, como se describe a continuación:

Imputación con Medidas Estadísticas

En algunas variables con un porcentaje moderado de datos faltantes (por ejemplo, menos del 30-40%), se optó por realizar un proceso de imputación [50]. Este proceso consistió en reemplazar los valores faltantes utilizando medidas estadísticas como la media, la moda o la mediana, dependiendo del tipo de variable y la naturaleza de los datos. En la mayoría de las variables cualitativas, se realizó la imputación con moda, utilizando las funciones `isnull()` y `mode()` de Pandas.

Imputación con palabras

En algunas variables se reemplazaron los valores nulos con respuestas como “Nunca”, “Ninguna”, “prefiero no responder” haciendo uso de la función `fillna()` de Pandas y `nan()` de Numpy, revisando que sea una característica adecuada para la variable según el contexto de cada atributo en la base de datos.

Eliminación

Se identificaron algunos atributos con un alto porcentaje de valores faltantes, que superaba el umbral aceptable (por ejemplo, más del 80-90%). Para estos atributos, no fue posible aplicar un proceso de imputación efectivo, por lo que se decidió proceder con su eliminación. Asimismo, se eliminaron atributos que no aportaban datos relevantes o certeros al análisis, como el código de los individuos. Para llevar a cabo la eliminación, se utilizó la función `drop()` de Pandas.

4.2.3 CONSTRUCCIÓN DE DATOS

Luego de haber verificado que la base de datos no contara con valores nulos, se desarrolló un proceso de análisis de los atributos a predecir con el fin de identificar aquellas variables que fueran susceptibles de volver binarias, esto es que sus categorías pasen a ser únicamente dos.

El proceso se desarrolla con el fin de ajustar la base de datos para el posterior modelado con técnicas de machine learning de clasificación, que funcionan de mejor manera para variables dicotómicas. En este sentido se usó la función de Pandas `Replace()` para ejecutar el proceso.

Codificación de los datos

Parte del proceso de alistamiento de la base de datos para técnicas de machine learning incluyen la codificación de variables categóricas, pues para los algoritmos de machine learning es más complejo reconocer las diferencias para este tipo de variables. El proceso que se implementó para codificar fue *onehot encoding*. El cual toma cada variable y crea una nueva columna por cada categoría que esta tome, para luego poner el valor 1 si aparece dicha categoría o el valor 0 si no lo hace [29].

Este proceso requiere de una nueva librería de python llamada Scikit-Learn, la cual integra una amplia gama de algoritmos de aprendizaje automático de última generación para problemas supervisados y no supervisados de escala mediana y que se usa tanto en entornos académicos como comerciales [31].

El proceso se desarrolló usando la función `get_dummies()` de Scikit-Learn arrojando una nueva versión de la base de datos sin valores nulos y con una columna adicional por cada categoría de cada variable cualitativa. Esta base de datos es la que se llama pre-procesada y será el insumo principal para el desarrollo de los modelos de machine learning.

4.3. MODELADO

Para abordar la fase de modelado, se eligieron los modelos de Machine Learning para clasificación, se partió de una revisión previa a los modelos elegidos y con mejor rendimiento en proyectos similares, optando por Decision Trees, Random Forest y XG-Boost.

4.3.1 ELECCIÓN DEL MODELO

Cada uno de los modelos presentados tienen ventajas y desventajas que junto con los resultados de las métricas de evaluación proporcionan la información necesaria para elegir el modelo que realiza de mejor forma y con mayor eficiencia la predicción de las variables respuesta.

La siguiente tabla muestra algunas ventajas y desventajas de los modelos Random Forest y XGBoost.

Tabla 6.
Ventajas y desventajas Random Forest y XGBoost – Fuente [39]

	Ventajas	Desventajas
Random Forest	-Pueden usarse para clasificar o predecir. -Es más simple de entrenar que otros modelos más complejos, pero con rendimiento similar. -Tiene desempeño eficiente en bases de datos grandes. -Maneja gran cantidad de variables predictoras y puede identificar las más importantes para el resultado. -Mantiene su eficiencia incluso con gran cantidad de datos faltantes.	-La visualización de los datos puede ser difícil de interpretar. -Se puede sobreajustar cuando hay ruido. -Cuando los predictores son categóricos con diferente número de niveles, el clasificador puede sesgarse con las variables con más niveles. -Solo admite resultados discretos para sus resultados. -Si en los datos de entrenamiento no se representa el rango completo de las variables respuesta se puede generar sesgo. -No se tiene control total sobre lo que hace el modelo.

XGBoost	-Puede manejar bases de datos grandes con diferentes tipos de variables. -No se le dificulta el manejo de valores faltantes. -Sus resultados son muy precisos. -Tiene una muy buena velocidad de ejecución.	-Gran consumo de recursos computacionales para bases de datos grandes. Es necesario el ajuste correcto de hiperparámetros para evitar un sobreajuste. -Solo trabaja con vectores numéricos por lo que requiere una conversión previa de las variables categóricas.

Se eligen modelos basados en árboles considerando que su interpretación puede ser más sencilla para profesionales sin formación específica en ciencia de datos, lo que mejoraría el impacto del modelado para crear proyectos a partir de la identificación de variables fundamentales para desarrollar las predicciones.

4.3.2 ENTRENAMIENTO Y VALIDACIÓN

Para realizar el entrenamiento de los modelos, se hace separar aleatoriamente el conjunto de datos en dos subconjuntos, uno de entrenamiento, el cual sirve para enseñar al modelo los resultados de las clasificaciones permitiéndole aprender a realizar dichas clasificaciones, en el conjunto de validación se pone a prueba lo aprendido por el modelo, contrastando su clasificación con datos reales y desconocidos. Para verificar la eficiencia del proceso se utilizan métricas para estimar el error generado, como por ejemplo el mean squared error (MSE) [51]. En este caso se dividió el conjunto de datos de forma aleatoria de la siguiente manera: 70% de los datos para entrenamiento y 30% para validación.

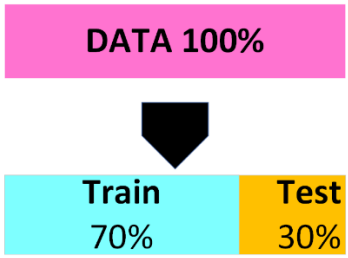


Ilustración 5. Particion train y test. Fuente Propia

En Python se usa la función train_test_split del módulo sklearn.model_selection de Scikit-Learn.

Sobreajuste en el entrenamiento

Existe un fenómeno al desarrollar el proceso de entrenamiento llamado sobreajuste o overfitting, el cual se da cuando en el conjunto de entrenamiento hay ruido en los datos o si presenta pocos datos. El modelo aprende sobre la base de este ruido y puede generar predicciones incorrectas. El sobreajuste se puede manifestar en métricas de evaluación muy altas y puede requerir de

procesos como reunir más datos de entrenamiento, reducir el ruido corrigiendo errores en los datos o eliminando valores atípicos [29].

4.3.3 CRITERIOS DE DESEMPEÑO PARA LOS MODELOS DE PREDICCIÓN

La matriz de confusión es un método utilizado para la evaluación del desempeño de modelos de clasificación como Random Forest y XGBoost, en ella se organizan los valores predichos por el modelo como verdaderos y falsos, y los valores reales en el conjunto de datos también como verdaderos y falsos [36], analizando las combinaciones que se den entre los resultados del modelo y los datos de la siguiente manera:

Tabla 7.
Matriz de confusión. Fuente [36]

		Predicciones	
		Negativos	Positivos
Datos reales	Negativos	VN	FP
	Positivos	FN	VP

Verdaderos positivos (VP): son aquellos valores que son positivos en el conjunto de datos y que el modelo predijo correctamente como positivos.
Verdaderos negativos (VN): son aquellos valores que son negativos en el conjunto de datos y que el modelo predijo correctamente como negativos.
Falsos positivos (FP): son aquellos valores que son negativos en el conjunto de datos y que el modelo predijo incorrectamente como positivos.
Falsos negativos (FN): son aquellos valores que son positivos en el conjunto de datos y que el modelo predijo incorrectamente como negativos.

Luego de construida la matriz de confusión para el modelo a evaluar se pueden definir las siguientes métricas para verificar su capacidad de predicción [25].

Accuracy

Muestra el desempeño global del modelo, midiendo la proporción de predicciones correctas sobre el total de predicciones.

$$Accuracy = \frac{VP + VN}{FP + FN + VP + VN}$$

Precision(P)

Mide la proporción de los valores predichos correctamente como positivos respecto al total de positivos predichos.

$$P = \frac{VP}{VP + FP}$$

Recall(R)

Mide la proporción de los valores predichos correctamente como positivos respecto al total de positivos reales.

$$R = \frac{VP}{FN + VP}$$

Specificity(S)

Mide la proporción de valores predichos correctamente como negativos respecto del total de negativos reales.

$$S = \frac{VN}{VN + FP}$$

F1-Score(F1)

Resume la Precision y Recall en una sola métrica, siendo obtenida como su media armónica y utilizada comúnmente en bases de datos desbalanceadas.

$$F1 = 2\frac{P * R}{P + R}$$

Curva ROC

La curva ROC (Receiver Operating Characteristics) relaciona los valores de R contra 1-S. Mostrando los valores identificados como positivos. El AUC_ROC (Area Under the Curve) muestra la probabilidad de que, al tomar dos muestras al azar, una positiva y una negativa, el modelo asigne una probabilidad más alta a la muestra positiva que a la muestra negativa, lo que indica una correcta clasificación [36].

La gráfica de la curva ROC muestra un mejor desempeño cuando la curva se aleja de la diagonal (línea de 45 grados) y se acerca a la esquina superior izquierda del gráfico. Esto indica que el modelo tiene una alta tasa de VP y una baja tasa de FP [29].

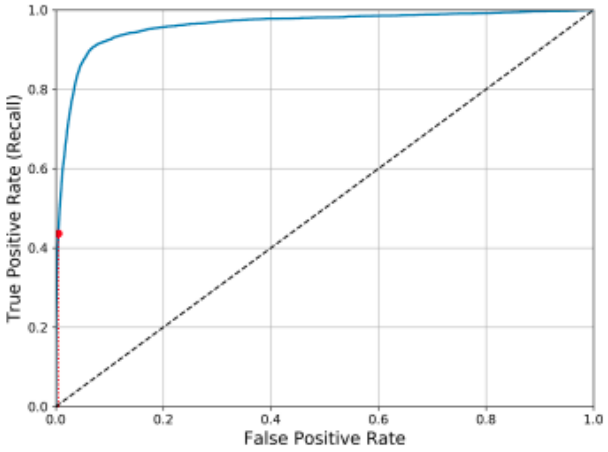


Ilustración 6. Gráfico curva ROC de un modelo con buen desempeño, Fuente [29]

Estas medidas otorgan un panorama general del desempeño de los modelos ejecutados y permite compararlos entre sí.

Para el cálculo de las métricas se usan las funciones confusion_matrix, recall_score, precision_score, f1_score, accuracy_score del módulo sklearn.metrics de Scikit-Learn, junto con la librería seaborn para generar los gráficos de las matrices. Del mismo módulo se usa roc_curve, auc para generar las curvas ROC y el cálculo del AUC junto con matplotlib para crear las gráficas.

4.3.4 INTERPRETACION DEL MODELO

Importancia relativa de las características de predictoras después del entrenamiento

En los modelos basados en Decision Trees no todas las variables predictoras aportan significativamente al resultado de la predicción. Existen variables que contribuyen de forma sustancial a la respuesta [52], identificar estas variables es un proceso importante en modelos como Random Forest y XGBoost.

4.3.4.1 METODO FEATURE IMPORTANCES

La propiedad “feature_importances_” es una técnica comúnmente utilizada en modelos de árboles de decisión para evaluar la importancia de las características. Implementada en librerías como scikit-learn, “feature_importances_” asigna un valor a cada característica según su contribución a la mejora de la precisión del modelo. Este valor se calcula como la reducción promedio de impureza (índice Gini o entropía) que la característica aporta en los nodos de los árboles donde se utiliza. Esta técnica ofrece una forma directa y eficiente de interpretar la relevancia de cada característica en el modelo, facilitando la identificación de las variables más influyentes [53].

4.3.4.2 MÉTODO SHAP

Una de las técnicas más avanzadas y precisas para la interpretación es SHAP (SHapley Additive exPlanations), esta se basa en la teoría de valores de Shapley de la teoría de juegos, SHAP proporciona una manera coherente de asignar la importancia de cada característica en una predicción específica, el gráfico "waterfall" de SHAP para XGBoost descompone la predicción de un modelo en contribuciones aditivas de las características. Comienza con el valor esperado de la predicción (base value) y muestra cómo cada característica aumenta o disminuye este valor hasta alcanzar la predicción final. Las contribuciones positivas y negativas se suman secuencialmente, lo que permite visualizar claramente el impacto individual de cada característica en la predicción del modelo [54].

5. RESULTADOS

5.1 COMPRENSIÓN DE LOS DATOS

En este apartado se presentó una visión panorámica de la base de datos proporcionada por el equipo de investigadores de la universidad, inicialmente se hizo una descripción del Dataset, su estructura, tamaño y naturaleza de las variables, luego se desarrolló un análisis exploratorio de los datos, con el fin de conocer a profundidad el Dataset y encontrar posibles problemas como: datos faltantes, inconsistencias en el formato de los datos, problema de escala, y demás. Lo que resulta de vital importancia a la hora de tomar decisiones sobre el proceso de limpieza de datos.

En la ilustración 12 se presenta un pequeño fragmento de los datos suministrados por la universidad

	CODIGO	NIVDEP	NIVANS	NIVEST	NIVSOLED	NIVIDEASUIC	NIVRESIL	NIVSATVIDA	NIVRECPSIC	RPSoptimis	...
0	1	normal	normal	normal	normal	Sin indicadores	alta	Alta	Alto	Casi todo el tiempo	...
1	5	normal	normal	leve / moderado	moderado	Sin indicadores	alta	Alta	Alto	Casi todo el tiempo	...
2	6	normal	normal	normal	normal	Sin indicadores	media	Alta	Alto	Casi todo el tiempo	...
3	10	normal	normal	normal	normal	Sin indicadores	alta	Alta	Alto	Casi todo el tiempo	...
4	11	normal	normal	normal	normal	Sin indicadores	alta	Alta	Alto	Casi todo el tiempo	...

Ilustración 7. Fragmento datos suministrados por la universidad. Fuente Propia

5.1.1 DESCRIPCIÓN DE LOS DATOS

La base de datos BD Colaboradores (JUNIO 2023) MCD.xlsx está formada por una población de colaboradores que incluye a todos los profesores con contrato de planta y hora cátedra, los trabajadores con cargos operativos y administrativos y los directivos académicos y administrativos de las unidades académicas de la Universidad, con contrato durante el periodo académico 2022-2, que aceptaran participar en el estudio. Además, el total de preguntas dentro de la encuesta colaboradores fue de 293. Por lo tanto, el conjunto de datos cuenta con 988 filas (colaboradores) y 297 columnas (atributos).

La base de datos está formada por 6 variables de respuesta divididas en indicadores positivos y negativos, las cuales son de tipo categórico binario, el resto de las columnas corresponden a las

variables de entrada que son determinantes sociales, algunos de tipo numérico y otros categóricos.

Al estudiar la base de datos (Anexo 1) se encuentra un 15% de datos nulos, los cuales están relacionados con datos faltantes por diligenciar en la encuesta, preguntas que están relacionadas con roles específicos en la universidad y que no responden todos los encuestados o relacionados a valores nulos que dejan preguntas que no aplican a ciertos participantes.

Los tipos de variables asociadas a la encuesta se distribuyen de la siguiente manera:

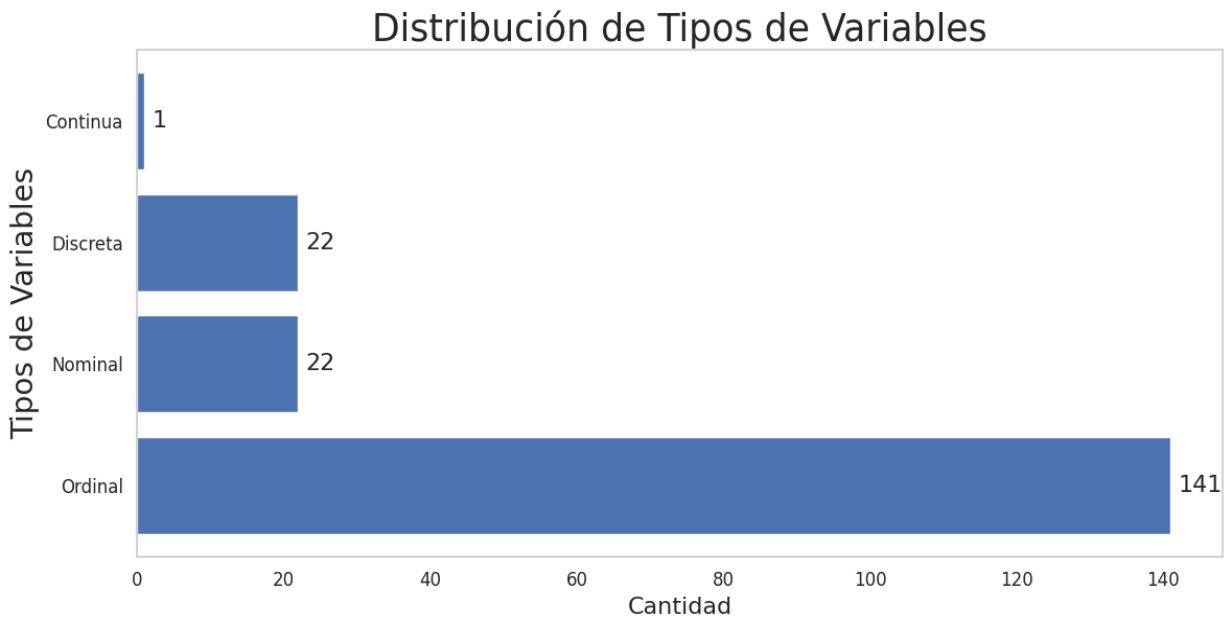


Ilustración 8. Tipo de variables. Fuente propia.

Se evidencia una gran cantidad de datos asociados a variables categóricas, en especial asociados a escalas ordinales y clasificaciones dicotómicas, en menor medida se ven variables categóricas nominales y variables cuantitativas discretas y continuas. Esto es producto del desarrollo de una encuesta que aborda principalmente preguntas en donde el participante responde de acuerdo con una escala tipo Likert o preguntas de respuesta es sí o no.

5.1.2 EXPLORAR LOS DATOS

En esta sección se conocerán la estructura del Dataset y las características de los datos, haciendo uso de medidas de tendencia central, gráficos y diferentes funciones.

A continuación, se describen los resultados encontrados luego de realizar el análisis:

- En la detección de datos faltantes se obtuvo que la base de datos cuenta con un total de 44740 datos faltantes es decir un 15.2469% del total de datos. Existen variables con 0

datos faltantes otras con menos de 50 e incluso algunas con más de 900 datos faltantes. En la siguiente gráfica se presenta la distribución de datos faltantes por variables.

- En la obtención de valores únicos por cada variable, se encontró que hay variables que tienen diferentes tipos de datos, las cuales deben ser codificadas para evitar problemas a lo largo del análisis.
- Se encontraron variables que presentaron diferentes escalas en los datos.
- Algunas variables fueron contestadas por un grupo específico de colaboradores.
- Teniendo en cuenta los resultados anteriores se plantearon unas estrategias generales para darle solución a los problemas encontrados en la base de datos.

5.1.3 VERIFICAR LA CALIDAD DE LOS DATOS

Como parte del análisis de la calidad de los datos, se realizó una gráfica que representa la cantidad de variables en función de la cantidad de valores nulos (Ver Ilustración 14) donde se observan variables con cero valores nulos, y otras donde no existe información y todos los valores son faltantes.

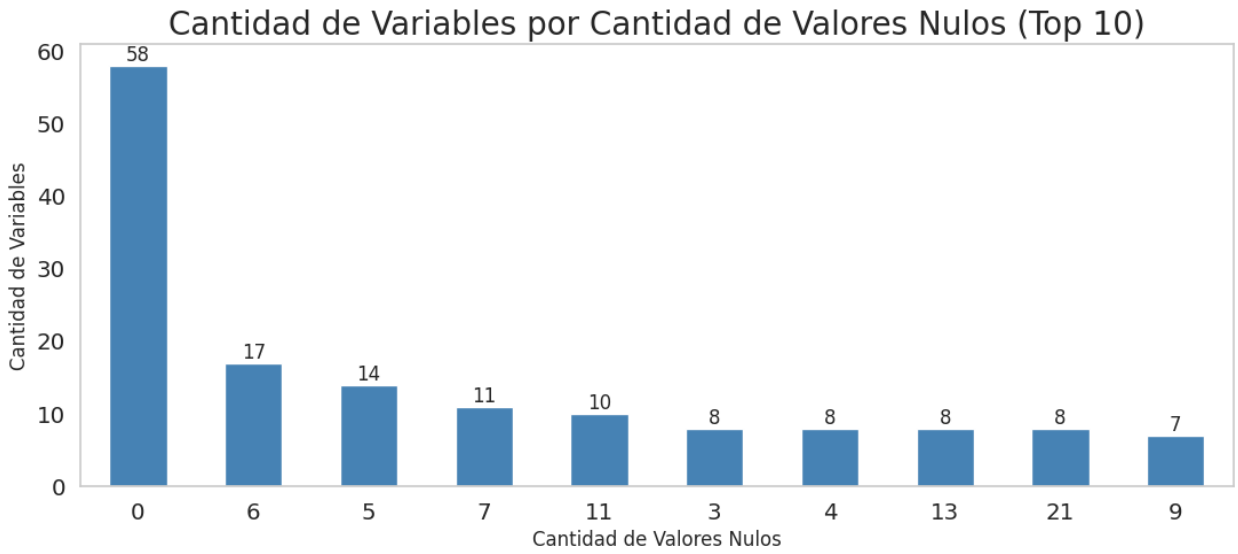


Ilustración 9. Cantidad de Variables por Cantidad de Valores Nulos Top 10. Fuente Propia.

