**TITULO PROYECTO:** Depression Risk Analysis

**INTEGRANTES:** Luis Alejandro Londoño Martínez, Simón Correa Marín

# 1. ENTENDIMIENTO DEL NEGOCIO

## 1.1 DESCRIPCIÓN DEL NEGOCIO

El proyecto **Depression Risk Analysis** se centra en analizar los factores de riesgo que pueden influir en la probabilidad de desarrollar depresión en adultos. Su objetivo principal es identificar patrones y correlaciones en las respuestas de una encuesta de salud mental para predecir el riesgo de depresión, basándose en datos de estilo de vida y demográficos de personas entre 18 y 60 años. A través de este análisis, se espera aportar a la comprensión de cómo variables como la satisfacción laboral, horas de estudio o trabajo, antecedentes familiares y otros factores de la vida cotidiana se relacionan con el riesgo de sufrir depresión.

Este proyecto se sitúa en el contexto de la investigación en salud mental**,** específicamente en la identificación de factores clave que afectan el bienestar psicológico en un entorno no clínico. La información obtenida se utilizará para desarrollar modelos predictivos de aprendizaje supervisado que podrían emplearse en la formulación de estrategias preventivas y en la toma de decisiones para intervenciones tempranas en el ámbito de la salud pública.

## 1.2 DESCRIPCIÓN DEL PROBLEMA

El problema específico que busca resolver este proyecto es identificar los factores cotidianos y demográficos que se asocian con el riesgo de depresión en adultos. A falta de evaluaciones clínicas y diagnósticos profesionales en este contexto, el desafío radica en emplear un modelo predictivo basado en datos de estilo de vida y satisfacción personal para predecir con precisión el riesgo de depresión en un grupo poblacional diverso. Esta herramienta permitirá entender mejor cómo ciertos factores pueden influir en la salud mental y proporcionar una base para la intervención temprana y la toma de decisiones informadas en materia de bienestar.

El problema que se busca abordar es la falta de herramientas predictivas para evaluar el riesgo de depresión en adultos con base en factores no clínicos y de estilo de vida.

## 1.3 OBJETIVOS DE LA MINERÍA

- Identificar los factores clave asociados al riesgo de depresión en adultos mediante un análisis exploratorio y predictivo de los datos recolectados.

- Construir modelos de aprendizaje supervisado para la clasificación del riesgo de depresión. Se emplearán cinco algoritmos de clasificación:

    - Máquinas de soporte vectorial para clasificación (SVM)
    - Red Neuronal para clasificación (ANN)
    - Árboles de decisión
    - K-Nearest Neighbors (KNN)
    - Regresión Logística

  Estos modelos se entrenarán sobre un conjunto de datos balanceado al 70%.

- Aplicar cuatro métodos de ensamble para mejorar la robustez y precisión de los modelos predictivos:

- Random Forest, que combina múltiples árboles de decisión de forma aleatoria.

- XGBoost y CatBoost, técnicas avanzadas de boosting que ajustan iterativamente los modelos para minimizar los errores residuales.

- Voting Hard, un ensamble que combina múltiples modelos de clasificación mediante un sistema de votación.

- Comparar y evaluar el rendimiento de cada modelo utilizando al menos cuatro métricas de calidad, como precisión, recall, F1-score y curva ROC, además de la matriz de confusión.

- Seleccionar los tres mejores modelos utilizando un análisis estadístico de diferencia significativa (ANOVA y prueba de Tukey) para garantizar que la elección de los modelos tenga una base estadísticamente válida.

- Optimizar los tres modelos seleccionados mediante hiperparametrización usando GridSearch y métodos avanzados como BayesSearchCV (optimización bayesiana) y GASearchCV (algoritmos genéticos) para maximizar el desempeño predictivo de cada modelo

- Desplegar el modelo final mediante un Pipeline de preparación de datos, integrándolo en una interfaz gráfica desarrollada con Streamlit**.** Esta interfaz permitirá al usuario cargar datos, ejecutar el modelo y visualizar las predicciones de riesgo, proporcionando una herramienta funcional y accesible para la toma de decisiones en salud mental y bienestar psicológico. El despliegue se realizará con LocalTunnel para generar un enlace público que facilite el acceso remoto a la aplicación, permitiendo que otros usuarios interactúen con la herramienta de forma sencilla y segura.

## 1.4 DISEÑO DE SOLUCIÓN

| Problema | Tipo de Análisis | Tipo de Aprendizaje | Tarea Analítica | Requerimientos en los Datos | Métodos | Evaluación | Calidad Esperada |
|---|---|---|---|---|---|---|---|
| Construir modelos de clasificación para predecir el riesgo de depresión e identificar los factores asociados al riesgo. | Predictivo | Supervisado | Análisis exploratorio y clasificación del riesgo | Datos de calidad con variables demográficas, sociales y laborales balanceados al 70% para evitar sesgos | Modelos: SVM, ANN, Árbol de decisión, KNN y Regresión Logística | Matriz de confusión, Precisión, recall, F1-score, curva ROC | Modelos de clasificación con precisión y sensibilidad superiores al 80%. |
| Mejorar la precisión de los modelos utilizando técnicas de ensamble | Predictivo | Supervisado | Ensamble para precisión y estabilidad | Datos limpios y preprocesados; balanceo previo al 70%; partición en conjuntos de entrenamiento y prueba | Ensambles: Random Forest, XGBoost, CatBoost, Voting Hard (con voto mayoritario) | Precisión, recall, F1-score, curva ROC | Modelos ensamblados con desempeño superior a modelos individuales. Aumento en precisión y robustez, con mejora del 5-10% en métricas sobre modelos individuales. |
| Seleccionar los mejores modelos mediante un análisis estadístico de significancia | No aplica | No aplica | Comparación y selección de modelos | Conjunto de datos de entrenamiento y validación, asegurando balanceo previo | ANOVA y prueba de Tukey | Diferencia estadística en rendimiento entre modelos | Selección de modelos basada en significancia estadística con una confiabilidad mínima del 95%. |
| Optimizar el rendimiento de los modelos seleccionados a través de hiperparametrización | No aplica | No aplica | Optimización de hiperparámetros | Conjunto de datos de validación balanceado para GridSearch y optimización avanzada de hiperparámetros | Hiperparametrización con GridSearch, BayesSearchCV y GASearchCV para modelos seleccionados | Matriz de confusión, Precisión, recall, F1-score, curva ROC tras optimización | Mejora en el rendimiento de los modelos seleccionados en al menos un 5% en comparación con los resultados iniciales. |
| Desplegar el modelo final en una interfaz gráfica para predicción del riesgo de depresión | No aplica | No aplica | Despliegue mediante una interfaz gráfica | Pipeline de datos con preprocesamiento completo, normalización, codificación y preparación de datos para despliegue | Implementación en Streamlit, con despliegue público mediante LocalTunnel | No aplica | Interfaz funcional y amigable para el usuario |

## 1.5 RECURSOS PARA CREACIÓN DEL MODELO Y PARA DESPLIEGUE

| Aspecto | Detalles |
|---|---|
| Entorno de Desarrollo | La creación del modelo se llevará a cabo de forma local utilizando Jupyter Notebook en Visual Studio Code. |
| Lenguaje de Programación | Se empleará Python tanto para la creación y entrenamiento del modelo como para el despliegue en Streamlit. |
| Librerías y Herramientas | - **Limpieza de datos**: pandas, numpy, sklearn<br>- **Entrenamiento y evaluación de modelos**: scikit-learn<br>- **Optimización y búsqueda de hiperparámetros**: GridSearchCV, BayesSearchCV, GASearchCV<br>- **Visualización de datos**: matplotlib, seaborn<br>- **Despliegue**: Streamlit, LocalTunnel |
| Repositorio y Control de Versiones | Los cambios del proyecto se subirán a un repositorio remoto en GitHub, permitiendo un control de versiones adecuado y colaboración si es necesario. |
| Despliegue del Modelo | - El modelo se desplegará en una aplicación de Streamlit, donde los usuarios podrán cargar datos, ejecutar el modelo y ver los resultados en una interfaz gráfica.<br><br>- Se utilizará LocalTunnel para generar un enlace público y permitir el acceso remoto a la aplicación, facilitando el uso del modelo en otros dispositivos o ubicaciones sin necesidad de configuración compleja. |
| Interacción Modelo-Interfaz | La interfaz de Streamlit cargará el modelo desde un archivo preentrenado almacenado localmente. Streamlit ejecutará el modelo en tiempo real y mostrará los resultados al usuario, asegurando una comunicación directa entre el modelo y la interfaz. |
| Licencias y Requisitos Legales | - **Licencia de Python y bibliotecas**: Python es de código abierto, y las bibliotecas seleccionadas tienen licencias que permiten el uso y modificación libre.<br>- **Visual Studio Code**: de código abierto, con licencias de uso gratuito para desarrolladores.<br>- **Streamlit y LocalTunnel**: también de código abierto y gratuitos. |
| IDE | Visual Studio Code para la creación y pruebas locales del modelo en Jupyter Notebooks. |

# 2. ENTENDIMIENTO DE LOS DATOS

## 2.1 CICLO DE LOS DATOS

### Generación de los datos

Los datos se generan a partir de una encuesta anónima que fue diseñada específicamente para evaluar factores demográficos, sociales y de estilo de vida relacionados con el riesgo de depresión en adultos. La encuesta fue distribuida en varias ciudades y capturó información autodeclarada de personas de entre 18 y 60 años sobre factores como la satisfacción laboral, horas de trabajo/estudio, antecedentes familiares, entre otros.