Por otro lado, también se validó la presencia de valores redundantes con una matriz de correlación de las variables de entrada (Ver Ilustración 15 y 16), concluyendo que no existe redundancia entre las variables predictoras.

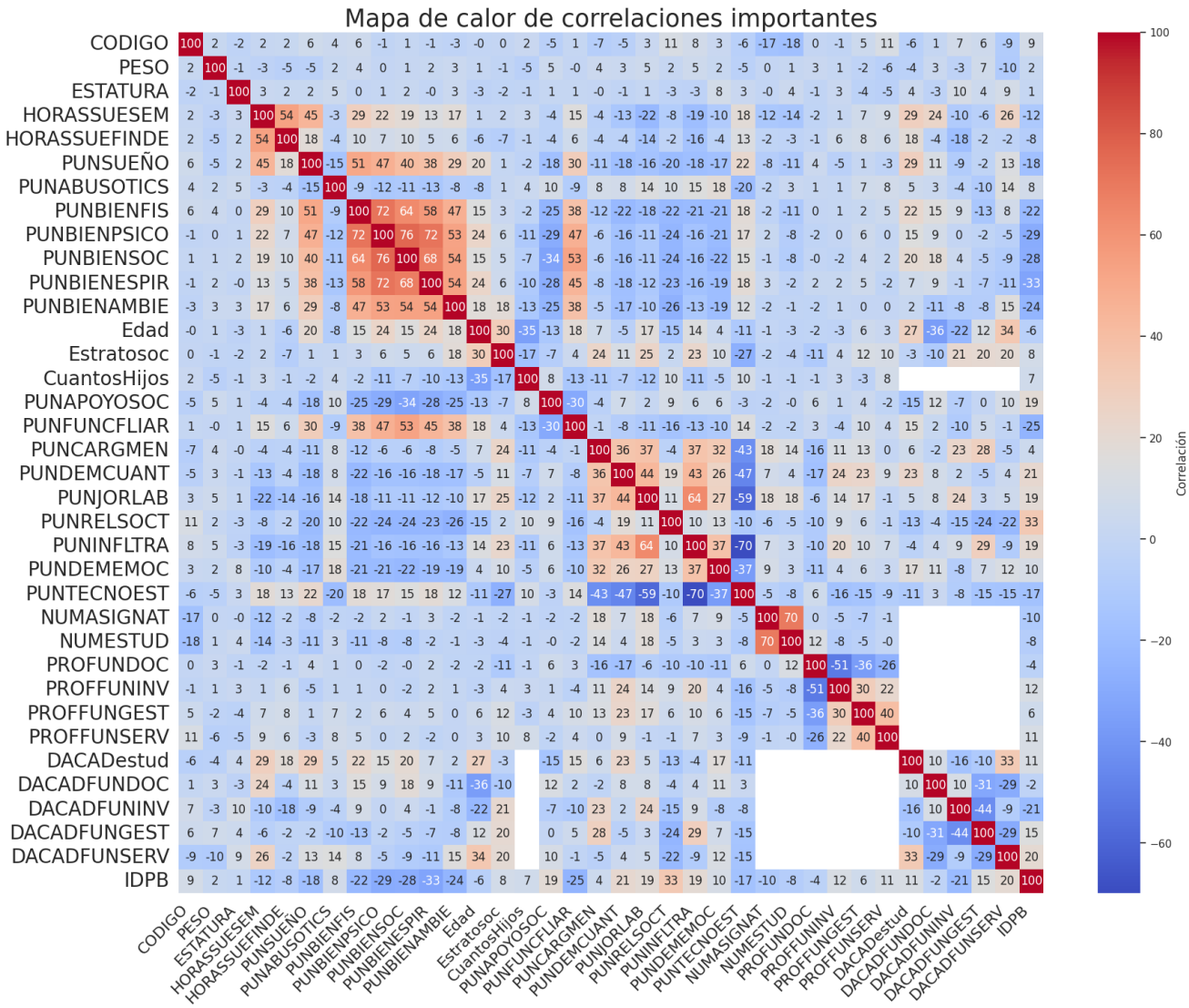


Ilustración 10. Mapa de calor Correlaciones. Fuente Propia.

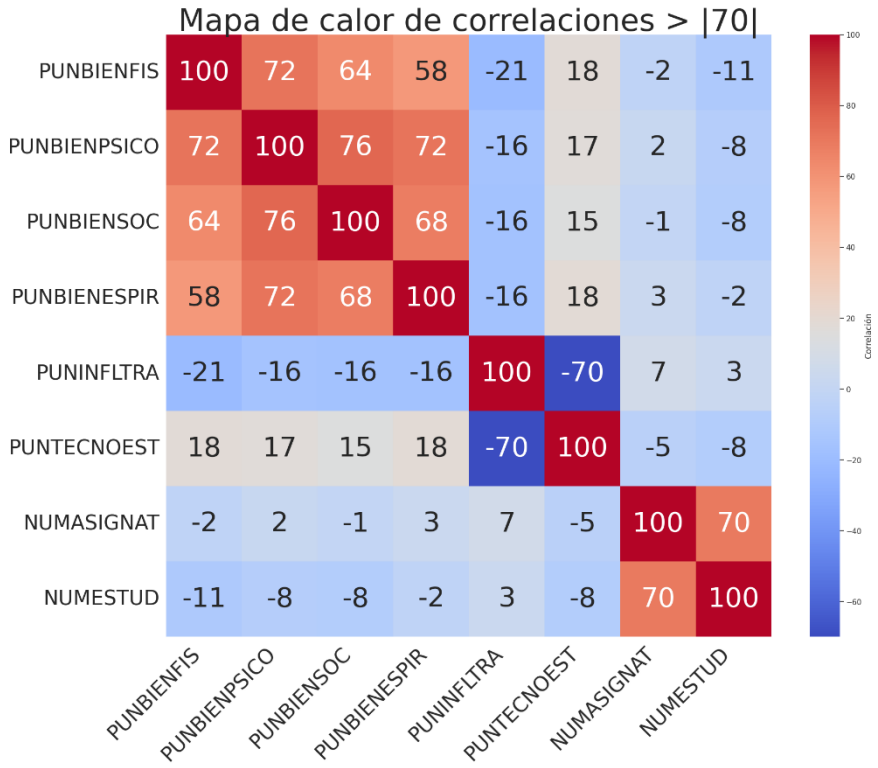


Ilustración 11. Mapa de calor Mayores correlaciones. Fuente Propia.

5.2. ANÁLISIS DE DATOS

Después de seleccionar y limpiar cuidadosamente los datos, construimos una base de datos robusta con información precisa y relevante. Este proceso minucioso aseguró la calidad de los datos, permitiéndonos realizar un análisis efectivo. A continuación, se presentan los resultados obtenidos en los procesos de selección, limpieza y construcción de los datos.

5.2.1 SELECCIÓN DE DATOS

Como resultado de esta fase, se generó un documento complementario, etiquetado como Anexo 4, que contiene el diccionario completo de variables, con una descripción detallada de cada una. Este documento servirá como referencia para los análisis posteriores y contribuirán significativamente a la comprensión y la interpretación de los resultados obtenidos.

5.2.2 LIMPIEZA DE DATOS

Una vez se ha realizado la etapa de exploración de los datos, se procede a realizar una limpieza de cada una de las variables contenidas dentro del Dataset, para este caso en específico se estudió cada variable por separado y se le categorizó según su estado, si presentaba o no valores nulos, y se generaron hipótesis sobre los motivos de los valores faltantes, de acuerdo a esto se plantearon estrategias que permitieron solucionar todos y cada uno de los problemas que

presentaba la base de datos, comenzando por las variables que tenían más del 90% de valores nulos, las cuales fueron eliminadas, en el caso de las variables que fueron contestadas por un rol específico (profesor, académico, administrativo), fueron imputadas.

La imputación se hizo dependiendo del tipo de variable, para las variables categóricas, se imputó usando la moda, palabras específicas o categorías que surgieron al analizar la base de datos. Para las variables cuantitativas se usó la media o mediana, también se estandarizaron algunos valores. Por último, se crearon variables que indican por quién fue contestada cada pregunta. Para acceder al análisis particular por variable y las estrategias de solución se puede observar el Anexo 6.

5.2.3 CONSTRUCCIÓN DE DATOS

Teniendo en cuenta que se obtiene una base de datos limpia, en la fase de construcción se busca crear una base de datos que admita la aplicación de modelos de Machine Learning, para este fin se deben transformar los valores de algunas variables de categóricas a numéricas, de tal manera que se puedan analizar de forma dicotómica o en escala.

El proceso se desarrolló identificando las variables dicotómicas (cuyas respuestas sean por ejemplo Si o No) y se reemplazaron por 0 y 1. De igual manera se buscaron las variables producto de escalas Likert, ya sean del tipo bajo, medio alto; leve, moderado, alto, etc. Y se transformaron a valores numéricos cada escala, esto teniendo en cuenta que para los algoritmos de machine learning se trabaja mejor con escalas numéricas. El procedimiento se encuentra descrito completamente en el anexo 7.

Tratamiento de las variables de respuesta:

Teniendo en cuenta que los algoritmos de clasificación que se usan en este proyecto mejoran su implementación al usar variables respuesta que sean dicotómicas se opta por agrupar los valores de dichas variables. Esta decisión simplifica el problema a dos posibles resultados, lo que no solo mejora el rendimiento del modelo, sino que también facilita la interpretación de los resultados. Al utilizar clasificación binaria, es más fácil ajustar el modelo y entender cómo se están haciendo las predicciones, lo que asegura que los resultados sean claros.

A continuación, se ilustra claramente cómo se agrupan los valores en categorías binarias.

Tabla 8.
Tratamiento de las variables de respuesta. Fuente Propia.

Variable	Tratamiento
NIVDEP: Depresión NIVANS: Ansiedad NIVEST: Estrés	“Normal” pasa a ser 0 “severo / extremadamente severo” y "leve / moderado" a ser 1 Se decide que en estado normal se toma como indicador de no requerir atención, más con “severo / extremadamente severo" y "leve / moderado" el valor 1 indica que requiere atención
NIVSOLED: Soledad	“Normal” pasa a ser 0 "severo" y "moderado" pasan a ser 1 Se decide que en estado normal se toma como indicador de no requerir atención, más con "severo" y "moderado" el valor 1 indica que requiere atención
NIVIDEASUIC: Ideación suicida	“sin indicadores”: 0 “Con algún indicador”: 1
NIVRESIL: Resiliencia NIVSATVIDA: Satisfacción con la vida NIVRECPSIC: Recursos psicológicos	“Alto”: 1 “medio” y “Bajo”: 0 Se toma el 1 como indicador de buen desarrollo de componentes positivos, y “medio” y “bajo” como indicador de alerta.

Teniendo en cuenta la construcción anterior se muestra a continuación las distribuciones comparadas:

Factores Negativos

Antes

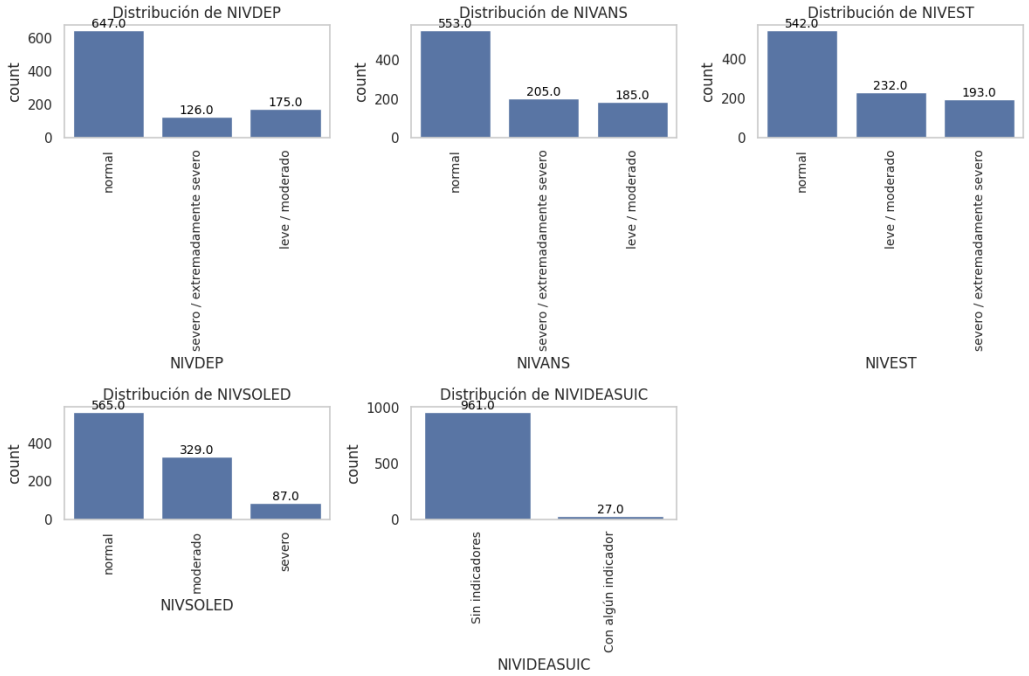


Ilustración 12. Distribución Factores Negativos por Clases. Fuente Propia.

Después

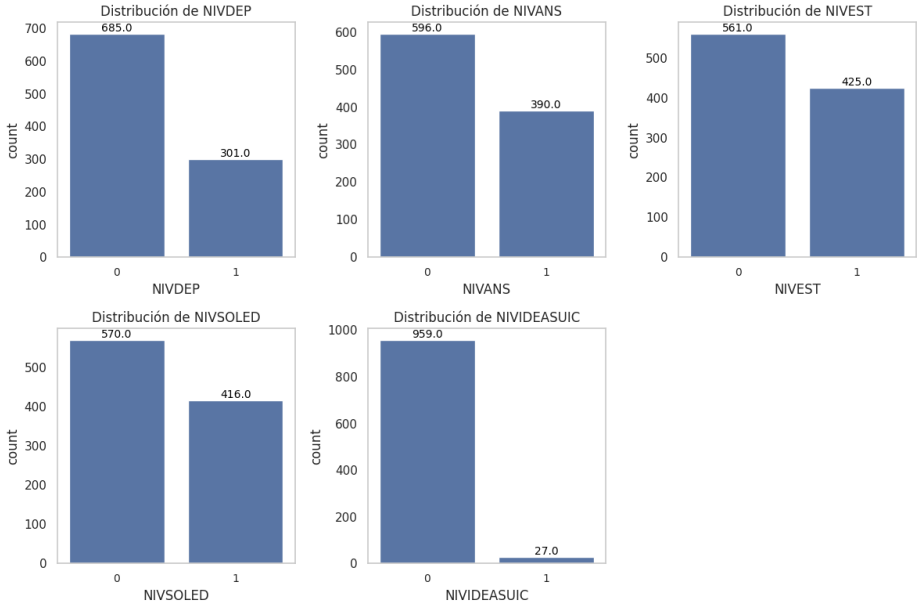


Ilustración 13. Distribución Factores Negativos Codificados. Fuente Propia.

Factores Positivos:

Antes:

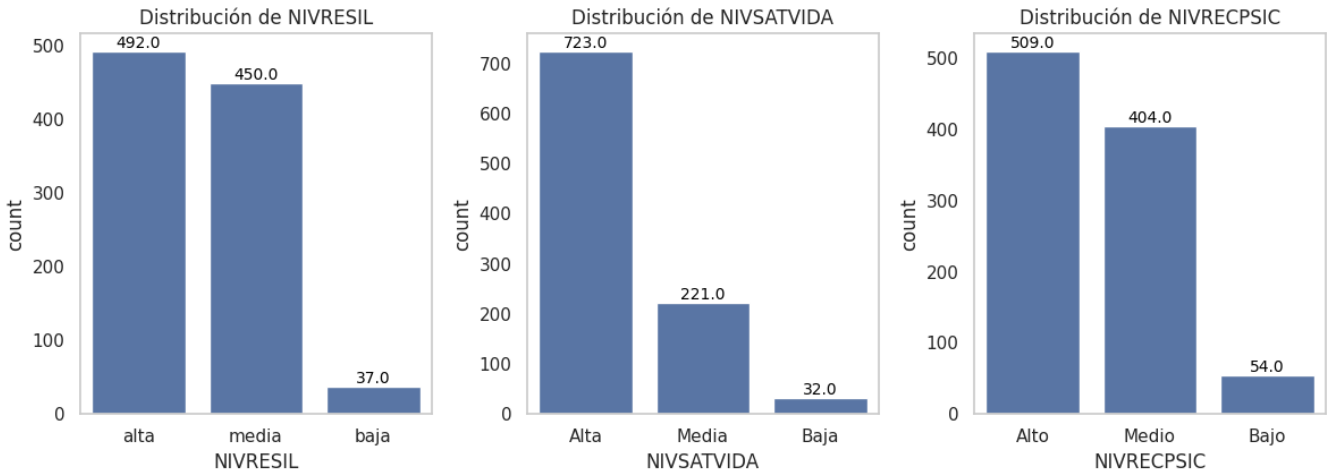


Ilustración 14. Distribución Factores Positivos por Clases. Fuente Propia.

Después:

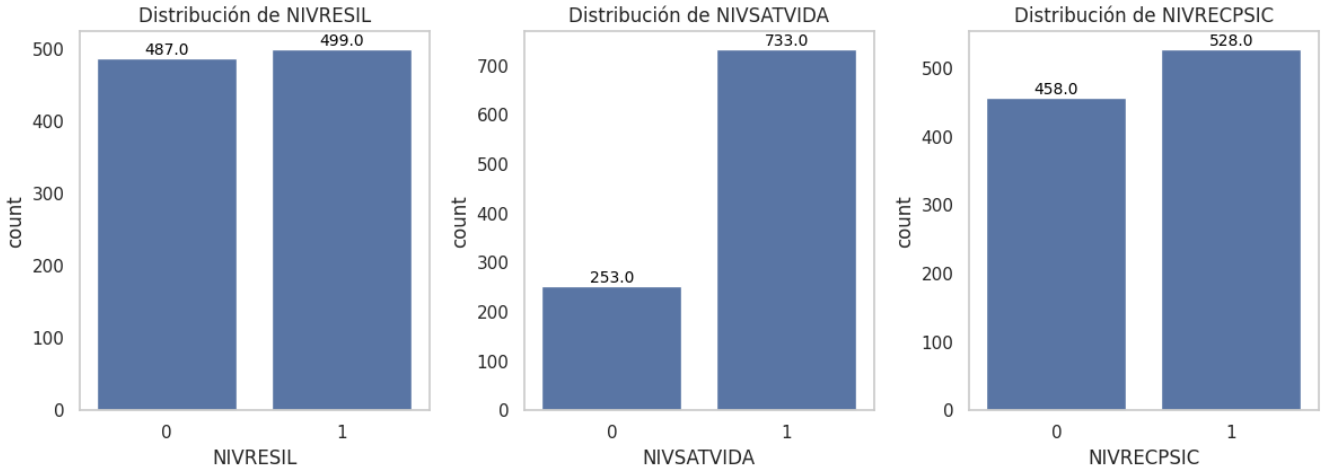


Ilustración 15. Distribución Factores Positivos Codificados. Fuente Propia.

Con este último tratamiento la base de datos queda lista para el proceso de modelado.

5.3. MODELADO

Es una etapa crucial en el proceso de aprendizaje automático, aquí se selecciona y ajusta un algoritmo para aprender de los datos y realizar predicciones o clasificaciones. En este caso, el objetivo es construir un modelo de clasificación, ya que se busca resolver un problema de categorización.

5.3.1 SELECCIÓN DE LA TÉCNICA DE MODELADO

Para la selección del modelo, se consideran varios factores cruciales. En primer lugar, se analiza el tipo de problema a resolver; en este caso, estamos abordando un problema de clasificación, lo que implica la necesidad de un algoritmo de aprendizaje supervisado. Además, se tienen en cuenta otros elementos determinantes, como el volumen de datos disponibles, la capacidad de interpretación del modelo y la naturaleza de las variables involucradas. Es esencial tener en cuenta que ciertos modelos pueden funcionar de manera más efectiva con ciertos tipos de variables. Asimismo, se presta atención al equilibrio entre las distintas clases de la variable objetivo, ya que esto puede influir significativamente en el rendimiento del modelo. Considerar cuidadosamente estos factores es crucial para seleccionar el modelo más adecuado y obtener resultados precisos y confiables en el proceso de modelado.

En la siguiente tabla se presenta una comparativa con algunas técnicas de aprendizaje supervisado [29], Ver tabla.

Tabla 9.
Ventajas y Desventajas Modelos de Aprendizaje Supervisado. Fuente [23].