### Almacenamiento de los datos

Actualmente, los datos recolectados están almacenados en un archivo CSV que ha sido descargado de Kaggle. Este archivo se gestiona de manera local, pero se almacena en un repositorio remoto en GitHub para facilitar el control de versiones y permitir la colaboración entre los integrantes del equipo.

**Modificación de los datos**
Dado que se trabaja con un archivo descargado, la manipulación de datos se hace por los miembros del equipo de análisis de datos. Las operaciones de preprocesamiento, limpieza y transformación de datos se realizarán mediante scripts en Python.

**Periodicidad de los datos y reentrenamiento del modelo.**
Los datos de la encuesta fueron recolectados en un solo periodo (enero a junio de 2023), lo que significa que el conjunto de datos es estático y no recibirá actualizaciones periódicas. En caso de futuras expansiones del proyecto o de la encuesta, se podrían establecer nuevas fases de recolección de datos. De ser así, el modelo podría ser reentrenado cada 2-3 años si se recopilan datos adicionales que reflejen cambios en los factores de riesgo para la depresión.

## 2.2 DICCIONARIO DE DATOS

| Variable | Tipo | Descripción |
|---|---|---|
| Name | Categórica | Nombre del participante (identificador anonimizado). |
| Gender | Categórica | Género del participante. |
| Age | Numérica | Edad del participante, en años. |
| City | Categórica | Ciudad de residencia del participante. |
| Working Professional or Student | Categórica | Estado laboral del participante. |
| Profession | Categórica | Profesión del participante, si aplica. |
| Academic Pressure | Categórica | Nivel de presión académica percibida. |
| Work Pressure | Categórica | Nivel de presión laboral percibida. |
| CGPA | Numérica | Calificación promedio acumulada (solo para estudiantes). |
| Study Satisfaction | Categórica | Nivel de satisfacción con los estudios, solo para estudiantes. |
| Job Satisfaction | Categórica | Nivel de satisfacción laboral, solo para profesionales en activo. |
| Sleep Duration | Categórica | Duración promedio del sueño del participante. |
| Dietary Habits | Categórica | Hábitos alimenticios. |
| Degree | Categórica | Grado académico máximo obtenido por el participante. |
| Have you ever had suicidal thoughts? | Categórica | Si el participante ha tenido pensamientos suicidas. |
| Work/Study Hours | Numérica | Promedio de horas dedicadas al trabajo o estudio diariamente. |
| Financial Stress | Categórica | Nivel de estrés financiero percibido. |
| Family History of Mental Illness | Categórica | Si el participante tiene antecedentes familiares de enfermedades mentales. |
| Depression | Categórica | Variable objetivo: riesgo de depresión, basada en la evaluación de factores demográficos y de vida. |

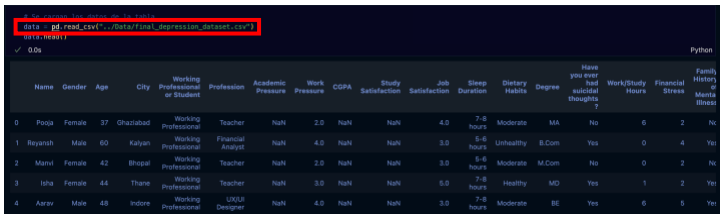
## 2.3 REGLAS DE CALIDAD

| Variable | Regla de calidad |
|---|---|
| Name | ------- |
| Gender | Entre 18 y 60 años. |
| Age | Male, Female. |
| City | Debe coincidir con una ciudad válida en el país. Cualquier otra entrada es un error. |
| Working Professional or Student | Working Professional, Student. |
| Profession | Solo se permiten valores entre las profesiones listadas (e.g., Teacher, HR Manager, Doctor, etc.). |
| Academic Pressure | 1, 2, 3, 4, 5. |
| Work Pressure | 1, 2, 3, 4, 5. |
| CGPA | Entre 5.0 y 10.0. Cualquier otro valor es un error. |
| Study Satisfaction | 1, 2, 3, 4, 5. |
| Job Satisfaction | 1, 2, 3, 4, 5. |
| Sleep Duration | Less than 5 hours, 5-6 hours, 7-8 hours, More than 8 hours. |
| Dietary Habits | Healthy, Moderate, Unhealthy. |
| Degree | Debe coincidir con un título académico válido listado en la base de datos (e.g., B.Tech, B.Sc, M.Tech, PhD, etc.). |
| Have you ever had suicidal thoughts? | Yes, No. |
| Work/Study Hours | Entre 0 y 12 horas diarias. |
| Financial Stress | 1, 2, 3, 4, 5. |
| Family History of Mental Illness | Yes, No. |
| Depression | Yes, No. |

# 3. PREPARACIÓN DE DATOS (Estadística)

## 3.1 INTEGRACIÓN

Se consolidará una única tabla (sábana de datos) en la que se integran todas las variables necesarias para el análisis. En este caso, los datos están en un solo archivo CSV, por lo que no se requieren uniones adicionales o joins.





Al inicio, se tienen 19 variables en el dataset.

## 3.2 SELECCIÓN DE VARIABLES

Se eliminan las variables irrelevantes o que no aporten al análisis, así como aquellas que puedan comprometer la privacidad de los participantes. En este caso, la variable *Name* no es útil para el modelo y será eliminada.

```python
# Variables irrelevantes para el proceso de minería
data = data.drop(['Name'], axis=1) # Axis = 1 para eliminar la columna
data.head()
```

Python

| | Gender | Age | City | Working Professional or Student | Profession | Academic Pressure | Work Pressure | CGPA | Study Satisfaction | Job Satisfaction | Sleep Duration | Dietary Habits | Degree | Have you ever had suicidal thoughts ? | Work/Study Hours | Financial Stress | Family History of Mental Illness | Depres |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Female | 37 | Ghaziabad | Working Professional | Teacher | NaN | 2.0 | NaN | NaN | 4.0 | 7-8 hours | Moderate | MA | No | 6 | 2 | No | |
| 1 | Male | 60 | Kalyan | Working Professional | Financial Analyst | NaN | 4.0 | NaN | NaN | 3.0 | 5-6 hours | Unhealthy | B.Com | Yes | 0 | 4 | Yes | |

## 3.3 DESCRIPCIÓN ESTADÍSTICA

Se realizará un análisis estadístico descriptivo de las variables numéricas y categóricas.

Dataset statistics

| | |
|---|---|
| **Number of variables** | 18 |
| **Number of observations** | 2556 |
| **Missing cells** | 7839 |
| **Missing cells (%)** | 17.0% |

Variable types

| | |
|---|---|
| **Categorical** | 15 |
| **Numeric** | 3 |

El dataset contiene 2556 registros y 19 variables, de las cuales 15 son categóricas y 3 numéricas. Hay 7839 celdas nulas que equivalen al 17% del dataset.

Alerts

| | |
|---|---|
| `Profession` has 673 (26.3%) missing values | Missing |
| `Academic Pressure` has 2054 (80.4%) missing values | Missing |
| `Work Pressure` has 502 (19.6%) missing values | Missing |
| `CGPA` has 2054 (80.4%) missing values | Missing |
| `Study Satisfaction` has 2054 (80.4%) missing values | Missing |
| `Job Satisfaction` has 502 (19.6%) missing values | Missing |
| `Work/Study Hours` has 204 (8.0%) zeros | Zeros |

En el conjunto de datos, algunas columnas tienen un porcentaje significativo de valores nulos, particularmente en las variables *Academic Pressure*, *CGPA* y *Study Satisfaction*. Esta situación se explica porque el conjunto de datos incluye tanto a estudiantes como a profesionales (variable *Working Professional or Student*), y estas variables son relevantes solo para los estudiantes.

**Academic Pressure:** Mide la presión académica percibida, pero únicamente aplica a los estudiantes. Como los profesionales no están sujetos a presión académica en este contexto, esta columna tiene valores nulos para ellos.