Modelo	Ventajas	Desventajas
Linear Regression	Simple de implementar y más fácil de interpretar y entrenar	Puede sufrir con valores atípicos, no captura relaciones no lineales
Random Forest	No es necesario transformar variables ni ajustar parámetros	Difícil de interpretar, sobreajuste con conjuntos grandes
SVM	Efectivas en espacios de alta dimensión, Buen desempeño con conjuntos de datos pequeños y medianos	Dificultad en la interpretación, no es adecuado para grandes conjuntos de datos
Neural Networks	Adaptabilidad a diferentes tipos de problemas, buen rendimiento para grandes conjuntos de datos	Tiempo y recursos computacionales, propenso al sobreajuste, difícil de interpretar
Decision Trees	Fácil interpretación, no requiere gran procesamiento	Propenso al sobreajuste, Tendencia al sesgo en clases minoritarias
XG Boost	Puede manejar grandes conjuntos de datos con eficiencia, Implementa técnicas de regularización para evitar el sobreajuste	Consumo de recursos computacionales, su entrenamiento puede ser lento en conjuntos de datos muy grandes

Parámetros de los modelos
Teniendo en cuenta la elección de los modelos basados en árboles Random Forest y XGBoost, se generaron los modelos estándar o por defecto que cuentan con los siguientes parámetros:

Random Forest
criterion='gini'
max_depth=None
min_samples_split=2
min_samples_leaf=1
max_features='auto'

XGBoost
booster='gbtree'
learning_rate=0.3
n_estimators=100
max_depth=6
subsample=1
colsample_bytree=1

5.3.2 DIVISIÓN DE DATOS

Se dividen los datos en conjuntos de entrenamiento y prueba. El conjunto de entrenamiento se utiliza para ajustar el modelo, y el conjunto de prueba se utiliza para evaluar el rendimiento final del modelo.

5.3.3 ENTRENAMIENTO DEL MODELO

Se aplica la técnica de modelado seleccionada al conjunto de entrenamiento y se ajustan los parámetros del modelo para minimizar el error en los datos de entrenamiento.

5.3.4 VALIDACIÓN DEL MODELO

Evaluar el rendimiento del modelo utilizando el conjunto de validación. Esto puede implicar la comparación de métricas de rendimiento como precisión, sensibilidad, especificidad, etc., dependiendo del tipo de problema que se esté abordando.

Para mejorar la validación del modelo se creó una base de datos en la que se balanceen los valores en las variables que presentaban mayor diferencia entre sí, (Ver ilustración 21) se muestra la distribución de los datos en las variables respuesta sin balancear.

Se procede a desarrollar un proceso de balanceo por submuestreo de la clase mayoritaria para las variables respuesta sin contar la variable NIVIDEASUIC ya que cuenta con un desbalance muy marcado que impide desarrollar el proceso, tampoco se aplica el proceso a la variable NIVRESIL

ya que cuenta con una distribución balanceada.

Al revisar los datos para la evaluación se comparará si el hecho de hacer el balance de los datos mejora el rendimiento del modelo.

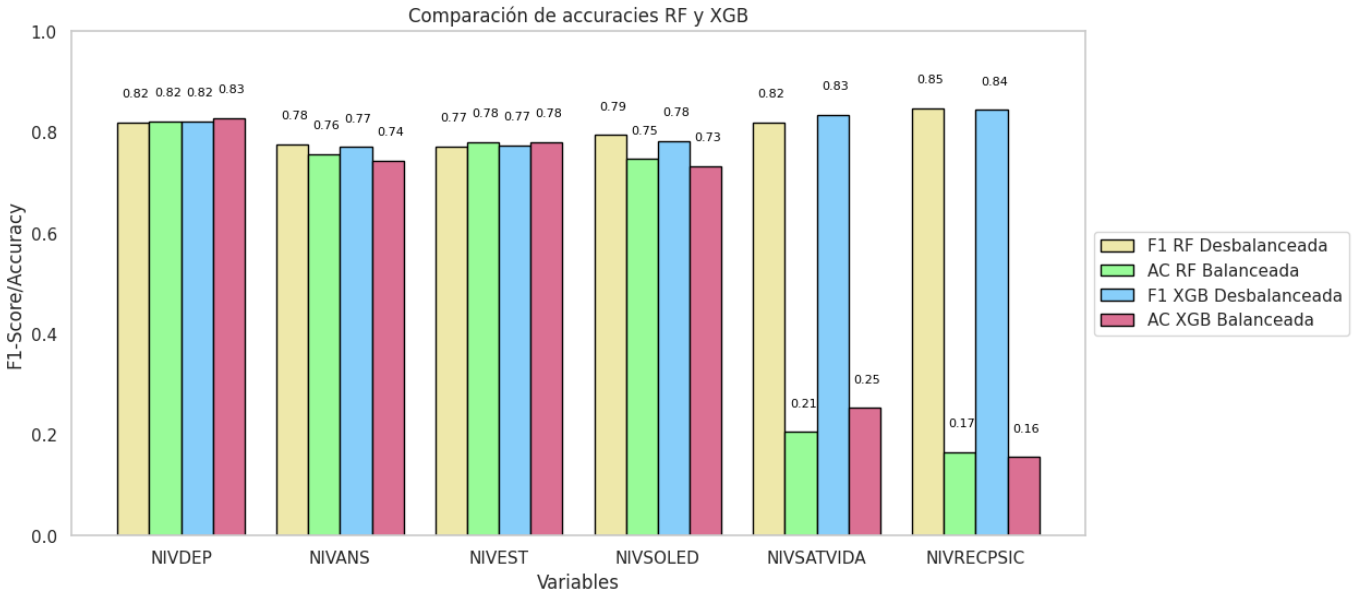


Ilustración 16. Comparación de rendimiento Random Forest (RF) Vs XGBoost (XGB), Datos Balanceados Vs Desbalanceados. Fuente Propia.

Como se puede observar, las diferencias entre las métricas de las bases de datos balanceadas y no balanceadas no son significativas en todas las variables y no justifican la pérdida de información que se hizo al generar los submuestreos.

5.3.5 ESTIMACIÓN DE HIPERPARÁMETROS

Se desarrollaron procesos de estimación de hiperparámetros definidos en [29] para afinar la selección de los parámetros de cada modelo, buscando la combinación que mejores resultados proporcione.

El primero es ‘Grid Search’ el cual evalúa el rendimiento de todas las combinaciones posibles de un conjunto establecido para los hiperparámetros. Por otro lado, ‘Random search’ evalua un número aleatorio en dicho conjunto de datos. El primero es más exhaustivo y así mismo más costoso en tiempo y recursos, el segundo es más eficiente, pero puede ignorar los valores que garanticen la mejor combinación.

Teniendo en cuenta lo anterior y haciendo pruebas de eficiencia para ambos métodos se decide establecer el método de Random Search, pues con pruebas en muestras aleatorias de variables se verificó la ineficiencia de Grid Search, su alto costo computacional y no se lograron aumentos significativos en las métricas obtenidas para las variables seleccionadas.

Random Search para Random forest

Se ejecutó el proceso de análisis de hiperparámetros para random search en la base de datos desbalanceada encontrando que los mejores modelos cuentan con la siguiente definición de conjuntos y parámetros:

Tabla 10.
Mejores Parametros Random Search - Random Forest. Fuente Propia.

Variable	n_estimators	max_depth	min_samples_split	min_samples_leaf	bootstrap
Depresión	153	30	7	2	False
Ansiedad	108	40	8	1	True
Estrés	225	40	2	1	False
Soledad	277	10	13	17	False
Ideación suicida	171	30	12	15	True
Resiliencia	262	20	9	17	True
Satisfacción con la vida	450	50	15	4	False
Recursos Psicológicos	212	50	12	1	False

Los resultados de estos modelos en comparación con los modelos default para random forest se muestran a continuación:

Tabla 11.
Comparación Resultados Random Search - Modelo Default. Fuente Propia.

	Random Search		Modelo por defecto	
Modelo	**Precisión**	**F1-score**	**Precisión**	**F1-score**
Depresión	**0.841216**	**0.836301**	0.827703	0.818846
Ansiedad	0.773649	0.771841	**0.777027**	**0.775469**
Estrés	**0.777027**	**0.776097**	0.773649	0.771719
Soledad	0.766892	0.762325	**0.797297**	**0.794318**
Ideación suicida	**0.983108**	**0.974734**	**0.983108**	**0.974734**
Resiliencia	**0.760135**	**0.759963**	0.733108	0.732806
Satisfacción con la vida	**0.834459**	**0.82249**	0.827703	0.819461
Recursos Psicológicos	0.841216	0.841069	**0.847973**	**0.847763**

Resalta el hecho de que no en todos los casos los casos los modelos por defecto son superados por los encontrados por Random Search, y cuando esto sucede no se genera un cambio significativo; la diferencia más notoria se da en el modelo de Resiliencia, en el cual el modelo encontrado por Random Search supera en tres puntos porcentuales las métricas del modelo estándar.

Random Search para XGBoost

Se ejecutó el proceso de análisis de hiperparámetros para random search en la base de datos desbalanceada con los siguientes conjuntos:

Tabla 12.
Mejores Parametros Random Search - XGBoost. Fuente Propia.

Variable	colsample_bytree	Learning_rate	max_depth	n_estimators	subsambple
Depresion	1.0	0.1	9	64	0.8
Ansiedad	1.0	0.01	4	179	0.8
Estrés	1.0	0.05	7	51	0.8
Soledad	1.0	0.1	9	101	0.6
Ideación suicida	0.6	0.1	9	192	1.0
Resiliencia	0.6	0.01	9	116	0.8
Satisfacción con la vida	0.8	0.05	6	70	0.6
Recursos Psicológicos	0.6	0.3	6	82	1.0

Los resultados de estos modelos en comparación con los modelos default para XGBoost se muestran a continuación:

Tabla 13.
Comparación Resultados XGBoost - Modelo Default. Fuente Propia.

	XGBoost		Modelo por defecto	
Modelo	**Precisión**	**F1-score**	**Precisión**	**F1-score**
Depresión	**0.841216**	**0.838455**	0.827703	0.818846
Ansiedad	0.773649	0.775414	**0.777027**	**0.775469**
Estrés	**0.790541**	**0.790541**	0.773649	0.771719
Soledad	0.770270	0.768383	**0.797297**	**0.794318**
Ideación suicida	**0.983108**	**0.982184**	**0.983108**	**0.974734**
Resiliencia	**0.766892**	**0.766852**	0.733108	0.732806

Satisfacción con la vida	**0.837838**	**0.829620**	0.827703	0.819461
Recursos Psicológicos	0.858108	0.857875	**0.847973**	**0.847763**

De manera similar a como sucede en los modelos Random Forest, no se encuentran diferencias significativas entre los modelos encontrados y los modelos estándar, nuevamente la mayor diferencia se da en resiliencia y en el resto de los modelos no hay diferencias significativas.

Por este motivo y priorizando la reproducibilidad de los modelos se decidió trabajar con los modelos estándar, de tal manera que se pueda asegurar su consistencia al poder replicarlo en otros contextos, así mismo se busca que el modelo sea sencillo de implementar y entender, además de buscar eficiencia y eficacia sin sacrificar tiempo y recursos computacionales.

5.4 EVALUACIÓN

Para llevar a cabo la evaluación de los modelos desarrollados en este proyecto, se utilizaron diferentes métricas como es el caso de Accuracy, F1 score, matriz de confusión y curva Roc. Además, para asegurarnos de que los modelos funcionen bien con datos nuevos y no solo con los datos de entrenamiento, empleamos un proceso de validación cruzada. A continuación, se presentan los resultados obtenidos por cada uno de los modelos estándar (Arboles de decisión, Random Forest y XGBoost) en términos de rendimiento.

Tabla 14.
Resultados Métricas de evaluación para el modelo Árbol de Decisión. Fuente Propia.

Variable	Accuracy	Recall	Precision	F1-Score	Specificity
Depresión (NIVDEP)	0.743243	0.743243	0.739218	0.740663	0.823834
Ansiedad (NIVANS)	0.685811	0.685811	0.697686	0.688949	0.688889
Estrés (NIVEST)	0.729730	0.729730	0.730542	0.730003	0.739130
Soledad (NIVSOLED)	0.635135	0.635135	0.634289	0.634596	0.679012
Ideación suicida (NIVIDEASUIC)	0.966216	0.966216	0.976585	0.970926	0.975945
Resiliencia (NIVRESIL)	0.635135	0.635135	0.635258	0.634935	0.657718
Satisfacción con la vida (NIVSATVIDA)	0.736486	0.736486	0.744482	0.739942	0.583333
Recursos psicológicos (NIVRECPSIC)	0.719595	0.719595	0.719685	0.719566	0.729730

Tabla 15.
Resultados Métricas de evaluación para el modelo Random Forest. Fuente Propia.

Variable	Accuracy	Recall	Precision	F1-Score	Specificity
Depresión (NIVDEP)	0.827703	0.827703	0.832326	0.818846	0.948187
Ansiedad (NIVANS)	0.777027	0.777027	0.775089	0.775469	0.838889
Estrés (NIVEST)	0.773649	0.773649	0.775032	0.771719	0.844720
Soledad (NIVSOLED)	0.797297	0.797297	0.802338	0.794318	0.888889
Ideación suicida (NIVIDEASUIC)	0.983108	0.983108	0.966502	0.383243	1.000000
Resiliencia (NIVRESIL)	0.733108	0.733108	0.733864	0.267473	0.765101
Satisfacción con la vida (NIVSATVIDA)	0.827703	0.827703	0.821772	0.819461	0.571429
Recursos psicológicos (NIVRECPSIC)	0.847973	0.847973	0.849906	0.847763	0.810811

Tabla 16.
Resultados Métricas de evaluación para el modelo XGBoost Fuente Propia.

Variable	Accuracy	Recall	Precision	F1-Score	Specificity
Depresión (NIVDEP)	0.824324	0.824324	0.821743	0.821539	0.896373
Ansiedad (NIVANS)	0.770270	0.770270	0.771022	0.770610	0.805556
Estrés (NIVEST)	0.773649	0.773649	0.773526	0.773578	0.795031
Soledad (NIVSOLED)	0.783784	0.783784	0.784937	0.782007	0.851852
Ideación suicida (NIVIDEASUIC)	0.983108	0.983108	0.981454	0.412876	0.993127
Resiliencia (NIVRESIL)	0.739865	0.739865	0.742872	0.738902	0.798658
Satisfacción con la vida (NIVSATVIDA)	0.841216	0.841216	0.836561	0.834497	0.607143
Recursos psicológicos (NIVRECPSIC)	0.844595	0.844595	0.845162	0.844531	0.824324

Resultados curva ROC-AUC

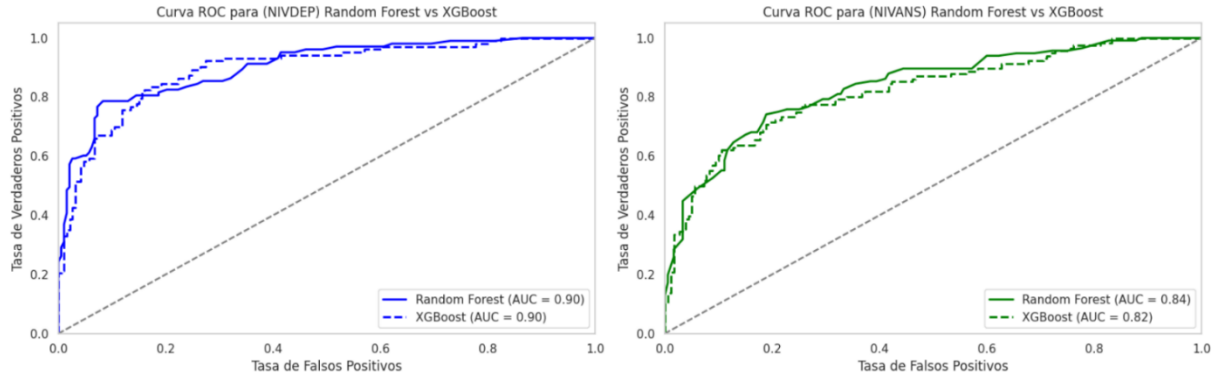


Ilustración 17. Curva Roc modelo Random Forest y XGBoost para depresión y ansiedad. Fuente propia.

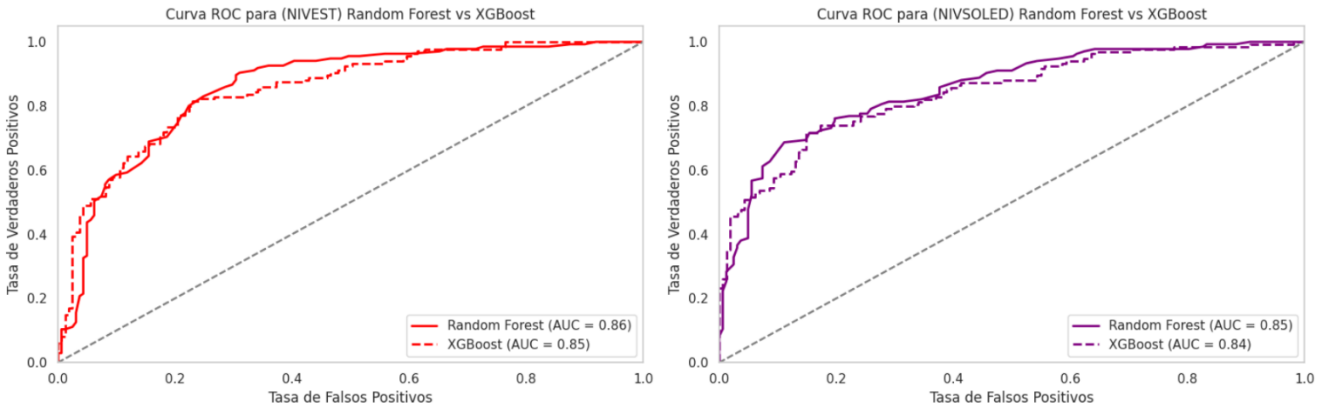


Ilustración 18. Curva Roc modelo Random Forest y XGBoost para estrés y soledad. Fuente propia.

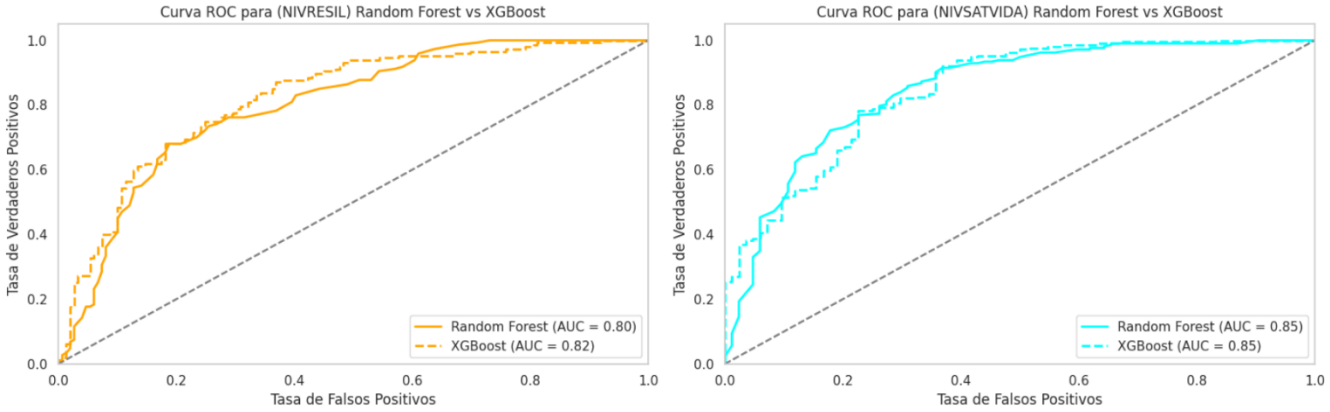


Ilustración 19. Curva Roc modelo Random Forest y XGBoost para resiliencia y satisfacción con la vida. Fuente propia.

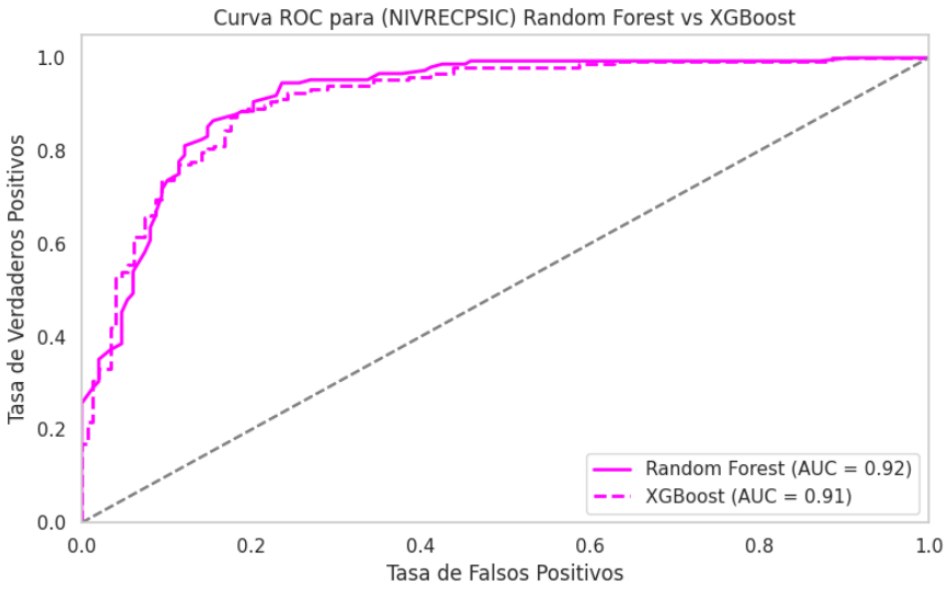


Ilustración 20. Curva Roc modelo Random Forest y XGBoost para recursos psicológicos. Fuente propia.

Resultados matrices de confusión

Adicional a las métricas de Accuracy, F1 score, Recall, y Curva Roc, también se aplicó la técnica de matriz de confusión, en este caso para el modelo XGBoost que fue el seleccionado para llevar a cabo esta investigación. En la ilustración 26, 27 se presentan los resultados.

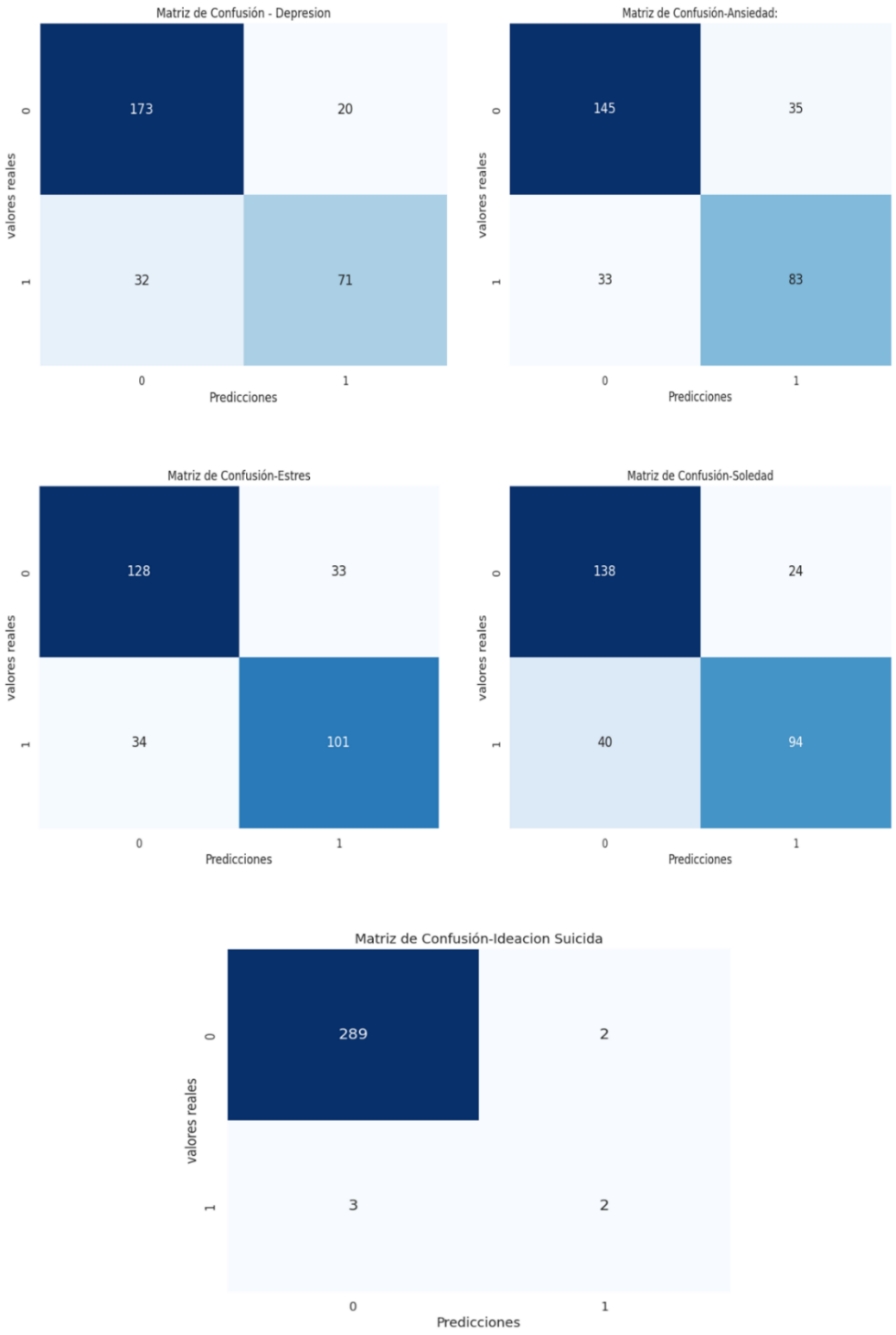


Ilustración 21. Resultados matrices de confusión para el modelo XGBoost – Indicadores Negativos. Fuente Propia.

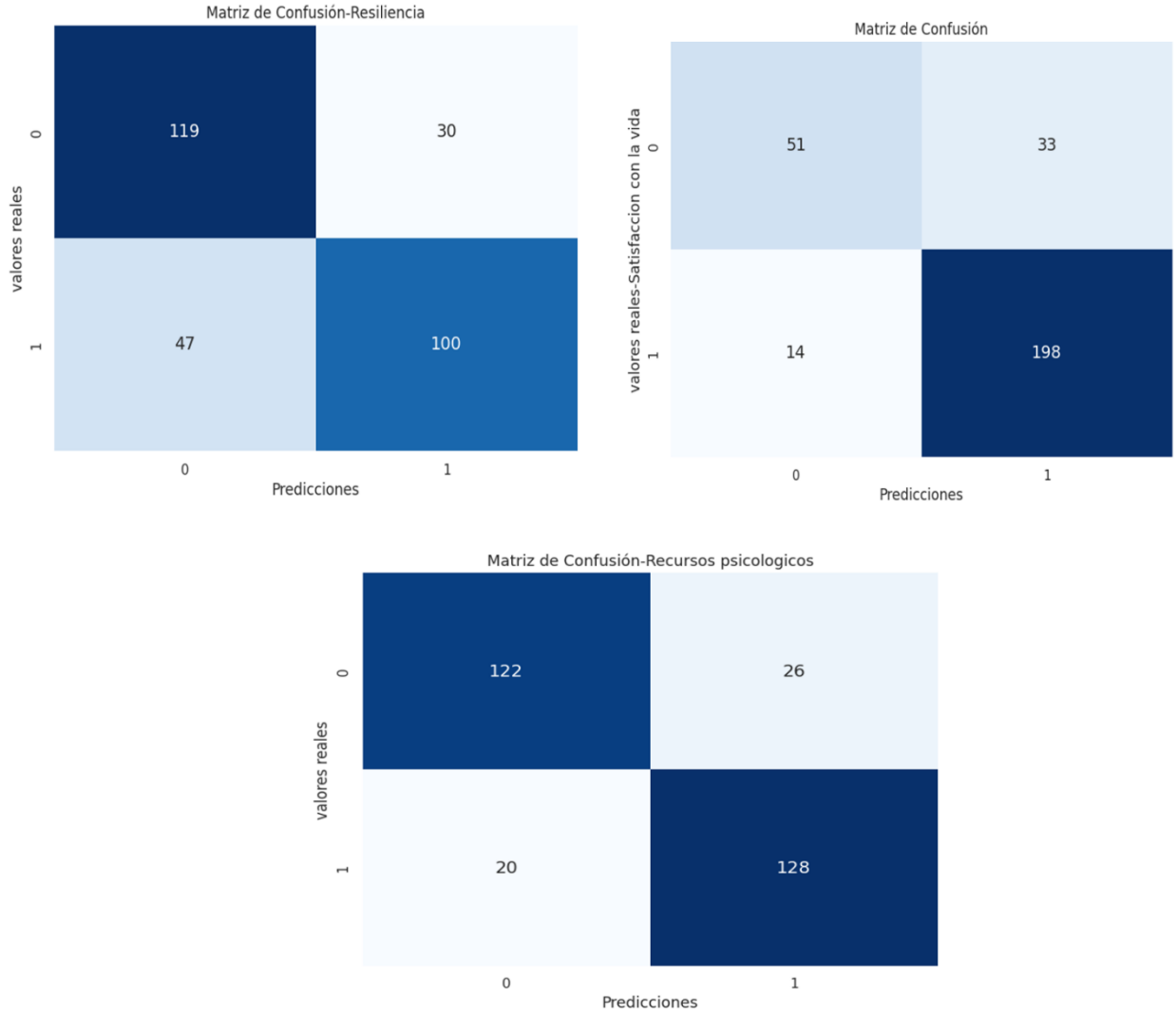


Ilustración 22. Resultados matrices de confusión para el modelo XGBoost – Indicadores Positivos. Fuente Propia.

Validación cruzada

Según [29] el proceso de evaluación por validación cruzada permite obtener una estimación del rendimiento de los modelos y además una medida de cuán precisa es dicha estimación con la desviación estándar, se desarrolla dividiendo en conjuntos de entrenamiento y de validación más pequeños, luego entrena y valida reiteradamente en dichos subconjuntos y compara los resultados. Este proceso se desarrolla por medio las funciones `cross_val_score`, `cross_validate` de Scikit-Learn.

Tabla 17.
Validación Cruzada Modelo Default. Fuente Propia.

Modelo	Accuracy Mean	Accuracy Std	Recall Mean	Recall Std	Precision Mean	Precision Std	F1-Score Mean	F1-Score Std	Specificity Mean	Specificity Std
NIVDEP	0.820	0.032	0.820	0.032	0.819	0.034	0.817	0.032	0.892	0.041
NIVANS	0.770	0.033	0.770	0.033	0.770	0.032	0.769	0.032	0.826	0.043
NIVEST	0.796	0.018	0.796	0.018	0.798	0.019	0.795	0.019	0.840	0.040
NIVSOLED	0.740	0.029	0.740	0.029	0.740	0.029	0.737	0.031	0.818	0.027
NIVRESIL	0.692	0.048	0.692	0.048	0.694	0.048	0.690	0.049	0.679	0.087
NIVSATVIDA	0.819	0.035	0.819	0.035	0.811	0.038	0.810	0.037	0.525	0.098
NIVRECPSIC	0.822	0.033	0.822	0.033	0.824	0.032	0.822	0.033	0.795	0.043

Respecto al modelo de predicción de la depresión, los niveles de precisión y recall indican que el modelo es efectivo para predecir la variable, la alta especificidad indica que el modelo identifica correctamente los casos negativos.

En cuanto al modelo para ansiedad, se puede ver un modelo robusto con alta especificidad, indicando buen manejo de los falsos positivos, de forma similar el modelo de estrés muestra buenos puntajes en precisión y especificidad, lo que muestra buen manejo de verdaderos positivos y negativos.

En resiliencia las métricas muestran un nivel inferior respecto a los otros modelos, sin ser ineficaz, el modelo tiene mayor dificultad para manejar los positivos como los negativos en comparación con los demás modelos.

En el modelo de satisfacción con la vida hay alta precisión y recall, pero la especificidad es cercana al 50% lo que indica que puede presentar problemas para manejar los casos negativos.

En recursos psicológicos el modelo presenta altas métricas, consolidándose como un modelo sólido para la predicción de positivos y negativos.

5.4.1 INTERPRETACIÓN DE RESULTADOS

Se entienden los resultados del modelo para extraer conocimientos y conclusiones significativas que puedan contribuir al objetivo de la investigación. Esto puede incluir la identificación de variables importantes, análisis de sensibilidad o visualización de resultados.

Para llevar a cabo la interpretación de los modelos creados se aplicaron 2 métodos feature_importances_ y shap. A continuación, se presentan los resultados obtenidos en cada uno.

Análisis de importancia relativa de las variables predictoras.

Por medio de la función feature_importances_ de Scikit-Learn, se puede evaluar la importancia relativa de las variables predictoras en el ejercicio de clasificación. Aplicando este modelo se obtuvieron los siguientes resultados:

Indicadores Negativos para el modelo de XGBoost

Depresión

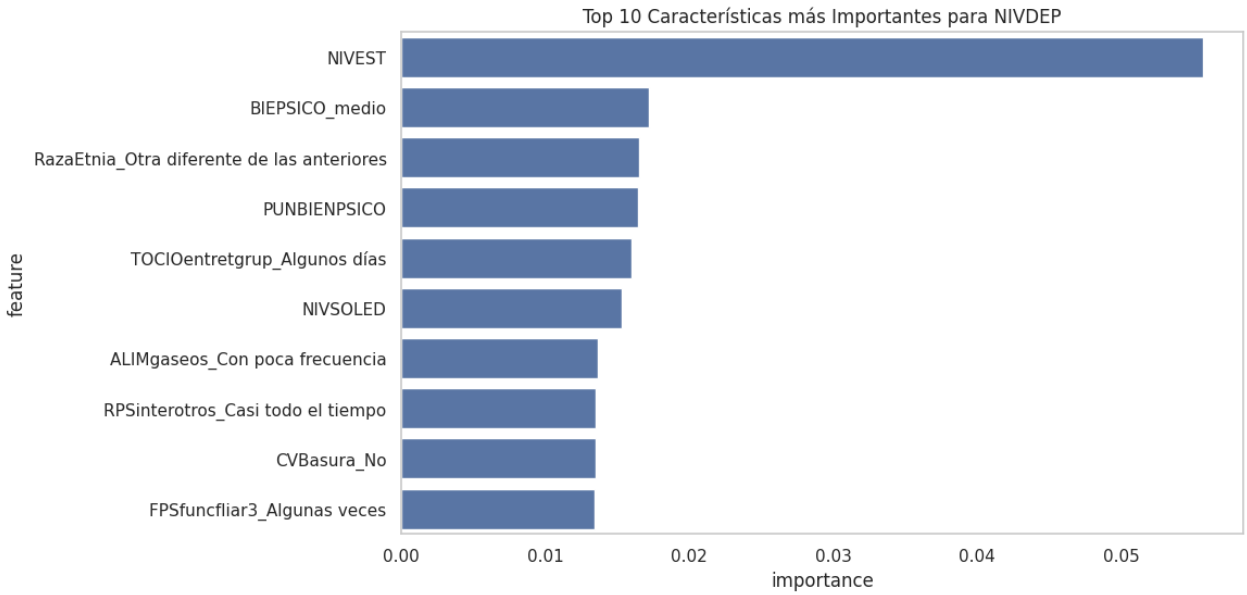


Ilustración 23. Top 10 características más relevantes para la depresión. Fuente Propia

Para la variable depresión, se tienen importancias relativas de variables de carácter individual como el estrés (NIVEST), el bienestar psicológico (BIEPSICO), la raza o etnia (RazaEtnia), el puntaje obtenido en bienestar psicológico (PUNBIENPSICO), el tiempo de ocio (TOCIOentregrup), la soledad (NIVSOLED), la frecuencia de consumo de alimentos (ALIMgaseos), los recursos psicológicos (RPSinterotros); a nivel externo hay dos variables presentes que son las condiciones de basura en el barrio donde vive (CVBasura) y el funcionamiento familiar (FPSfuncfliar3).

Ansiedad

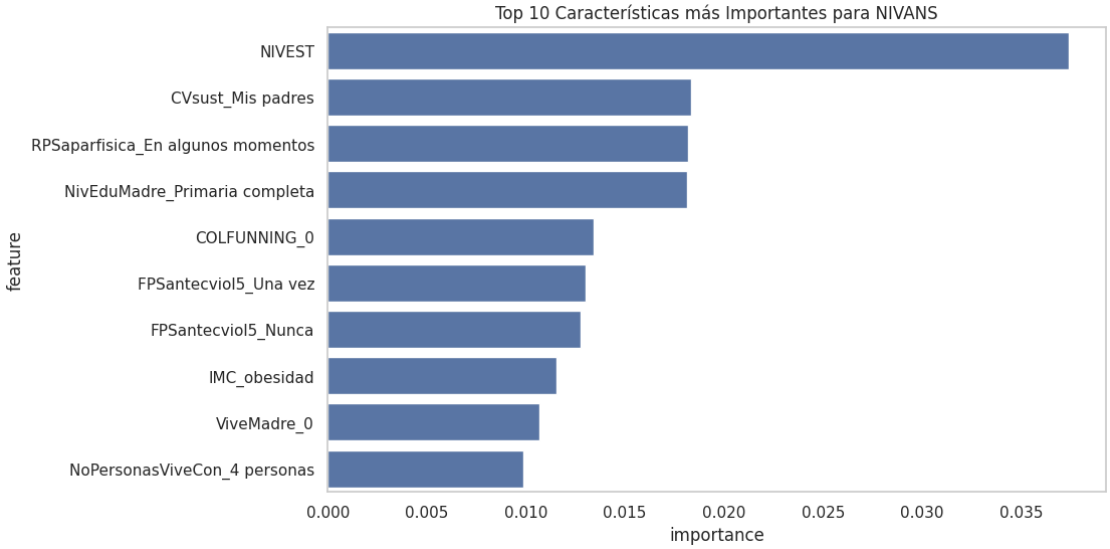


Ilustración 24. Top 10 características más relevantes para la ansiedad. Fuente Propia.

Para la variable ansiedad se puede observar que las variables que más afectan la predicción son relativos a factores individuales como el nivel de estrés (NIVEST), recursos psicológicos (RPSaparfisica), antecedentes de violencia y de abuso sexual (FPSantecviol5) y el índice de masa corporal (IMC). Pero también aparecen factores externos al individuo, por ejemplo, quién sustenta el hogar (CVsust), el nivel de educación de la madre (NivEduMadre), la cantidad de funciones asignadas en el trabajo administrativo (COLFUNNING), si vive con la madre (ViveMadre) y la cantidad de personas con las que vive (NoPersonasViveCon).

Estrés

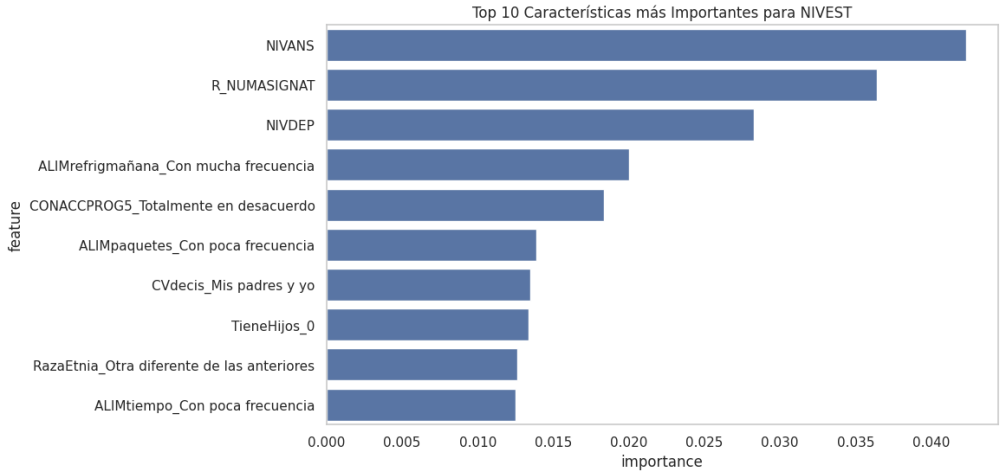


Ilustración 25. Top 10 características más relevantes para el estrés. Fuente Propia.

Respecto al estrés las variables de carácter individual que aparecen en el top 10 son la ansiedad (NIVANS), el número de asignaturas a cargo en el caso de los profesores (R_NUMASIGNAT), la

depresión (NIVDEP), respecto a salud y hábitos de salud emerge la alimentación (ALIMrefrigmañana, ALIMpaquetes y ALIMtiempo), en cuestiones estructurales se presenta el conocimiento y acceso a programas de salud y bienestar de la universidad (CONACCPROG5), en condiciones de vida surge quien toma las decisiones del hogar (CVdecis), en la factores individuales sociodemográficos el hecho de si tiene hijos (Tienehijos) y la raza (RazaEtnia).

Soledad

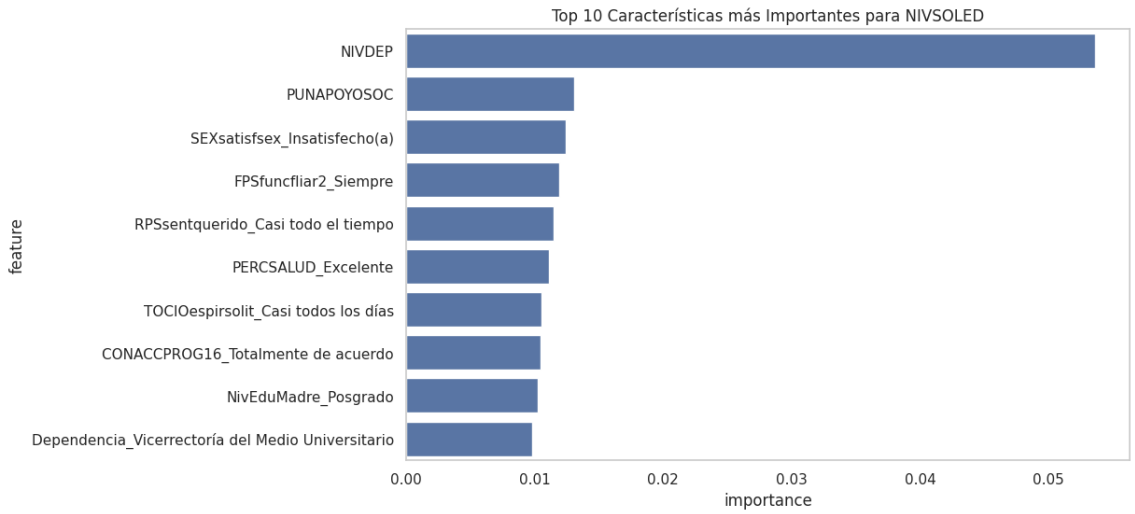


Ilustración 26. Top 10 características más relevantes para la soledad. Fuente Propia.

Respecto a la soledad, las variables importantes muestran que la depresión es un factor que impacta de mayor manera que las demás, en el resto se puede ver que aparecen variables relacionadas con factores individuales psicosociales como el puntaje obtenido en apoyo social (PUNAPOYOSOC), el funcionamiento familiar (PSfuncfliar), en salud y hábitos de salud se presenta la satisfacción sexual (SEXsatisfsex), la percepción general de salud (PERCSALUD) y el tiempo de ocio (TOCIOespirsolit), en salud mental emergen variables como recursos psicológicos (RPSsentquerido), por último aparecen el nivel educativo de la madre (NivEduMadre) y la dependencia donde se desempeña la labor (Dependencia).