**CGPA:** Representa el promedio acumulado, una medida que se utiliza para estudiantes en contexto académico.

**Study Satisfaction:** Mide el nivel de satisfacción con los estudios, relevante solo para estudiantes.

Esto se puede demostrar viendo los datos

| Working Professi... | Profession | # Academic Press... | # Work Pressure | # CGPA | # Study |
|---|---|---|---|---|---|
| Working Professi... 80% | [null] 26% | | | | |
| Student 20% | Teacher 13% | | | | |
| | Other (1561) 61% | 1 5 | 1 5 | 5.03 10 | 1 |
| Working Professional | Teacher | | 2 | | |
| Working Professional | Financial Analyst | | 4 | | |
| Working Professional | Teacher | | 2 | | |
| Working Professional | Teacher | | 3 | | |
| Working Professional | UX/UI Designer | | 4 | | |
| Working Professional | Civil Engineer | | 1 | | |
| Working Professional | Accountant | | 4 | | |
| Working Professional | Teacher | | 1 | | |
| Working Professional | Lawyer | | 1 | | |
| Working Professional | Content Writer | | 3 | | |
| Working Professional | Mechanical Engineer | | 4 | | |
| Working Professional | Civil Engineer | | 3 | | |
| Working Professional | Consultant | | 2 | | |
| Student | | 1 | | 9.9 | 4 |
| Working Professional | Data Scientist | | 5 | | |
| Student | | 1 | | 5.97 | 5 |
| Working Professional | Accountant | | 3 | | |
| Working Professional | Pharmacist | | 4 | | |
| Working Professional | Software Engineer | | 1 | | |
| Working Professional | Teacher | | 5 | | |
| Student | | 4 | | 9.05 | 4 |
| Student | | 4 | | 9.96 | 2 |
| Working Professional | Teacher | | 1 | | |
| Student | | 4 | | 6.17 | 1 |
| Working Professional | Travel Consultant | | 1 | | |

Al observar los datos, se confirma que los registros asociados a *Working Professional* tienen valores nulos en estas columnas, mientras que los registros categorizados como *Student* contienen valores válidos. Este comportamiento refleja que el conjunto de datos fue diseñado para capturar información específica según el rol del participante (estudiante o profesional), lo cual es la causa de los valores faltantes en ciertas columnas.

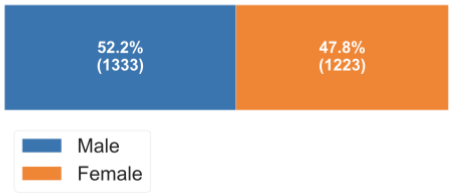
## Estudio de variables – Feature Engineering

### Gender

Categorical

| | |
|---|---|
| **Distinct** | 2 |
| **Distinct (%)** | 0.1% |
| **Missing** | 0 |
| **Missing (%)** | 0.0% |
| **Memory size** | 2.7 KiB |

Common Values (Plot)

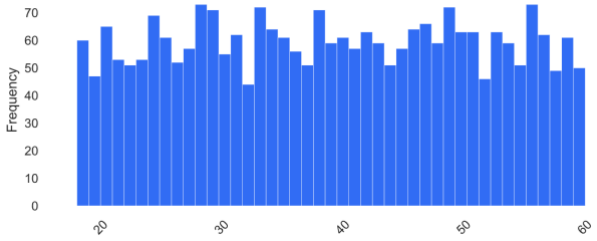



Contiene 2 categorías distintas, no tiene nulos, está balanceada.

### Age

Real number (ℝ)

| | | | |
|---|---|---|---|
| **Distinct** | 43 | **Minimum** | 18 |
| **Distinct (%)** | 1.7% | **Maximum** | 60 |
| **Missing** | 0 | **Zeros** | 0 |
| **Missing (%)** | 0.0% | **Zeros (%)** | 0.0% |
| **Infinite** | 0 | **Negative** | 0 |
| **Infinite (%)** | 0.0% | **Negative (%)** | 0.0% |
| **Mean** | 39.04303599 | **Memory size** | 20.1 KiB |

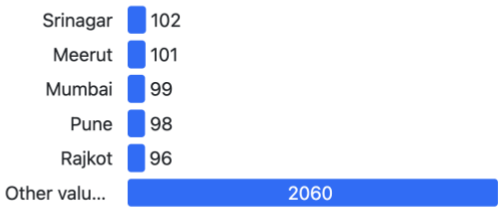



Rango de edad entre 18-60, no tiene nulos, media 39.04, se distribuye balanceadamente.

### City

Categorical

| | |
|---|---|
| **Distinct** | 30 |
| **Distinct (%)** | 1.2% |
| **Missing** | 0 |
| **Missing (%)** | 0.0% |
| **Memory size** | 3.9 KiB |



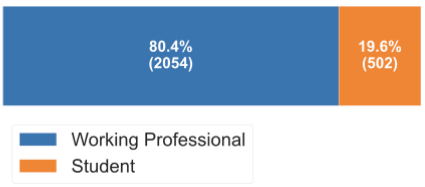

30 cuidades distintas, no contiene nulos.