Ideación suicida

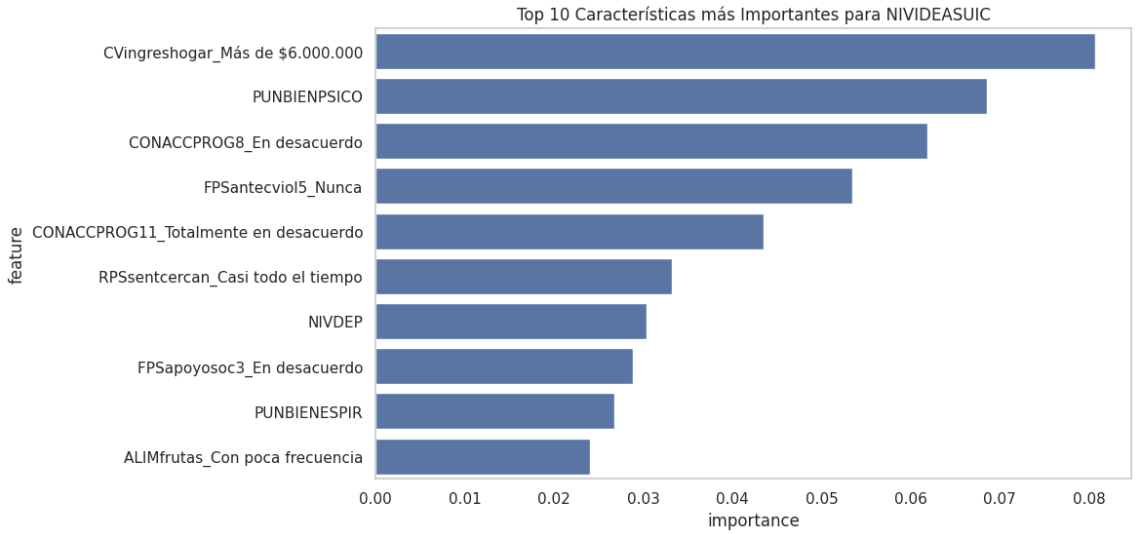


Ilustración 27. Top 10 características más relevantes para ideación suicida. Fuente Propia.

En ideación suicida aparecen como importantes las variables de condiciones de vida reflejada en los ingresos del hogar (CVingreshogar), el puntaje de bienestar psicológico, en variables estructurales surge el conocimiento y acceso a programas de salud y bienestar en la universidad (CONACCPROG8 Y CONACCPROG11), en factores individuales aparecen los antecedentes de violencia y abuso sexual (FPSantecviol5) el apoyo social (FPSapoyosoc3), en salud mental están los recursos psicológicos (RPSsentcercan) y la depresión (NIVDEP). El bienestar espiritual (PUNBIENESPIR) y la alimentación (ALIMfrutas) también se presentan con importancia relativa.

Indicadores Positivos para el modelo de XGBoost

Resiliencia

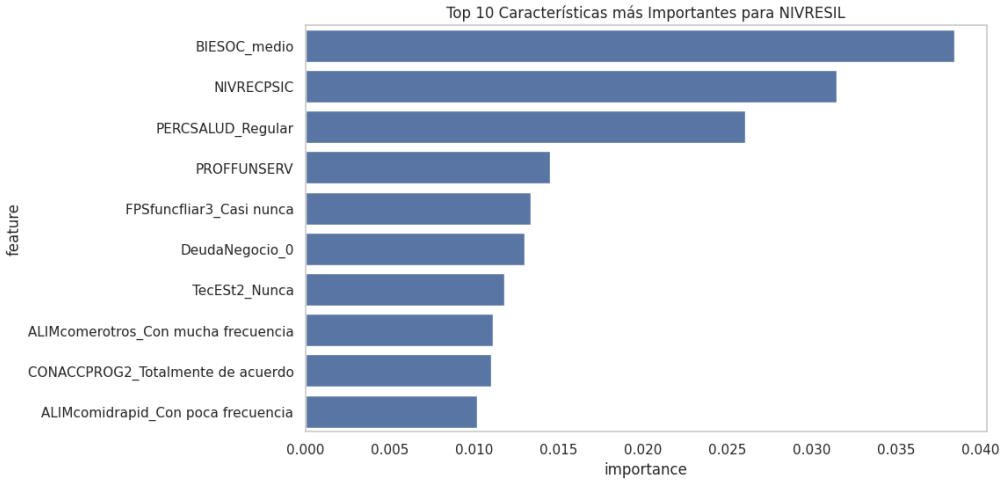


Ilustración 28. Top 10 características más relevantes para la resiliencia. Fuente Propia.

Respecto a la resiliencia se puede observar la importancia relativa de variables de bienestar social (BIESOC), en salud mental los recursos psicológicos (NIVRECPSIC), en salud está la percepción de la propia salud (PERCSALUD), en los profesores aparece la variable de funciones docentes (PROFFUNSERV), respecto a los factores individuales psicosociales surge el funcionamiento familiar (FPSfuncfliar), en condiciones aparece el endeudamiento como variable importante, la alimentación (ALIMcomerotros y ALIMcomidrapid) y el conocimiento y acceso a programas de salud y bienestar en la universidad (CONACCPROG2).

Satisfacción con la vida

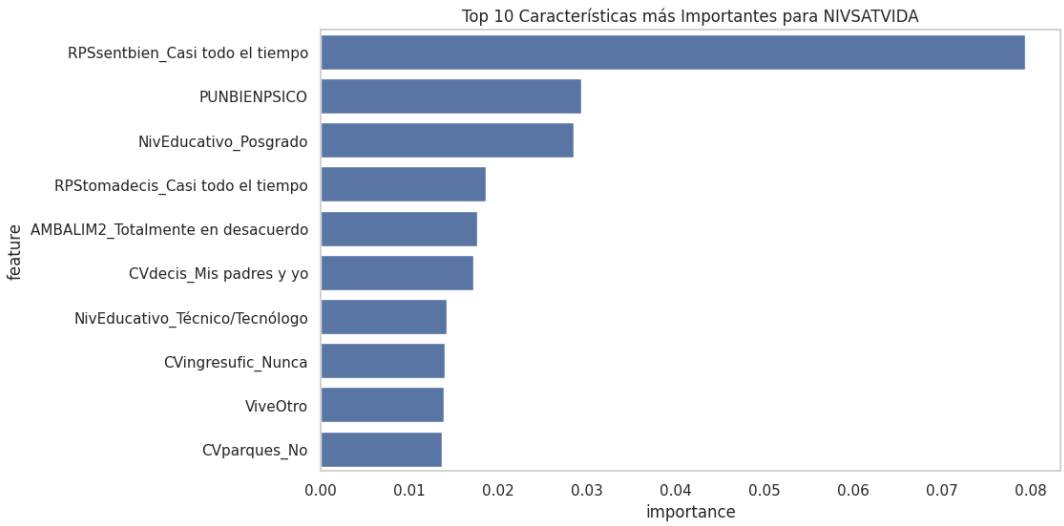


Ilustración 29. Top 10 características más relevantes para la satisfacción con la vida. Fuente Propia.

En satisfacción con la vida se tiene como variable importante la relacionada con salud mental, específicamente con los recursos psicológicos (RPSsentbien), luego el puntaje en el bienestar psicológico (PUNBIENPSICO), el nivel educativo (NivEducativo) también surge dentro de los factores individuales sociodemográficos junto con la composición familiar (ViveOtro). Emergen también cuestiones estructurales como el ambiente alimentario universitario (AMBALIM2), en condiciones de vida aparecen variables relacionadas con la toma de decisiones del hogar (CVdecis), suficiencia de los ingresos en el hogar (CVingresufic) y el acceso a espacios en las condiciones de vivienda (CVparques).

Recursos psicológicos

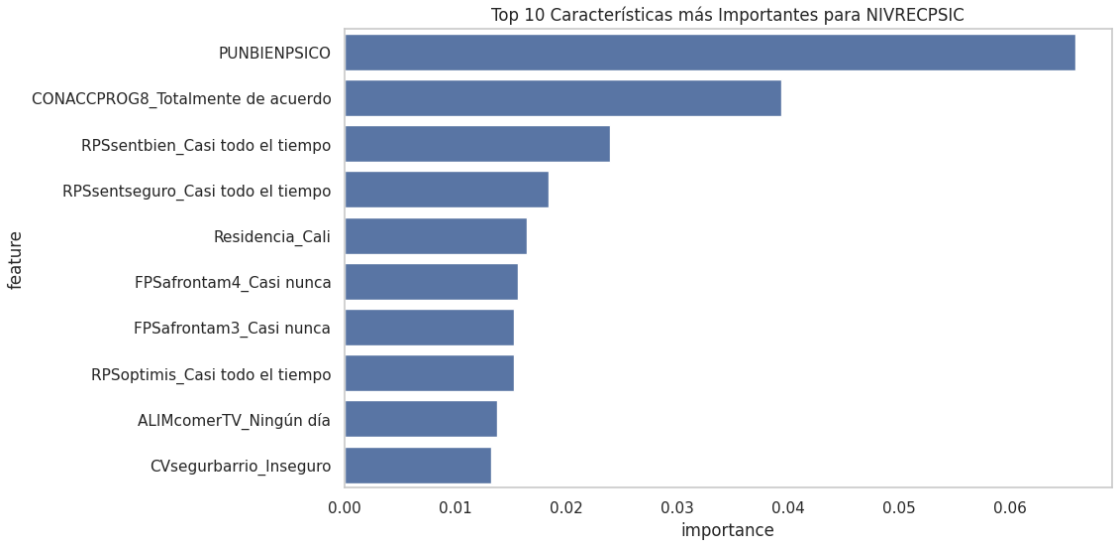


Ilustración 30. Top 10 características más relevantes para los recursos psicológicos. Fuente Propia.

En recursos psicológicos la variable que relativamente aporta más importancia a la predicción es el puntaje obtenido en bienestar psicológico (PUNBIENPSICO), seguido del conocimiento y acceso a programas de salud y bienestar en la universidad (CONACCPROG8), se destacan los recursos psicológicos (RPSsentbien, RPSsentseguro y RPSoptimismo), también surgen factores individuales como la residencia (Residencia), factores individuales psicosociales como el afrontamiento (FPSafrontam4 y FPSafrontam3), la alimentación (ALIMcomerTv) y las condiciones de vida respecto al barrio (CVsegurbarrio)

Análisis Shap

En esta sección, se presenta la interpretación de los resultados obtenidos mediante el análisis de SHAP. Esta librería se usó para analizar individualmente cada predicción del modelo, lo que permite una comprensión detallada de cómo cada característica contribuye a las decisiones del modelo para cada individuo.

La función "shap.plots.waterfall" permite visualizar las contribuciones de las características en las predicciones del modelo. Al generar un gráfico de cascada, cada barra representa una característica y muestra cómo esa característica en particular impacta la predicción final en comparación con una línea de base. Esto ofrece una visión detallada de la contribución de cada característica a la salida del modelo.

A continuación, se presentan los gráficos de cascada para el primer individuo del conjunto de prueba, que muestran la contribución de cada característica a la predicción del modelo.

Indicadores Negativos

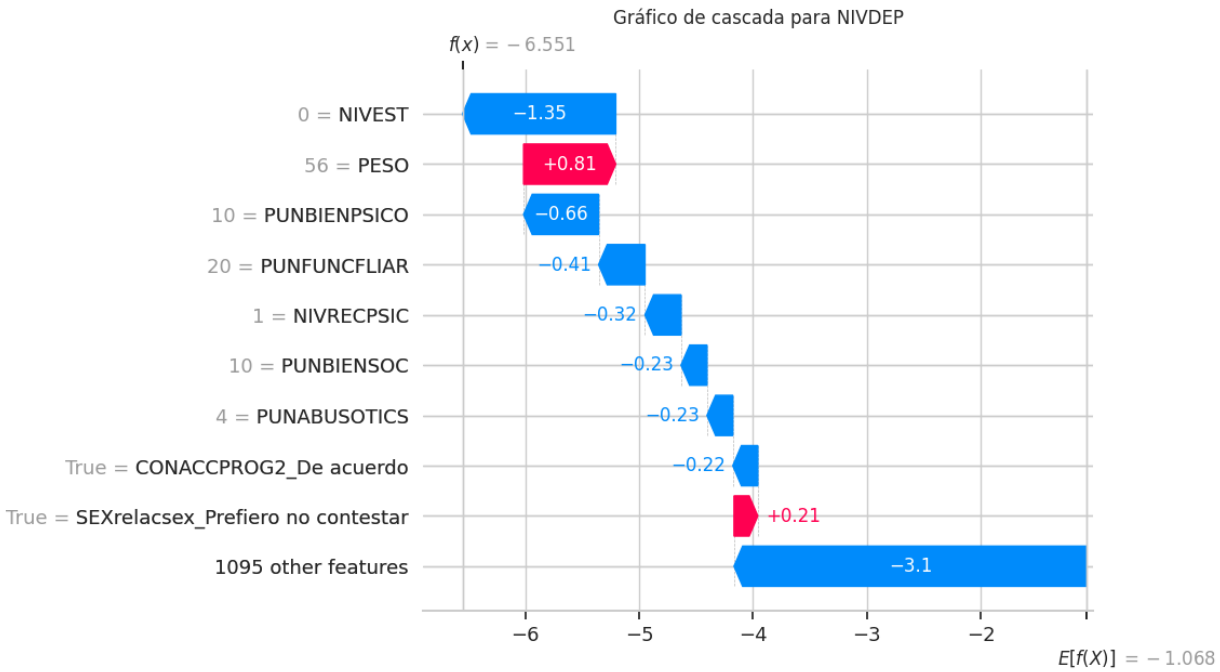


Ilustración 31. Grafica shap para la variable nivel de depresión. Fuente Propia.

La visualización de SHAP revela que el peso corporal y las relaciones sexuales del individuo son las características que más contribuyen positivamente a la probabilidad de depresión. En contraste, el nivel de estrés, bienestar psicológico, funcionamiento familiar, recursos psicológicos, bienestar social, uso abusivo de TICs, y el conocimiento y acceso a programas de salud y bienestar en la universidad tienen una influencia negativa.

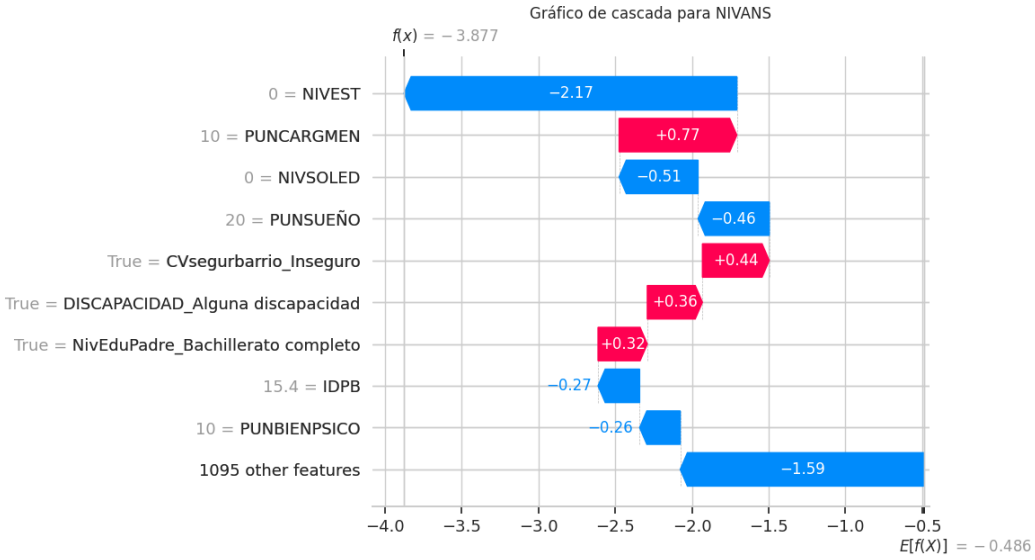


Ilustración 32. Grafica shap para la variable nivel de ansiedad. Fuente Propia.

La visualización de SHAP revela que la carga mental, seguridad del barrio, discapacidad y el nivel de educación del padre son las características que más contribuyen positivamente a la probabilidad de ansiedad. En contraste, el nivel de estrés, nivel de soledad, calidad del sueño, el desconocimiento de programas de Salud y bienestar y el bienestar psicológico tienen una influencia negativa.

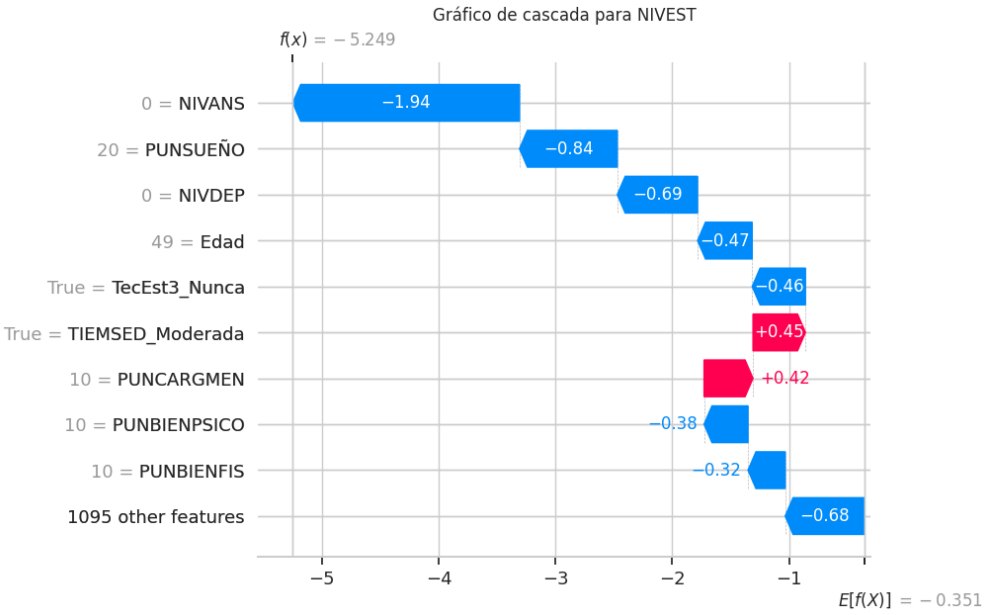


Ilustración 33. Grafica shap para la variable nivel de estrés. Fuente Propia.

La visualización de SHAP revela que el sedentarismo y la carga mental son las características que más contribuyen positivamente a la probabilidad de estrés. En contraste, el nivel de ansiedad, calidad del sueño, nivel de depresión, edad, el estrés tecnológico, bienestar psicológico y el bienestar físico tienen una influencia negativa.

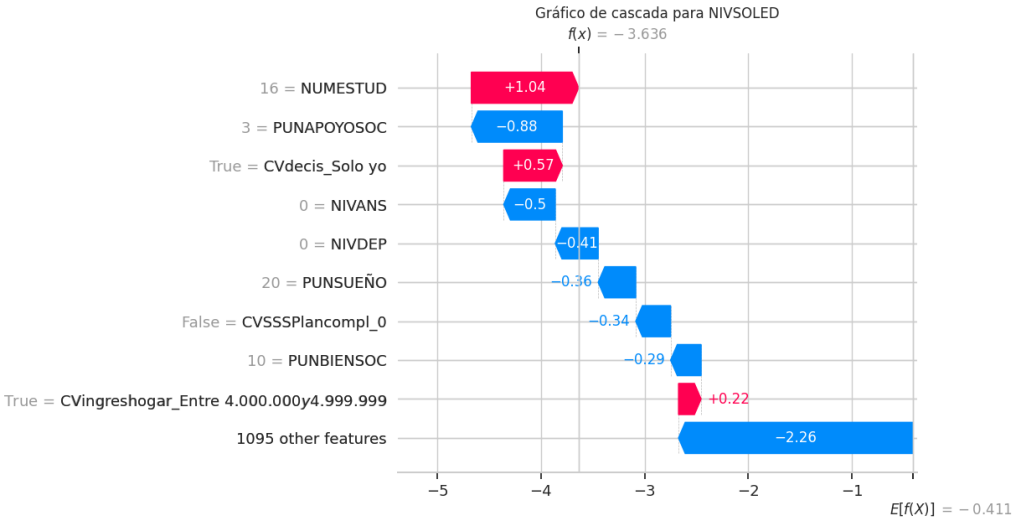


Ilustración 34. Grafica shap para la variable nivel de soledad. Fuente Propia.

La visualización de SHAP revela que el apoyo Social, nivel de ansiedad, nivel de depresión, calidad de sueño, seguridad social en salud, bienestar social son las características que más contribuyen positivamente a la probabilidad de soledad. En contraste, el número de estudiantes, toma de las decisiones en el hogar e ingresos del hogar tienen una influencia negativa.

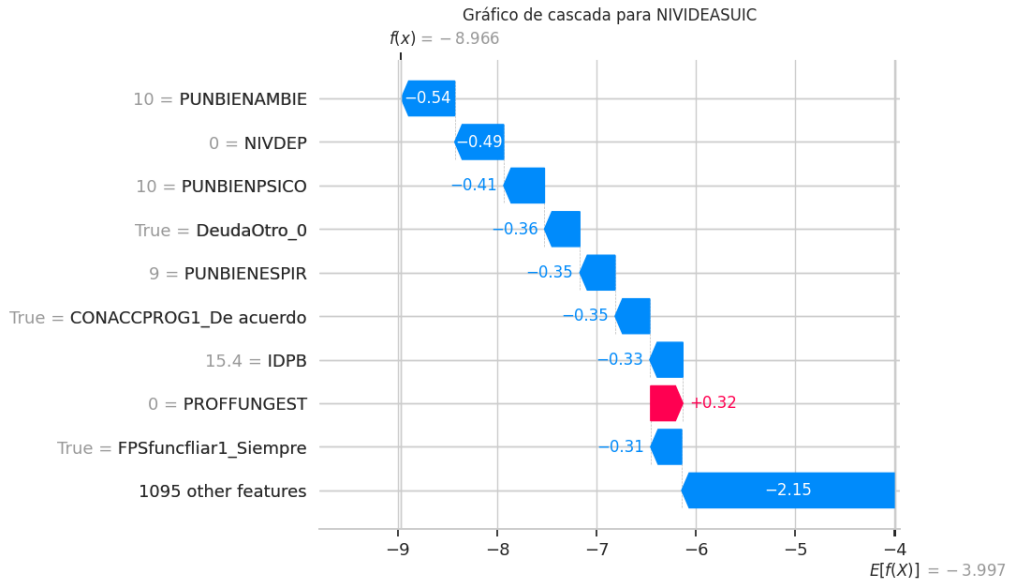


Ilustración 35. Grafica shap para la variable nivel de ideación suicida. Fuente Propia.