### Working Professional or Student

Categorical

| | |
|---|---|
| **Distinct** | 2 |
| **Distinct (%)** | 0.1% |
| **Missing** | 0 |
| **Missing (%)** | 0.0% |
| **Memory size** | 2.7 KiB |

Common Values (Plot)



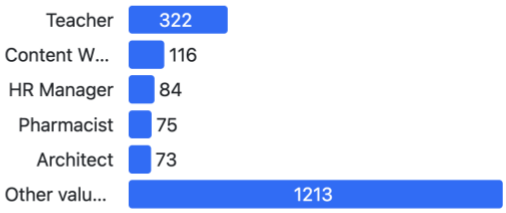

El 80.4% de los encuestados son profesionales, el 19.6% son estudiantes, está desbalanceada y no tiene nulos.

## Profession

Categorical

Missing

| | |
|---|---|
| **Distinct** | 35 |
| **Distinct (%)** | 1.9% |
| **Missing** | 673 |
| **Missing (%)** | 26.3% |
| **Memory size** | 3.9 KiB |




Se tienen 35 profesiones distintas, hay 673 (26.3%) datos nulos que corresponden a los registros de estudiantes que no tienen una profesión aun y a otros datos vacíos.

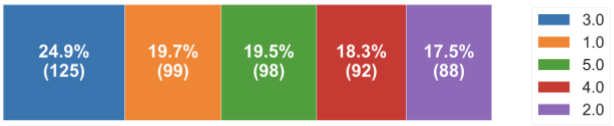
## Academic Pressure

Categorical

Missing

| | |
|---|---|
| **Distinct** | 5 |
| **Distinct (%)** | 1.0% |
| **Missing** | 2054 |
| **Missing (%)** | 80.4% |
| **Memory size** | 2.8 KiB |

Common Values (Plot)




Hay 2054 valores nulos, como se mencionó anteriormente, la mayoría de encuestados son profesionales, razón por la cual esta columna tiene un número significativo de datos nulos.

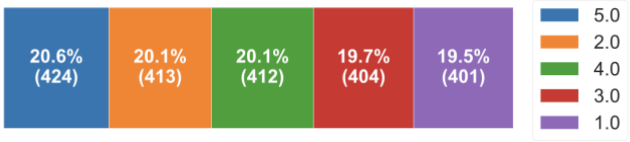
## Work Pressure

Categorical

Missing

| | |
|---|---|
| **Distinct** | 5 |
| **Distinct (%)** | 0.2% |
| **Missing** | 502 |
| **Missing (%)** | 19.6% |
| **Memory size** | 2.8 KiB |

Common Values (Plot)



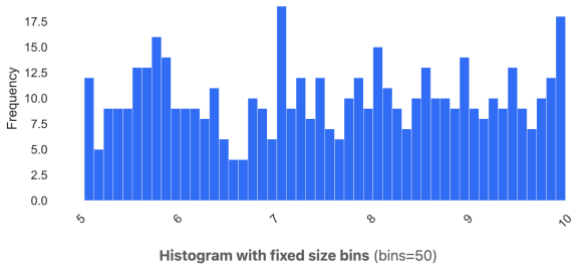

Hay 502 valores nulos que corresponden a los registros de los estudiantes que no presentan presión laboral sino presión académica (variable anterior). Más delante se deciden unificar estas 2 columas en una sola llamada *Work/Academic Pressure.*

## CGPA

Real number (ℝ)

Missing

| | | | |
|---|---|---|---|
| Distinct | 312 | Minimum | 5.03 |
| Distinct (%) | 62.2% | Maximum | 10 |
| Missing | 2054 | Zeros | 0 |
| Missing (%) | 80.4% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 7.567808765 | Memory size | 20.1 KiB |



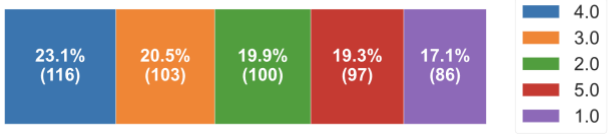

El CGPA es el promedio acumulado de los estudiantes. Se tiene la misma situación, esta variable solo puede ser calculada para estudiantes, por ello tiene tantos nulos.

## Study Satisfaction

Categorical

Missing

| | |
|---|---|
| Distinct | 5 |
| Distinct (%) | 1.0% |
| Missing | 2054 |
| Missing (%) | 80.4% |
| Memory size | 2.8 KiB |



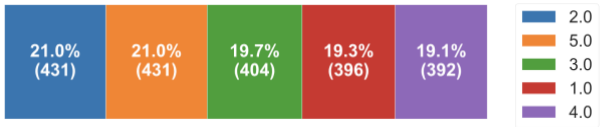

Para la satisfacción con el estudio sucede lo mismo, solo se llena con datos de estudiantes. Por esto tiene tanto nulos.