La visualización de SHAP revela que la función docente de gestión es la característica que más contribuye positivamente a la probabilidad de la ideación suicida. En contraste el bienestar ambiental, Nivel de depresión, bienestar psicológico, endeudamiento, bienestar espiritual, desconocimiento y acceso a programas de salud y bienestar y el funcionamiento familiar tienen una influencia negativa.

Indicadores Positivos

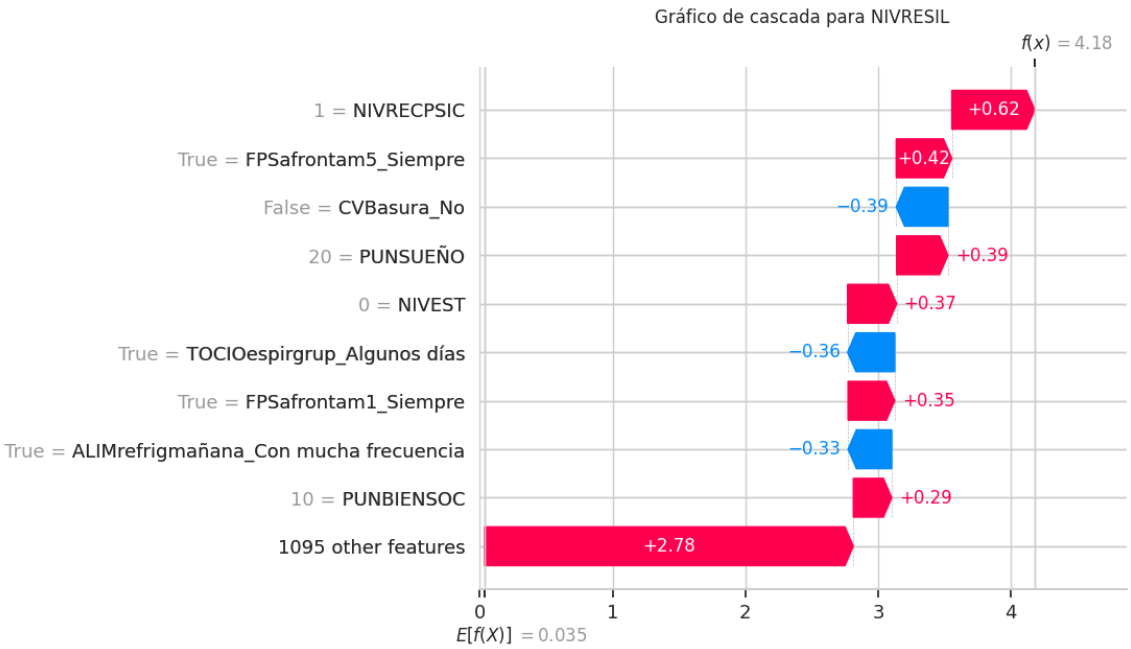


Ilustración 36. Grafica shap para la variable nivel de resiliencia. Fuente Propia.

La visualización de SHAP revela que el nivel de recursos psicológicos, el afrontamiento, calidad del sueño, nivel de estrés, Factores individuales psicosociales y Bienestar social son las características que más contribuyen positivamente a la probabilidad de resiliencia. En contraste, condiciones del barrio, tiempos de alimentación y Tiempo de ocio tienen una influencia negativa.

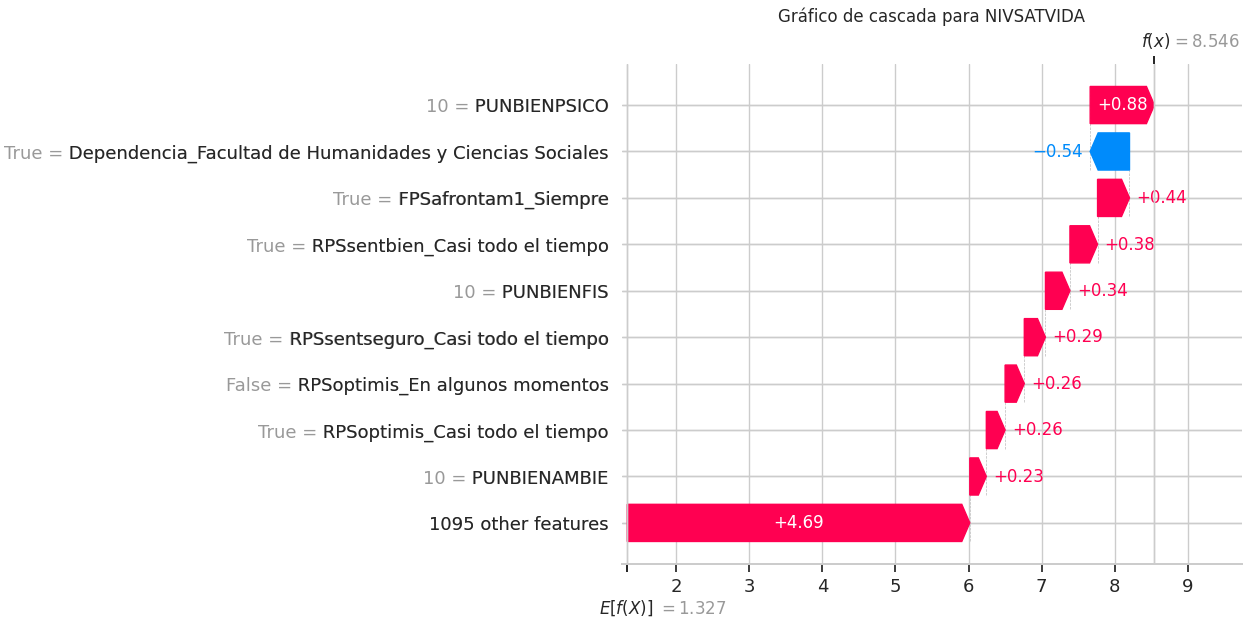


Ilustración 37. Grafica shap para la variable nivel de satisfacción con la vida. Fuente Propia.

La visualización de SHAP revela que el bienestar psicológico, el afrontamiento1, sentirse bien (recursos psicológicos), bienestar físico, sentirse seguro, sentirse optimista, el bienestar ambiental son las características que más contribuyen positivamente a la probabilidad de satisfacción con la vida. En contraste, la Dependencia facultad de humanidades y ciencias sociales tienen una influencia negativa.

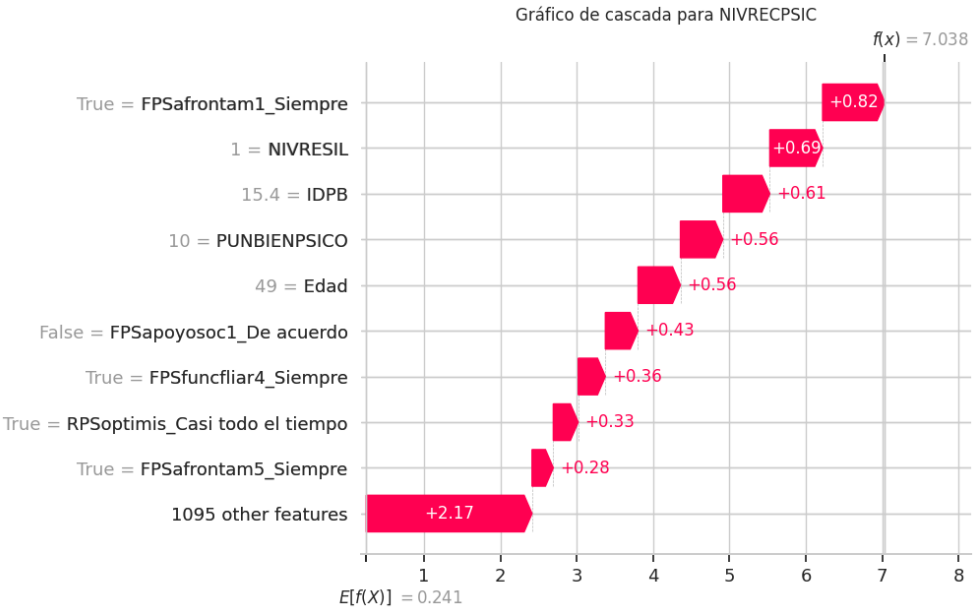


Ilustración 38. Grafica shap para la variable nivel de recursos psicológicos. Fuente Propia.

La visualización de SHAP revela que el afrontamiento1, nivel de resiliencia, desconocimiento de programas de salud y bienestar, bienestar psicológico, edad, apoyo social, funcionamiento familiar, sentirse optimista son las características que más contribuyen positivamente a la probabilidad de recursos psicológicos.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. CONCLUSIONES

De acuerdo con los resultados obtenidos durante el desarrollo del proyecto, se determinó que el mejor modelo para la predicción de variables de salud mental en la universidad fue XGBoost en términos de F1 score con los datos desbalanceados. A pesar de eso, los otros modelos de Machine Learning creados también obtuvieron un buen rendimiento. El modelo XGBoost logró un rendimiento superior al 72% en cada una de las variables, y que varía en un rango de 73% a 82%.

Los modelos de clasificación Random Forest y XGBoost obtuvieron desempeños similares en relación con la predicción de los indicadores de requerir o no atención respecto a las variables depresión, ansiedad, estrés y soledad, de la misma manera sucedió para la predicción de componentes positivos o indicadores de alerta en las variables Resiliencia, Satisfacción con la vida y recursos psicológicos.

Las variables predictoras mostraron relación con la posibilidad de activar alertas frente a la presencia o no de alguna de las variables. Estos resultados pueden ser utilizados para la formulación de proyectos que analicen y focalicen dichas variables predictoras.

La utilidad de esta investigación para la universidad es que ahora cuenta con una herramienta de Machine Learning capaz de predecir variables de salud mental, tanto factores positivos como factores negativos, de igual forma cuenta con un análisis de relevancia donde se presentaran las variables que más influyen en cada uno de los modelos creados. Lo que les permitirá tomar decisiones informadas sobre temas relacionados con la salud mental generando planes de intervención y prevención.

Estos resultados muestran la importancia de usar modelos de Machine Learning para predecir variables de salud mental a partir del análisis de bases de datos de la Universidad.

Los modelos Random Forest y XGBoost, con base en los Decision Trees, muestran resultados que resaltan su importancia para este tipo de estudios, alternativamente a los que se habían revisado y en los cuales se utilizaban otros tipos modelos.

El análisis de relevancia realizado en esta investigación por medio de las herramientas Feature_importances y shap, revelo que algunas de las características más influyentes en diversas variables de salud mental, como los ingresos del hogar (CVingreshogar), el funcionamiento familiar (FPSfuncfliar), la dependencia donde se desempeña la labor (Dependencia), apoyo social (FPSapoyosoc3), composición familiar (ViveOtro), carga mental, seguridad del barrio y demás, no son únicamente propias del individuo, sino que están relacionadas con su entorno social. Este descubrimiento puede tener importantes implicaciones para las estrategias de intervención en psicología.

En cuanto a la interpretabilidad de los modelos utilizados, tanto Random Forest como XGBoost pueden presentar cierto grado de dificultad, sin embargo, usando procesos como el análisis de variables importantes y SHAP, se puede mejorar en cierto grado la interpretación de los resultados al aportar herramientas para entender el impacto de cada variable predictora en las predicciones realizadas por el modelo.

Teniendo en cuenta lo anterior, se puede verificar el cumplimiento de los objetivos del proyecto, puesto que se desarrolló un módulo de procesamiento y análisis de datos con el Dataset obtenido, generando una base de datos que permite experimentar tanto con los modelos utilizados en este proyecto como otro tipo de modelos para diferentes propósitos.

Sumado a lo anterior, se logró crear modelos de predicción usando técnicas de aprendizaje automático, estos modelos mostraron un buen rendimiento para hacer predicciones con los datos, y con los cuales se podrían generar futuros análisis de importancia relativa a diferentes instancias usando técnicas Shap para obtener un panorama de cómo afectan las variables predictoras a varios casos específicos, como podrían ser los casos en los que se presente alguna condición de salud mental de interés para próximos investigadores.

Respecto a los procesos de validación de resultados estadísticos para definir el rendimiento, se encontró que en general los procesos de validación usados aportaron significativamente a decisiones importantes para el proyecto como la elección del modelo a trabajar como predictor. La teoría encontrada para implementar estos modelos es amplia y permitió validar los rendimientos de cada modelo de diferentes maneras.

6.2. TRABAJOS FUTUROS

Para complementar este proyecto se podrían separar las bases de datos para hacer un análisis específico en población docente, considerando variables adicionales como las calificaciones obtenidas, el promedio de reprobación por curso, entre otras que pueden obtenerse del cruce con otras bases de datos.

Al predecir la variable Ideación suicida se encontró un conjunto de datos con un desbalance que dificultó el proceso de predicción, sin embargo, por la naturaleza de la misma variable se podría considerar un análisis usando modelos que permitan hacer la predicción manteniendo la proporción de los datos iniciales.

Se propone emplear técnicas comunes de búsqueda de hiperparámetros, como Grid Search o Random Search, con el propósito de mejorar el rendimiento del modelo. Implementar estas técnicas de sintonización podría mejorar significativamente la precisión y la capacidad de generalización del modelo, logrando así una predicción más precisa de variables en salud mental.

Con los modelos ejecutados y el análisis de importancia relativa de las variables predictoras se podrían generar proyectos de prevención que contrasten los resultados de estos modelos con la

teoría psicológica actual y puedan dar pie a mejorar la comprensión del fenómeno dentro de la universidad.

REFERENCIAS BIBLIOGRÁFICAS

[1] Varela Maria Teresa, Cepeda Iván Leonardo, Uribe Ana Marcela, Cadavid Natalia, y Botero Jimena, "Resumen ejecutivo proyecto Salud y bienestar PUJ", 2023.

[2] J. J. Del Pozo-Antúnez, A. Ariza-Montes, F. Fernández-Navarro, y H. Molina-Sánchez, "Effect of a Job Demand-Control-Social Support Model on Accounting Professionals' Health Perception", *Int J Environ Res Public Health*, vol. 15, núm. 11, nov. 2018, doi: 10.3390/IJERPH15112437.

[3] Organizacion Mundial de la salud, "Salud mental: fortalecer nuestra respuesta". Consultado: el 11 de abril de 2024. [En línea]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/mental-health-strengthening-our-response

[4] world health organization, "Mental health". Consultado: el 6 de junio de 2024. [En línea]. Disponible en: https://www.who.int/health-topics/mental-health#tab=tab_2

[5] H. A. Whiteford *et al.*, "Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010", *Lancet*, vol. 382, núm. 9904, pp. 1575–1586, 2013, doi: 10.1016/S0140-6736(13)61611-6.

[6] W. de Paula, G. S. Breguez, E. L. Machado, y A. L. Meireles, "Prevalence of anxiety, depression, and suicidal ideation symptoms among university students: a systematic review / Prevalência de sintomas ansiedade, depressão e ideação suicida entre estudantes universitários: uma revisão sistemática", *Brazilian Journal of Health Review*, vol. 3, núm. 4, pp. 8739–8756, jul. 2020, doi: 10.34119/BJHRV3N4-119.

[7] T. Teismann, T. Forkmann, J. Brailovskaia, P. Siegmann, H. Glaesmer, y J. Margraf, "Positive mental health moderates the association between depression and suicide ideation: A longitudinal study", *Int J Clin Health Psychol*, vol. 18, núm. 1, p. 1, ene. 2018, doi: 10.1016/J.IJCHP.2017.08.001.

[8] C. Goetz, R. Bavaresco, R. Kunst, y J. Barbosa, "Industrial intelligence in the care of workers' mental health: A review of status and challenges", *Int J Ind Ergon*, vol. 87, p. 103234, ene. 2022, doi: 10.1016/J.ERGON.2021.103234.

[9] A. B. R. Shatte, D. M. Hutchinson, y S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications", *Psychol Med*, vol. 49, núm. 9, pp. 1426–1448, jul. 2019, doi: 10.1017/S0033291719000151.

[10] X. Wang *et al.*, "Prediction of Mental Health in Medical Workers During COVID-19 Based on Machine Learning", *Front Public Health*, vol. 9, p. 697850, sep. 2021, doi: 10.3389/FPUBH.2021.697850/FULL.

[11] OMS, *Constitución de la Organización Mundial de la Salud*. 2014. Consultado: el 9 de junio de 2023. [En línea]. Disponible en: https://apps.who.int/gb/bd/PDF/bd48/basic-documents-48th-edition-sp.pdf?ua=1#page=7,%202014

[12] M. R. Paredes, V. Apaolaza, C. Fernandez-Robin, P. Hartmann, y D. Yañez-Martinez, "The impact of the COVID-19 pandemic on subjective mental well-being: The interplay of perceived threat, future anxiety and resilience", *Pers Individ Dif*, vol. 170, feb. 2021, doi: 10.1016/J.PAID.2020.110455.

[13] J. Allen, R. Balfour, R. Bell, y M. Marmot, "Social determinants of mental health", *Int Rev Psychiatry*, vol. 26, núm. 4, pp. 392–407, 2014, doi: 10.3109/09540261.2014.928270.

[14] C. Haquin F, M. Larraguibel Q, y J. Cabezas A, "Factores protectores y de riesgo en salud mental en niños y adolescentes de la ciudad de Calama", *Rev Chil Pediatr*, vol. 75, núm. 5, pp. 425–433, sep. 2004, doi: 10.4067/S0370-41062004000500003.

[15] American Psychiatric Association, *Diagnostic And Statistical Manual Of Mental Disorders, Fifth Edition*, 5a ed. 2013.

[16] S. E. Hobfoll, "Social and Psychological Resources and Adaptation", *Review of General Psychology*, vol. 6, núm. 4, pp. 307–324, 2002, doi: 10.1037/1089-2680.6.4.307.

[17] D. V. Gutiérrez, I. C. García, M. Z. Gutiérrez, R. M. Gilchrist, M. C. R. Torres, y A. C. Montecino, "SOCIAL DETERMINANTS OF HEALTH AND LIFESTYLES IN ADULT POPULATION CONCEPCIÓN, CHILE", *Ciencia y enfermería*, vol. 20, núm. 1, pp. 61–74, 2014, doi: 10.4067/S0717-95532014000100006.

[18] N. Cornejo Espinoza *et al.*, "Asociación entre determinantes sociales y salud mental: efecto de la doble carga laboral y doméstica", *MediSur*, vol. 20, núm. 5, pp. 907–916, 2022, Consultado: el 11 de abril de 2024. [En línea]. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1727-897X2022000500907&lng=es&nrm=iso&tlng=es

[19] L. Knifton y G. Inglis, "Poverty and mental health: policy, practice and research implications", *BJPsych Bull*, vol. 44, núm. 5, pp. 193–196, oct. 2020, doi: 10.1192/BJB.2020.78.

[20] D. Kestel, "La salud mental en el lugar de trabajo: orientaciones a nivel mundial". Consultado: el 11 de abril de 2024. [En línea]. Disponible en: https://www.who.int/es/news-room/commentaries/detail/mental-health-in-the-workplace

[21] R. Allande-Cussó, J. Jesús García-Iglesias, J. Fagundo-Rivera, Y. Navarro-Abal, J. Antonio Climent-Rodríguez, y J. Gómez-Salgado, "SALUD MENTAL Y TRASTORNOS MENTALES EN LOS LUGARES DE TRABAJO", vol. 96, ene. 2022, Consultado: el 11 de abril de 2024. [En línea]. Disponible en: https://rabida.uhu.es/dspace/bitstream/handle/10272/20924/RESP%20Trast%20mentales%20trabajo%202022%20Q4.pdf?sequence=2

[22] A. Calvo Soto, "Salud mental en la actualidad", *Revista Colombiana de Salud Ocupacional*, vol. 10, núm. 1, pp. 6457–6457, mar. 2021, doi: 10.18041/2322-634x/rcso.1.2020.6457.

[23] A. J. Oswald, E. Proto, y D. Sgroi, "Happiness and Productivity", *https://doi.org/10.1086/681096*, vol. 33, núm. 4, pp. 789–822, oct. 2015, doi: 10.1086/681096.

[24] J. M. Ortega Candel, "Big data, machine learning y data science en python", 2023.

[25] J. Bobadilla Sancho, "Machine Learning y Deep Learning", 2020.

[26] W. Rahman, "Como funciona la inteligencia artificial y el aprendizaje automatico", *AI and Machine Learning*, feb. 2020, doi: 10.4135/9789354791796.

[27] M. I. Jordan y T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects", *Science (1979)*, vol. 349, núm. 6245, pp. 255–260, jul. 2015, doi: 10.1126/SCIENCE.AAA8415.

[28] N. Silaparasetty, "An Overview of Machine Learning", *Machine Learning Concepts with Python and the Jupyter Notebook Environment*, pp. 21–39, 2020, doi: 10.1007/978-1-4842-5967-2_2.

[29] A. Gerón, "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems SECOND EDITION", 2019, Consultado: el 30 de enero de 2024. [En línea]. Disponible en: https://powerunit-ju.com/wp-content/uploads/2021/04/Aurelien-Geron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-OReilly-Media-2019.pdf

[30] V. R. Konasani y S. Kadre, *INTRODUCTION TO MACHINE LEARNING AND DEEP LEARNING*. McGraw-Hill Education, 2021. Consultado: el 9 de junio de 2023. [En línea]. Disponible en: https://www.accessengineeringlibrary.com/content/book/9781260462296/chapter/chapter1

[31] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, Consultado: el 26 de mayo de 2024. [En línea]. Disponible en: http://scikit-learn.sourceforge.net.

[32] R. E. B. Martínez *et al.*, "Árboles de decisión como herramienta en el diagnóstico médico", *Revista Médica de la Universidad Veracruzana*, vol. 9, núm. 2, pp. 19–24, 2009, Consultado: el 26 de mayo de 2024. [En línea]. Disponible en: https://www.medigraphic.com/cgi-

bin/new/resumen.cgi?IDARTICULO=27872&id2=
[33] C. Arana, “Modelos de Aprendizaje Automático Mediante Árboles de Decisión”, *CEMA Working Papers: Serie Documentos de Trabajo.*, 2021, Consultado: el 4 de mayo de 2024. [En línea]. Disponible en: https://ideas.repec.org/p/cem/doctra/778.html
[34] F. Provost, B. Lange, y T. Fawcett, “Data science for business : what you need to know about data mining and data-analytic thinking”, 2021, Consultado: el 4 de mayo de 2024. [En línea]. Disponible en: https://www.oreilly.com/library/view/data-science-for/9781663728265/
[35] J. D. Hunter, “Matplotlib: A 2D graphics environment”, *Comput Sci Eng*, vol. 9, núm. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.
[36] M. Á. Morales Hernández *et al.*, “Algoritmos de aprendizaje automático para la predicción del logro académico”, *RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, vol. 12, núm. 24, p. 341, abr. 2022, doi: 10.23913/RIDE.V12I24.1180.
[37] T. Chen y C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, ago. 2016, doi: 10.1145/2939672.2939785.
[38] H. T. T. Nguyen, L. H. Chen, V. S. Saravanarajan, y H. Q. Pham, “Using XG Boost and Random Forest Classifier Algorithms to Predict Student Behavior”, *2021 IEEE International Conference on Emerging Trends in Industry 4.0, ETI 4.0 2021*, 2021, doi: 10.1109/ETI4.051663.2021.9619217.
[39] J. J. Espinosa Zúñiga, “Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito”, *Ingeniería Investigación y Tecnología*, vol. 21, núm. 3, pp. 1–16, jul. 2020, doi: 10.22201/FI.25940732E.2020.21.3.022.
[40] V. M. Marchant Contreras, “UN MODELO PREDICTIVO INTERPRETABLE PARA LA ESTIMACIÓN DEL INGRESO MONETARIO DE CLIENTES BANCARIOS BASADO EN XGBOOST Y SHAP.”, oct. 2022, Consultado: el 30 de mayo de 2024. [En línea]. Disponible en: http://repositorio.udec.cl/jspui/bitstream/11594/10173/1/Tesis_Vicente_Marchant.pdf
[41] IBM, “Conceptos básicos de ayuda de CRISP-DM - Documentación de IBM”. Consultado: el 9 de junio de 2023. [En línea]. Disponible en: https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview
[42] T. Jain, A. Jain, P. S. Hada, H. Kumar, V. K. Verma, y A. Patni, “Machine Learning Techniques for Prediction of Mental Health”, *Proceedings of the 3rd International Conference on Inventive Research in Computing Applications, ICIRCA 2021*, pp. 1606–1613, sep. 2021, doi: 10.1109/ICIRCA51532.2021.9545061.
[43] S. O. Castrillon, L. M. G. Marín, H. H. J. Villegas, y C. C. P. Escobar, “Machine learning aplicado en la clasificación y predicción de la depresión: Una revisión sistemática”, *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, núm. Special Issue 47, pp. 363–375, 2022, Consultado: el 9 de junio de 2023. [En línea]. Disponible en: https://investigaciones-pure.udem.edu.co/es/publications/machine-learning-aplicado-en-la-clasificaci%C3%B3n-y-predicci%C3%B3n-de-la-
[44] Perez Alvarez Susana, “Sistema de estratificación y predicción de salud mental en trabajadores tecnológicos”. Consultado: el 9 de junio de 2023. [En línea]. Disponible en: https://openaccess.uoc.edu/bitstream/10609/99106/6/marrueTFM0619memoria.pdf
[45] K. Vaishnavi, U. N. Kamath, B. A. Rao, y N. V. S. Reddy, “Predicting Mental Health Illness using Machine Learning Algorithms”, *J Phys Conf Ser*, vol. 2161, núm. 1, p. 012021, ene. 2022, doi: 10.1088/1742-6596/2161/1/012021.
[46] Y. Zhou, W. Han, X. Yao, J. J. Xue, Z. Li, y Y. Li, “Developing a machine learning model for detecting depression, anxiety, and apathy in older adults with mild cognitive impairment using speech and facial expressions: A cross-sectional observational study”, *Int J Nurs Stud*, vol. 146, p. 104562, oct. 2023, doi: 10.1016/J.IJNURSTU.2023.104562.

[47] Pandas, “Pandas - Python Data Analysis Library”. Consultado: el 25 de mayo de 2024. [En línea]. Disponible en: https://pandas.pydata.org/
[48] C. R. Harris *et al.*, “Array programming with NumPy”, *Nature*, vol. 585, núm. 7825, pp. 357–362, sep. 2020, doi: 10.1038/S41586-020-2649-2.
[49] M. Waskom, “Seaborn: statistical data visualization”, *J Open Source Softw*, vol. 6, núm. 60, p. 3021, abr. 2021, doi: 10.21105/JOSS.03021.
[50] W. Mckinney, “Python for Data Analysis-O’Reilly Media (2012)”, *The effects of brief mindfulness intervention on acute pain experience: An examination of individual difference*, vol. 1, pp. 1689–1699, 2015, Consultado: el 26 de mayo de 2024. [En línea]. Disponible en: https://www.oreilly.com/library/view/python-for-data/9781449323592/
[51] G. James, D. Witten, T. Hastie, y R. Tibshirani, “An Introduction to Statistical Learning with Applications in R”, 2021, doi: 10.1007/978-1-0716-1418-1.
[52] T. Hastie, R. Tibshirani, y J. Friedman, “The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition”, 2009, doi: 10.1007/978-0-387-84858-7.
[53] SCIKIT-LEARN, “Feature importances with a forest of trees — scikit-learn 1.5.0 documentation”. Consultado: el 2 de junio de 2024. [En línea]. Disponible en: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html
[54] Shap, “Waterfall plot — SHAP”. Consultado: el 31 de mayo de 2024. [En línea]. Disponible en: https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/waterfall.html

ANEXOS

ANEXO 1. DICCIONARIO DE VARIABLES

Tabla 18.

Tabla Diccionario de Variables. Fuente PUJC.

Variables			BD colaboradores
Salud Mental	Indicadores negativos	Depresión	NIVDEP
		Ansiedad	NIVANS
		Estrés	NIVEST
		Soledad	NIVSOLED
		Ideación suicida	NIVIDEASUIC
	Indicadores positivos	Resiliencia	NIVRESIL
		Satisfacción con la vida	NIVSATVIDA
		Recursos psicológicos	**NIVRECPSIC** RPSoptimis RPSinterotros RPSresolprob RPSsentbien RPSsentcercan RPSsentseguro RPStomadecis RPSsentquerido RPSaparfisica
Salud y hábitos de salud	Percepción general de salud		PERCSALUD **PERCSALUDreco**
	Estado general de salud	IMC	IMC
		Enfermedades diagnosticadas	ENFERMEDAD ENFERMEDADCual CONDPSIQU CONDPSIQUCual
		Dolor	DOLOR DOLORCual
		Discapacidad	DISCAPACIDAD DISCAPfisica DISCAPauditiva DISCAPvisual DISCAPintelect DISCAPotra
	Actividad física		NIVACTFIS
	Sedentarismo		TIEMSED

	Sueño	Horas de sueño entre semana y fin de semana	HORASSUESEM HORASSUEFINDE
		Calidad del sueño	NIVSUEÑO
	Tiempo de ocio		TOCIOrelaj TOCIOartist TOCIOmusic TOCIOmanual TOCIOespirsolit TOCIOespirgrup TOCIOentretsolit TOCIOentretgrup
	Consumo de SPA		SPAalcohol SPAcigarrillo SPAvapeo SPAmarihuana SPAilegales
	Alimentación	Frecuencia de consumo de alimentos	ALIMfrutas ALIMverdur ALIMembutid ALIMpaquetes ALIMcomidrapid ALIMgaseos ALIMdulces ALIMcomidprepar
		Prácticas de alimentación	ALIMcafeteriaU ALIMcomerTV ALIMmaquinas ALIMtiempo ALIMhoras ALIMcomerotros
		Tiempos de alimentación	ALIMdesayuno ALIMrefrigmañana ALIMalmuerzo ALIMrefrigtarde ALIMcena ALIMdespucena
	Sexualidad	Ha tenido relaciones sexuales	SEXrelacsex
		Uso del preservativo	SEXpreserv
		Satisfacción sexual	SEXsatisfsex
		Orientación sexual	SEXorientacsex

	Uso abusivo de TICs		NIVABUSOTICS
Bienestar	Físico		BIEFISICO
	Psicológico		BIEPSICO
	Social		BIESOC
	Espiritual		BIENESPIR
	Ambiental		BIEAMBI
Variables Determinantes sociales de la salud y bienestar			
Factores individuales sociodemográficos	Sexo		Sexo
	Género		Género **GéneroReco**
	Nivel educativo		NivEducativo
	Edad		Edad RangoEdad
	Raza o etnia		RazaEtnia
	Estrato socioeconómico		Estratosoc NivelSocioec
	Estado civil		Estadocivil
	Procedencia		Nacioen
	Residencia		Residencia Residencia_valle Residencia_cauca Residencia_otro
	Zona residencia rural/urbana		ZonaResidencia
	Composición familiar	# de personas con las que vive	NoPersonasViveCon
		Tiene hijos/cuántos	TieneHijos CuantosHijos
		Con quién vive	ViveSolo VivePadre ViveMadre VivePareja ViveHermanos ViveAbuelo VivePrimos ViveSobrinos ViveFamiliares ViveConocidos ViveCompañeros ViveHijos ViveOtro

Factores individuales psicosociales	Afrontamiento		FPSafrontam1 FPSafrontam2 FPSafrontam3 FPSafrontam4 FPSafrontam5
	Apoyo social		FPSapoyosoc1 FPSapoyosoc2 FPSapoyosoc3 **APOYOSOC**
	Funcionamiento familiar		FPSfuncfliar1 FPSfuncfliar2 FPSfuncfliar3 FPSfuncfliar4 **FUNCFLIAR**
	Antecedentes de violencia y de abuso sexual		FPSantecviol1 FPSantecviol2 FPSantecviol3 FPSantecviol4 FPSantecviol5 FPSantecviol6
Condiciones de vida	Condiciones de la vivienda y del barrio	Acceso a servicios básicos	CVservpub Cvinternet
		Acceso a espacios	CVzonasocial CVcentrodepor CVtransp CVparques CVcentrossalud CVespaccomunit
		Seguridad del barrio	CVsegurbarrio
		Violencia en el barrio	Cvviolenbarrio
		Condiciones del barrio	CVInundac CVRuido CVBasura CVInvasespac Cvvías
	Transporte a la Universidad		CVTransVehicprop CVTransVehicompar CVTransPublicoMas CVTransPublicoTax CVTransBici CVTransCamina

	Situación económica del grupo familiar	Toma de las decisiones en el hogar	Cvdecis
		Sustento económico del hogar	CVsust
		Dependencia económica	Cvdepecon
		Suficiencia de los ingresos del hogar	CVingresufic
		Ingresos del hogar	Cvingreshogar
		Endeudamiento	**Endeudamiento** (# de deudas) DeudaVivienda DeudaVehiculo DeudaEducacion DeudaViaje DeudaNegocio DeudaOtro
	Seguridad social en salud		CVSSSsisben CVSSSEPS CVSSSPlancompl CVSSSMedPrep CVSSSOtro
	Seguridad alimentaria		CVSegAlimen1 CVSegAlimen2
Determinantes intermedios (colaboradores)			
Aspectos laborales en el contexto universitario	Condiciones laborales	Rol en la Universidad	ROLPrincipal ROLreco
		Dependencia de trabajo	Dependencia
		Tipo de contrato	Contrato
		Tiempo de vinculación	TiempoVincul
	Condiciones de trabajo	Carga mental	NIVCARGMEN
		Demandas cuantitativas	NIVDEMCUANT
		Demanda de la Jornada Laboral	NIVJORLAB
		Relaciones sociales en el trabajo	NIVRELSOCT

		Influencia del trabajo sobre el entorno extralaboral	NIVINFLTRA
		Demandas emocionales	NIVDEMEMOC
	Satisfacción laboral		SATISLABOR
	Tecno-estrés		NIVTECNOEST
	Trabaja y estudia		TrabajaEstudia HorasEstudia
	Deja de tomar vacaciones		VACACnotoma VACACrazones VACACrazonesotras
Aspectos laborales en el contexto universitario **específicos según el rol**	Tipo de contrato profesor		TIPOCONTRPROF
	Número de asignaturas a cargo (pre y posgrado)		NUMASIGNAT
	Número de estudiantes		NUMESTUD
	Funciones docentes	Solo profesores	PROFUNDOC PROFFUNINV PROFFUNGEST PROFFUNSERV PROFFUNFUERA PROFFUNFUERAdoc PROFFUNFUERAEdCon PROFFUNFUERACons PROFFUNFUERADirTG PROFFUNFUERAInv PROFFUNFUERAProd PROFFUNFUERARegCal PROFFUNFUERAOtro
	Funciones adicionales	Solo colaboradores administrativos	COLFUNDOC COLFUNINV COLFUNSERV COLFUNNING
		Solo directivos administrativos	DADMFUNDOC DADMFUNINV DADMFUNSERV

		Solo directivos académicos	DACAasign DACADestud DACADFUNDOC DACADFUNINV DACADFUNGEST DACADFUNSERV DACADFUNFUERA DACADFUNFUERAdoc DACADFUNFUERAEdCon DACADFUNFUERACons DACADFUNFUERADirTG DACADFUNFUERAInv DACADFUNFUERAProd DACADFUNFUERAOtro
	Relaciones sociales en el trabajo	Solo directivos administrativos	DADMRELSOC1 DADMRELSOC2 DADMRELSOC3 DADMRELSOC4 DADMRELSOC5 DADMRELSOC6 DADMRELSOC7 DADMRELSOC8 DADMRELSOC9 DADMRELSOC10 DADMRELSOC11
		Solo directivos académicos	DACARELSOC1 DACARELSOC2 DACARELSOC3 DACARELSOC4 DACARELSOC5 DACARELSOC6 DACARELSOC7 DACARELSOC8 DACARELSOC9 DACARELSOC10 DACARELSOC11

Determinantes estructurales	Conocimiento y acceso a programas de salud y bienestar en la universidad		**IDBP** (desconocimiento de programas de SyB) CONACCPROG1 CONACCPROG2 CONACCPROG3 CONACCPROG4 CONACCPROG5 CONACCPROG6 CONACCPROG7 CONACCPROG8 CONACCPROG9 CONACCPROG10 CONACCPROG11 CONACCPROG12 CONACCPROG13 CONACCPROG14 CONACCPROG15
	Ambiente alimentario universitario		AMBALIM1 AMBALIM2 AMBALIM3 AMBALIM4 AMBALIM5
Total preguntas			**293**

ANEXO 2. METODOLOGÍA DEL PROYECTO

Tabla 19.

Tabla Metodología. Fuente Propia.

OBJETIVO	FASE	ACTIVIDADES	ENTREGABLES	HERRAMIENTAS
. Desarrollar un módulo de procesamiento y análisis de datos con el	1. Comprensión de los datos	1.1 Describir los datos	Documentación	Base de datos otorgada por el equipo de investigación de la universidad -Encuesta bienestar salud
		1.2 Explorar los datos.		
		1.3 Verificar la calidad de los datos.		
	Análisis de datos	2.1 Selección de datos	Dataset preprocesado con variables de	Software de limpieza de datos,
		2.2 Limpieza de datos		

Dataset obtenido.		2.3 Construcción de datos	interés para el estudio aplicando métodos de limpieza de datos.	máquina virtual
		2.4 Integración de datos		
2. Desarrollar modelos de predicción a partir de técnicas de aprendizaje automático.	3. Modelado.	3.1 Selección de la técnica de modelado	Los modelos con la mayor precisión de todo el análisis realizado.	Tecnologías de machine Learning, máquina virtual
		3.2 Diseño del test		
		3.3 Construcción de los modelos		
		3.4 Evaluación de los modelos		
3. Validar los resultados estadísticos de los modelos.	4. Evaluación	4.1 Evaluación de los resultados	Documentación final con los resultados obtenidos en la investigación.	Dispositivo de cómputo, internet.
		4.2 Revisión del proceso		
		4.3 Determinación de los próximos pasos		
		4.4 Documentación del proyecto		

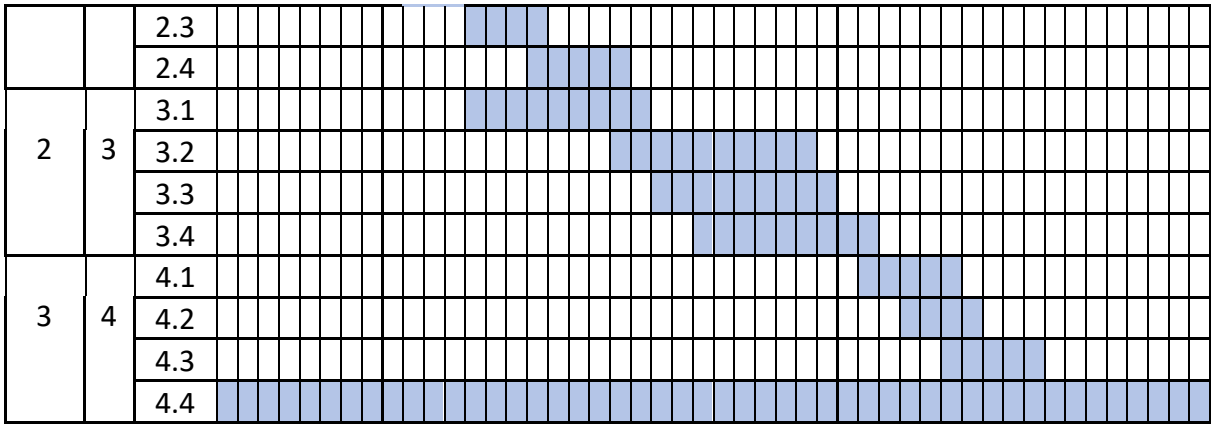
ANEXO 3. CRONOGRAMA DE ACTIVIDADES

Aquí se puede evidenciar como están distribuidas las actividades y fases correspondientes a los objetivos planteados por los diferentes meses

Tabla 20.
Tabla cronograma de actividades. Fuente Propia.

CRONOGRAMA DEL PROYECTO

Objetivos	Fases	Actividades	Mes 1				Mes 2				Mes 3				Mes 4				Mes 5				Mes 6				Mes 7				Mes 8				Mes 9				Mes 10				Mes 11				Mes 12			
			1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	1	1.1	■	■	■	■																																												
		1.2	■	■	■	■																																												
		1.3				■	■	■	■																																									
	2	2.1							■	■	■	■																																						
		2.2										■	■	■	■																																			



A continuación, se presenta una tabla con las variables pertenecientes a la base de datos colaboradores, clasificadas por tipo (cualitativa o cuantitativa), y cada una con una descripción.

ANEXO 4. TIPOS DE VARIABLES EN LA ENCUESTA DE SALUD MENTAL.

Tabla 21.
Tabla de Variables De Salud Mental. Fuente Propia.