## Job Satisfaction

Categorical

Missing

| | |
|---|---|
| Distinct | 5 |
| Distinct (%) | 0.2% |
| Missing | 502 |
| Missing (%) | 19.6% |
| Memory size | 2.8 KiB |

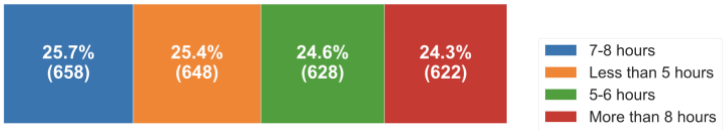



Esta columna solo tiene los valores de satisfacción del trabajo (solo profesionales). Tiene 502 valores nulos que corresponden a los registros de los estudiantes. Luego se unifican también estas 2 columnas.

## Sleep Duration

Categorical

| | |
|---|---|
| Distinct | 4 |
| Distinct (%) | 0.2% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 2.8 KiB |

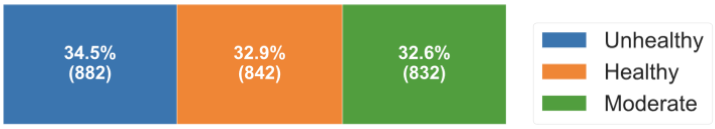



Esta variable no tiene nulos y está distribuida balanceadamente.

Dietary Habits

Categorical

| | |
|---|---|
| Distinct | 3 |
| Distinct (%) | 0.1% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 2.8 KiB |

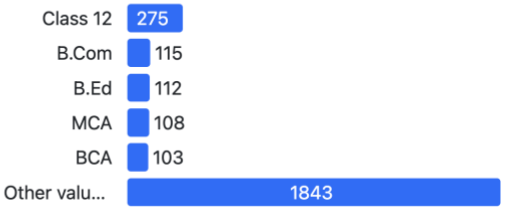



La variable Dietary Habits está balanceda, posee 3 categorías y no tiene nulos.

Degree

Categorical

| | |
|---|---|
| Distinct | 27 |
| Distinct (%) | 1.1% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 3.9 KiB |



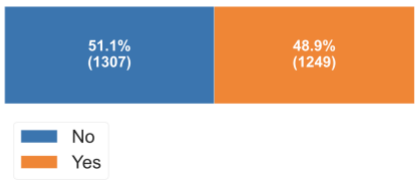

Degree es el nivel educativo del encuestado, esta variable no tiene nulos y tiene 27 categorías.

Have you ever had suicidal thoughts ?

Categorical

| | |
|---|---|
| Distinct | 2 |
| Distinct (%) | 0.1% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 2.7 KiB |

Common Values (Plot)



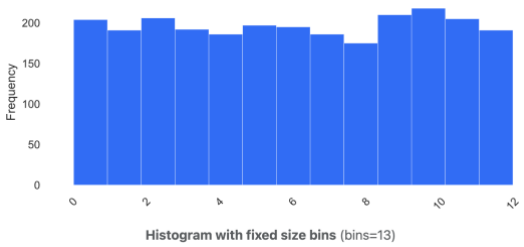

La variable “Has tenido pensamientos suicidas?” no contiene nulos, está distribuida balanceadamente.

Work/Study Hours

Real number (ℝ)

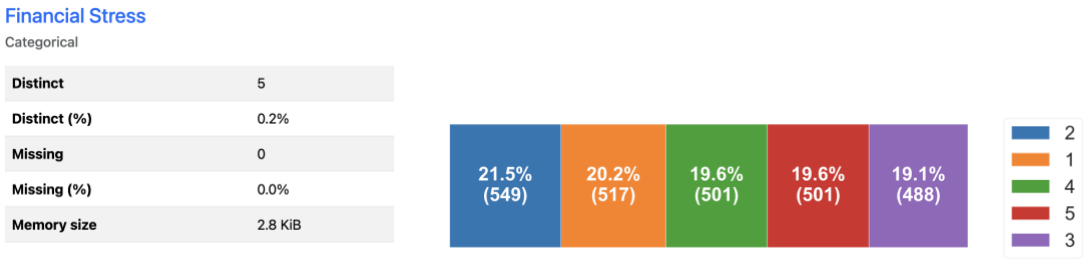
Zeros

| | | | |
|---|---|---|---|
| Distinct | 13 | Minimum | 0 |
| Distinct (%) | 0.5% | Maximum | 12 |
| Missing | 0 | Zeros | 204 |
| Missing (%) | 0.0% | Zeros (%) | 8.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 6.023865415 | Memory size | 20.1 KiB |