Variables BD	**Definición**	**Tipo**	**Medición y unidad de medición**
NIVDEP	Depresión, Indicador negativo, Salud Mental.	Cualitativa	Ordinal
NIVANS	Ansiedad, Indicador negativo, Salud Mental.	Cualitativa	Ordinal
NIVEST	Estrés, Indicador negativo, Salud Mental.	Cualitativa	Ordinal
NIVSOLED	Soledad, Indicador negativo, Salud Mental.	Cualitativa	Ordinal
NIVIDEASUIC	Ideación suicida, Indicador negativo, Salud Mental.	Cualitativa	Nominal
NIVRESIL	Resiliencia, Indicador positivo, Salud Mental.	Cualitativa	Ordinal
NIVSATVIDA	Satisfacción con la vida, Indicador positivo, Salud Mental.	Cualitativa	Ordinal
NIVRECPSIC	Recursos psicológicos, Indicador positivo, Salud Mental.	Cualitativa	Ordinal
RPSoptimis	Recursos psicológicos, Indicador positivo, Salud Mental.	Cualitativa	Nominal
RPSinterotros	Recursos psicológicos, Indicador	Cualitativa	Nominal

	positivo, Salud Mental.		
RPSresolprob	Recursos psicológicos, Indicador positivo, Salud Mental.	Cualitativa	Nominal
RPSsentbien	Recursos psicológicos, Indicador positivo, Salud Mental.	Cualitativa	Nominal
RPSsentcercan	Recursos psicológicos, Indicador positivo, Salud Mental.	Cualitativa	Nominal
RPSsentseguro	Recursos psicológicos, Indicador positivo, Salud Mental.	Cualitativa	Nominal
RPStomadecis	Recursos psicológicos, Indicador positivo, Salud Mental.	Cualitativa	Nominal
RPSsentquerido	Recursos psicológicos, Indicador positivo, Salud Mental.	Cualitativa	Nominal
RPSaparfisica	Recursos psicológicos, Indicador positivo, Salud Mental.	Cualitativa	Nominal
PERCSALUD	Percepción general de salud, Salud y hábitos de salud	Cualitativa	Ordinal
PERCSALUDreco	Percepción general de salud, Salud y hábitos de salud	Cualitativa	Ordinal
IMC	IMC, Estado general de salud, Salud y hábitos de salud	Cualitativa	Ordinal
ENFERMEDAD	Enfermedades diagnosticadas, Estado general de salud, Salud y hábitos de salud	Cualitativa	Nominal
ENFERMEDADCual	Enfermedades diagnosticadas, Estado general de salud, Salud y hábitos de salud	Cualitativa	Nominal
CONDPSIQU	Enfermedades diagnosticadas, Estado general de salud, Salud y hábitos de salud	Cualitativa	Nominal
CONDPSIQUCual	Enfermedades diagnosticadas, Estado general de salud, Salud y hábitos de salud	Cualitativa	Nominal
DOLOR	Dolor, Estado general de salud, Salud y hábitos de salud	Cualitativa	Nominal
DOLORCual	Dolor, Estado general de salud, Salud y hábitos de salud	Cualitativa	Nominal
DISCAPACIDAD	Discapacidad, Estado general de salud, Salud y hábitos de salud	Cualitativa	Nominal
DISCAPfisica	Discapacidad, Estado general de salud, Salud y hábitos de salud	Cualitativa	Nominal
DISCAPauditiva	Discapacidad, Estado general de	Cualitativa	Nominal

	salud, Salud y hábitos de salud		
DISCAPvisual	Discapacidad, Estado general de salud, Salud y hábitos de salud	Cualitativa	Nominal
DISCAPintelect	Discapacidad, Estado general de salud, Salud y hábitos de salud	Cualitativa	Nominal
DISCAPotra	Discapacidad, Estado general de salud, Salud y hábitos de salud	Cualitativa	Nominal
NIVACTFIS	Actividad física, Salud y hábitos de salud	Cualitativa	Ordinal
TIEMSED	Sedentarismo, , Salud y hábitos de salud	Cualitativa	Ordinal
HORASSUESEM	Horas de sueño entre semana y fin de semana, Sueño, Salud y hábitos de salud	Cuantitativa	Discreta
HORASSUEFINDE	Horas de sueño entre semana y fin de semana, Sueño, Salud y hábitos de salud	Cuantitativa	Discreta
NIVSUEÑO	Calidad del sueño, Sueño, Salud y hábitos de salud	Cualitativa	Ordinal
TOCIOrelaj	Tiempo de ocio, Salud y hábitos de salud	Cualitativa	Ordinal
TOCIOartist	Tiempo de ocio, Salud y hábitos de salud	Cualitativa	Ordinal
TOCIOmusic	Tiempo de ocio, Salud y hábitos de salud	Cualitativa	Ordinal
TOCIOmanual	Tiempo de ocio, Salud y hábitos de salud	Cualitativa	Ordinal
TOCIOespirsolit	Tiempo de ocio, Salud y hábitos de salud	Cualitativa	Ordinal
TOCIOespirgrup	Tiempo de ocio, Salud y hábitos de salud	Cualitativa	Ordinal
TOCIOentretsolit	Tiempo de ocio, Salud y hábitos de salud	Cualitativa	Ordinal
TOCIOentretgrup	Tiempo de ocio, Salud y hábitos de salud	Cualitativa	Ordinal
SPAalcohol	Consumo de SPA, Salud y hábitos de salud	Cualitativa	Nominal
SPAcigarrillo	Consumo de SPA, Salud y hábitos de salud	Cualitativa	Nominal
SPAvapeo	Consumo de SPA, Salud y hábitos de salud	Cualitativa	Nominal
SPAmarihuana	Consumo de SPA, Salud y hábitos de	Cualitativa	Nominal

	salud		
SPAilegales	Consumo de SPA, Salud y hábitos de salud	Cualitativa	Nominal
ALIMfrutas	Frecuencia de consumo de alimentos, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMverdur	Frecuencia de consumo de alimentos, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMembutid	Frecuencia de consumo de alimentos, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMpaquetes	Frecuencia de consumo de alimentos, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMcomidrapid	Frecuencia de consumo de alimentos, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMgaseos	Frecuencia de consumo de alimentos, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMdulces	Frecuencia de consumo de alimentos, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMcomidprepar	Frecuencia de consumo de alimentos, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMcafeteriaU	Prácticas de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMcomerTV	Prácticas de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMmaquinas	Prácticas de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMtiempo	Prácticas de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMhoras	Prácticas de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal

ALIMcomerotros	Prácticas de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMdesayuno	Tiempos de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMrefrigmañana	Tiempos de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMalmuerzo	Tiempos de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMrefrigtarde	Tiempos de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMcena	Tiempos de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
ALIMdespucena	Tiempos de alimentación, Alimentación, Salud y hábitos de salud	Cualitativa	Ordinal
SEXrelacsex	Ha tenido relaciones sexuales,Sexualidad, Salud y hábitos de salud	Cualitativa	Nominal
SEXpreserv	Uso del preservativo, Sexualidad, Salud y hábitos de salud	Cualitativa	Ordinal
SEXsatisfsex	Satisfacción sexual, Sexualidad, Salud y hábitos de salud	Cualitativa	Ordinal
SEXorientacsex	Orientación sexual, Sexualidad, Salud y hábitos de salud	Cualitativa	Nominal
NIVABUSOTICS	Uso abusivo de TICs, Salud y hábitos de salud	Cualitativa	Nominal
BIEFISICO	Físico, Bienestar	Cualitativa	Ordinal
BIEPSICO	Psicológico, Bienestar	Cualitativa	Ordinal
BIESOC	Social, Bienestar	Cualitativa	Ordinal
BIENESPIR	Espiritual, Bienestar	Cualitativa	Ordinal
BIEAMBI	Ambiental, Bienestar	Cualitativa	Ordinal
Sexo	Sexo, Factores individuales sociodemográficos	Cualitativa	Nominal
Género	Género, Factores individuales sociodemográficos	Cualitativa	Nominal
GéneroReco	Género, Factores individuales	Cualitativa	Nominal

	sociodemográficos		
NivEducativo	Nivel educativo, Factores individuales sociodemográficos	Cualitativa	Ordinal
Edad	Edad, Factores individuales sociodemográficos	Cuantitativa	Discreta
RangoEdad	Edad, Factores individuales sociodemográficos	Cuantitativa	Discreta
RazaEtnia	Raza o etnia, Factores individuales sociodemográficos	Cualitativa	Nominal
Estratosoc	Estrato socioeconómico, Factores individuales sociodemográficos	Cualitativa	Ordinal
NivelSocioec	Estrato socioeconómico, Factores individuales sociodemográficos	Cualitativa	Ordinal
Estadocivil	Estado civil, Factores individuales sociodemográficos	Cualitativa	Nominal
Nacioen	Procedencia, Factores individuales sociodemográficos	Cualitativa	Nominal
Residencia	Residencia, Factores individuales sociodemográficos	Cualitativa	Nominal
Residencia_valle	Residencia, Factores individuales sociodemográficos	Cualitativa	Nominal
Residencia_cauca	Residencia, Factores individuales sociodemográficos	Cualitativa	Nominal
Residencia_otro	Residencia, Factores individuales sociodemográficos	Cualitativa	Nominal
ZonaResidencia	Zona residencia rural/urbana, Factores individuales sociodemográficos	Cualitativa	Ordinal
NoPersonasViveCon	# de personas con las que vive, Composición familiar, Factores individuales sociodemográficos	Cuantitativa	Discreta
TieneHijos	Tiene hijos/cuántos, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
CuantosHijos	Tiene hijos/cuántos, Composición familiar, Factores individuales sociodemográficos	Cuantitativa	Discreta
ViveSolo	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
VivePadre	Con quién vive, Composición familiar, Factores individuales	Cualitativa	Nominal

	sociodemográficos		
ViveMadre	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
VivePareja	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
ViveHermanos	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
ViveAbuelo	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
VivePrimos	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
ViveSobrinos	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
ViveFamiliares	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
ViveConocidos	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
ViveCompañeros	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
ViveHijos	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
ViveOtro	Con quién vive, Composición familiar, Factores individuales sociodemográficos	Cualitativa	Nominal
FPSafrontam1	Afrontamiento, Factores individuales psicosociales	Cualitativa	Ordinal
FPSafrontam2	Afrontamiento, Factores individuales psicosociales	Cualitativa	Ordinal
FPSafrontam3	Afrontamiento, Factores individuales psicosociales	Cualitativa	Ordinal
FPSafrontam4	Afrontamiento, Factores individuales psicosociales	Cualitativa	Ordinal

FPSafrontam5	Afrontamiento, Factores individuales psicosociales	Cualitativa	Ordinal
FPSapoyosoc1	Apoyo social, Factores individuales psicosociales	Cualitativa	Ordinal
FPSapoyosoc2	Apoyo social, Factores individuales psicosociales	Cualitativa	Ordinal
FPSapoyosoc3	Apoyo social, Factores individuales psicosociales	Cualitativa	Ordinal
APOYOSOC	Apoyo social, Factores individuales psicosociales	Cualitativa	Ordinal
FPSfuncfliar1	Funcionamiento familiar, Factores individuales psicosociales	Cualitativa	Ordinal
FPSfuncfliar2	Funcionamiento familiar, Factores individuales psicosociales	Cualitativa	Ordinal
FPSfuncfliar3	Funcionamiento familiar, Factores individuales psicosociales	Cualitativa	Ordinal
FPSfuncfliar4	Funcionamiento familiar, Factores individuales psicosociales	Cualitativa	Ordinal
FUNCFLIAR	Funcionamiento familiar, Factores individuales psicosociales	Cualitativa	Ordinal
FPSantecviol1	Antecedentes de violencia y de abuso sexual, Factores individuales psicosociales	Cualitativa	Ordinal
FPSantecviol2	Antecedentes de violencia y de abuso sexual, Factores individuales psicosociales	Cualitativa	Ordinal
FPSantecviol3	Antecedentes de violencia y de abuso sexual, Factores individuales psicosociales	Cualitativa	Ordinal
FPSantecviol4	Antecedentes de violencia y de abuso sexual, Factores individuales psicosociales	Cualitativa	Ordinal
FPSantecviol5	Antecedentes de violencia y de abuso sexual, Factores individuales psicosociales	Cualitativa	Ordinal
FPSantecviol6	Antecedentes de violencia y de abuso sexual, Factores individuales psicosociales	Cualitativa	Ordinal
CVservpub	Acceso a servicios básicos, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
Cvinternet	Acceso a servicios básicos,	Cualitativa	Nominal

	Condiciones de la vivienda y del barrio, Condiciones de vida		
CVzonasocial	Acceso a espacios, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
CVcentrodepor	Acceso a espacios, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
CVtransp	Acceso a espacios, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
CVparques	Acceso a espacios, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
CVcentrossalud	Acceso a espacios, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
CVespaccomunit	Acceso a espacios, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
CVsegurbarrio	Seguridad del barrio, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Ordinal
Cvviolenbarrio	Violencia en el barrio, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Ordinal
CVInundac	Condiciones del barrio, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
CVRuido	Condiciones del barrio, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
CVBasura	Condiciones del barrio, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
CVInvasespac	Condiciones del barrio, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
Cvvías	Condiciones del barrio, Condiciones de la vivienda y del barrio, Condiciones de vida	Cualitativa	Nominal
CVTransVehicprop	Transporte a la Universidad,	Cualitativa	Ordinal

	Condiciones de vida		
CVTransVehicompar	Transporte a la Universidad, Condiciones de vida	Cualitativa	Ordinal
CVTransPublicoMas	Transporte a la Universidad, Condiciones de vida	Cualitativa	Ordinal
CVTransPublicoTax	Transporte a la Universidad, Condiciones de vida	Cualitativa	Ordinal
CVTransBici	Transporte a la Universidad, Condiciones de vida	Cualitativa	Ordinal
CVTransCamina	Transporte a la Universidad, Condiciones de vida	Cualitativa	Ordinal
Cvdecis	Toma de las decisiones en el hogar, Situación económica del grupo familiar, Condiciones de vida	Cualitativa	Nominal
CVsust	Sustento económico del hogar, Situación económica del grupo familiar, Condiciones de vida	Cualitativa	Nominal
Cvdepecon	Dependencia económica, Situación económica del grupo familiar, Condiciones de vida	Cuantitativa	Discreta
CVingresufic	Suficiencia de los ingresos del hogar, Situación económica del grupo familiar, Condiciones de vida	Cualitativa	Ordinal
Cvingreshogar	Ingresos del hogar, Situación económica del grupo familiar, Condiciones de vida	Cuantitativa	Continua
Endeudamiento (# de deudas)	Endeudamiento, Situación económica del grupo familiar, Condiciones de vida	Cualitativa	Ordinal
DeudaVivienda	Endeudamiento, Situación económica del grupo familiar, Condiciones de vida	Cualitativa	Nominal
DeudaVehiculo	Endeudamiento, Situación económica del grupo familiar, Condiciones de vida	Cualitativa	Nominal
DeudaEducacion	Endeudamiento, Situación económica del grupo familiar, Condiciones de vida	Cualitativa	Nominal
DeudaViaje	Endeudamiento, Situación económica del grupo familiar, Condiciones de vida	Cualitativa	Nominal
DeudaNegocio	Endeudamiento, Situación	Cualitativa	Nominal

	económica del grupo familiar, Condiciones de vida		
DeudaOtro	Endeudamiento, Situación económica del grupo familiar, Condiciones de vida	Cualitativa	Nominal
CVSSSsisben	Seguridad social en salud, Condiciones de vida	Cualitativa	Nominal
CVSSSEPS	Seguridad social en salud, Condiciones de vida	Cualitativa	Nominal
CVSSSPlancompl	Seguridad social en salud, Condiciones de vida	Cualitativa	Nominal
CVSSSMedPrep	Seguridad social en salud, Condiciones de vida	Cualitativa	Nominal
CVSSSOtro	Seguridad social en salud, Condiciones de vida	Cualitativa	Nominal
CVSegAlimen1	Seguridad alimentaria, Condiciones de vida	Cualitativa	Nominal
CVSegAlimen2	Seguridad alimentaria, Condiciones de vida	Cualitativa	Nominal
ROLPrincipal	Rol en la Universidad, Condiciones laborales , Aspectos laborales en el contexto universitario	Cualitativa	Ordinal
ROLreco	Rol en la Universidad, Condiciones laborales , Aspectos laborales en el contexto universitario	Cualitativa	Ordinal
Dependencia	Dependencia de trabajo, Condiciones laborales , Aspectos laborales en el contexto universitario	Cualitativa	Nominal
Contrato	Tipo de contrato, Condiciones laborales , Aspectos laborales en el contexto universitario	Cualitativa	Nominal
TiempoVincul	Tiempo de vinculación, Condiciones laborales , Aspectos laborales en el contexto universitario	Cuantitativa	Discreta
NIVCARGMEN	Carga mental, Condiciones de trabajo, Aspectos laborales en el contexto universitario	Cualitativa	Ordinal
NIVDEMCUANT	Demandas cuantitativas, Condiciones de trabajo, Aspectos laborales en el contexto universitario	Cualitativa	Ordinal

NIVJORLAB	Demanda de la Jornada Laboral, Condiciones de trabajo, Aspectos laborales en el contexto universitario	Cualitativa	Ordinal
NIVRELSOCT	Relaciones sociales en el trabajo, Condiciones de trabajo, Aspectos laborales en el contexto universitario	Cualitativa	Ordinal
NIVINFLTRA	Influencia del trabajo sobre el entorno extralaboral, Condiciones de trabajo, Aspectos laborales en el contexto universitario	Cualitativa	Ordinal
NIVDEMEMOC	Demandas emocionales, Condiciones de trabajo, Aspectos laborales en el contexto universitario	Cualitativa	Ordinal
SATISLABOR	Satisfacción laboral, Aspectos laborales en el contexto universitario	Cualitativa	Ordinal
NIVTECNOEST	Tecno-estrés, Aspectos laborales en el contexto universitario	Cualitativa	Ordinal
TrabajaEstudia	Trabaja y estudia, Aspectos laborales en el contexto universitario	Cualitativa	Nominal
HorasEstudia	Trabaja y estudia, Aspectos laborales en el contexto universitario	Cuantitativa	Discreta
VACACnotoma	Deja de tomar vacaciones, Aspectos laborales en el contexto universitario	Cualitativa	Ordinal
VACACrazones	Deja de tomar vacaciones, Aspectos laborales en el contexto universitario	Cualitativa	Nominal
VACACrazonesotras	Deja de tomar vacaciones, Aspectos laborales en el contexto universitario	Cualitativa	Nominal
TIPOCONTRPROF	Tipo de contrato profesor, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
NUMASIGNAT	Número de asignaturas a cargo (pre y posgrado), Aspectos laborales en el contexto universitario específicos según el rol	Cuantitativa	Discreta
NUMESTUD	Número de estudiantes, Aspectos	Cuantitativa	Discreta

	laborales en el contexto universitario específicos según el rol		
PROFUNDOC	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cuantitativa	Discreta
PROFFUNINV	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cuantitativa	Discreta
PROFFUNGEST	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cuantitativa	Discreta
PROFFUNSERV	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cuantitativa	Discreta
PROFFUNFUERA	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
PROFFUNFUERAdoc	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol Cualitativa	Cualitativa	Nominal
PROFFUNFUERAEdCon	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
PROFFUNFUERACons	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
PROFFUNFUERADirTG	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
PROFFUNFUERAInv	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal

PROFFUNFUERAProd	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
PROFFUNFUERARegCal	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
PROFFUNFUERAOtro	Solo profesores, Funciones docentes, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
COLFUNDOC	Solo colaboradores administrativos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
COLFUNINV	Solo colaboradores administrativos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
COLFUNSERV	Solo colaboradores administrativos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
COLFUNNING	Solo colaboradores administrativos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
DADMFUNDOC	Solo directivos administrativos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
DADMFUNINV	Solo directivos administrativos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
DADMFUNSERV	Solo directivos administrativos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
DACAasign	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto	Cualitativa	Nominal

	universitario específicos según el rol		
DACADestud	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cuantitativa	Discreta
DACADFUNDOC	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cuantitativa	Discreta
DACADFUNINV	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cuantitativa	Discreta
DACADFUNGEST	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cuantitativa	Discreta
DACADFUNSERV	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cuantitativa	Discreta
DACADFUNFUERA	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
DACADFUNFUERAdoc	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
DACADFUNFUERAEdCon	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
DACADFUNFUERACons	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
DACADFUNFUERADirTG	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
DACADFUNFUERAInv	Solo directivos académicos,	Cualitativa	Nominal

	Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol		
DACADFUNFUERAProd	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario Cualitativa específicos según el rol	Cualitativa	Nominal
DACADFUNFUERAOtro	Solo directivos académicos, Funciones adicionales, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Nominal
DADMRELSOC1	Solo directivos administrativos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DADMRELSOC2	Solo directivos administrativos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DADMRELSOC3	Solo directivos administrativos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DADMRELSOC4	Solo directivos administrativos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DADMRELSOC5	Solo directivos administrativos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DADMRELSOC6	Solo directivos administrativos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DADMRELSOC7	Solo directivos administrativos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario Cualitativa específicos según el rol	Cualitativa	Ordinal
DADMRELSOC8	Solo directivos administrativos, Relaciones sociales en el trabajo,	Cualitativa	Ordinal

	Aspectos laborales en el contexto universitario específicos según el rol		
DADMRELSOC9	Solo directivos administrativos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DADMRELSOC10	Solo directivos administrativos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DADMRELSOC11	Solo directivos administrativos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DACARELSOC1	Solo directivos académicos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DACARELSOC2	Solo directivos académicos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DACARELSOC3	Solo directivos académicos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DACARELSOC4	Solo directivos académicos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DACARELSOC5	Solo directivos académicos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DACARELSOC6	Solo directivos académicos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DACARELSOC7	Solo directivos académicos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal

DACARELSOC8	Solo directivos académicos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DACARELSOC9	Solo directivos académicos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DACARELSOC10	Solo directivos académicos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
DACARELSOC11	Solo directivos académicos, Relaciones sociales en el trabajo, Aspectos laborales en el contexto universitario específicos según el rol	Cualitativa	Ordinal
IDBP (desconocimiento de programas de SyB)	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cuantitativa	Continua
CONACCPROG1	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG2	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG3	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG4	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG5	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG6	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes	Cualitativa	Ordinal

	estructurales		
CONACCPROG7	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG8	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG9	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG10	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG11	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG12	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG13	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG14	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
CONACCPROG15	Conocimiento y acceso a programas de salud y bienestar en la universidad, Determinantes estructurales	Cualitativa	Ordinal
AMBALIM1	Ambiente alimentario universitario, Determinantes estructurales	Cualitativa	Ordinal
AMBALIM2	Ambiente alimentario universitario, Determinantes estructurales	Cualitativa	Ordinal
AMBALIM3	Ambiente alimentario universitario,	Cualitativa	Ordinal

	Determinantes estructurales		
AMBALIM4	Ambiente alimentario universitario, Determinantes estructurales	Cualitativa	Ordinal
AMBALIM5	Ambiente alimentario universitario, Determinantes estructurales	Cualitativa	Ordinal

ANEXO 5. DESCRIPCIÓN DE LOS DATOS

https://drive.google.com/file/d/16CfH16W2pFin24B5kE-9OsXJJ-nG5vOM/view?usp=sharing

ANEXO 6. ANÁLISIS PARA LIMPIEZA DE DATOS

https://drive.google.com/file/d/1WJuXtwicag5gbp9gcz5yfxRiGGIfXKR3/view?usp=sharing

ANEXO 7. CONSTRUCCIÓN DE DATOS

https://drive.google.com/file/d/1HlcRV_M5TBycRKYh7A9zlscvgySy9OK8/view?usp=sharing