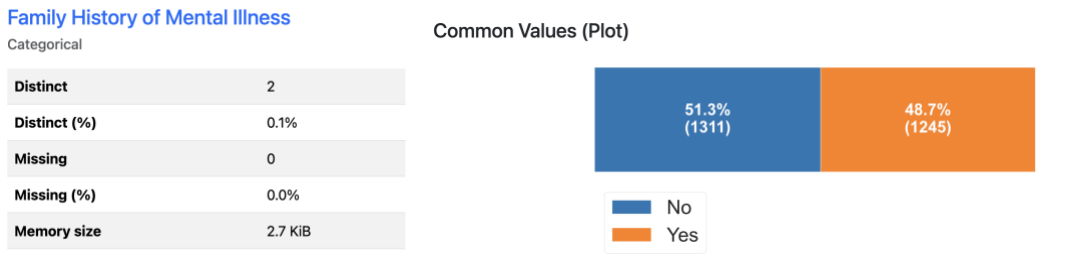



Esta columna numérica no tiene valores nulos, el rango de horas de sueño es de 0-12. La media de horas de sueño es de 6.02 horas. Se identifica que

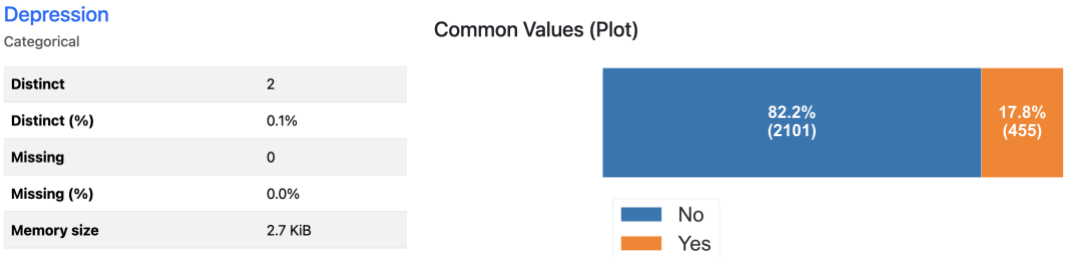
hay 204 valores que son 0, esto significa que hay registros de participantes que trabajan o estudian 0 horas, sin embargo, esto puede ser posible.




La variable estrés financiero está posee una escala de 1-5, esta balanceada y no tiene nulos.




La variable indica si se tienen antecedentes familiares con enfermedades mentales. No tiene nulos, está balanceada.




Depresión es la variable objetivo que indica si hay riesgo o no de que un participante sufra depresión. No tiene nulos pero está desbalanceada.

### 3.4 LIMPIEZA DE ATÍPICOS

El dataset no contiene datos atípicos.

## 3.5 LIMPIEZA DE NULOS

En el análisis y descripción estadística de los datos, se identificó un número significativo de valores nulos en varias columnas, especialmente en aquellas que corresponden exclusivamente a estudiantes.

| | Nulos | Porcentaje (%) |
|---|---|---|
| CGPA | 2054 | 80.359937 |
| Academic Pressure | 2054 | 80.359937 |
| Study Satisfaction | 2054 | 80.359937 |
| Profession | 673 | 26.330203 |
| Work Pressure | 502 | 19.640063 |
| Job Satisfaction | 502 | 19.640063 |
| Age | 0 | 0.000000 |
| Gender | 0 | 0.000000 |
| City | 0 | 0.000000 |
| Working Professional or Student | 0 | 0.000000 |
| Sleep Duration | 0 | 0.000000 |

Para abordar este problema, se tomaron las siguientes medidas:

- Se decidió eliminar la columna *CGPA* (promedio acumulado de estudiantes) debido a que presenta un porcentaje de valores nulos superior al 30%, lo cual no permite su imputación sin comprometer la calidad de los datos

- Se creó una nueva columna llamada *Academic/Work Pressure*, que combina los datos de *Academic Pressure* (para estudiantes) y *Work Pressure* (para profesionales). De esta forma, se obtiene una columna unificada sin valores nulos. De manera similar, se creó la columna *Study/Job Satisfaction* unificando *Study Satisfaction* (para estudiantes) y *Job Satisfaction* (para profesionales), logrando así una variable completa y sin nulos.

  Estas combinaciones fueron posibles gracias a la relación observada entre estudiantes y profesionales durante el análisis de perfilado de datos, permitiendo crear columnas que engloban ambas perspectivas de presión y satisfacción.

```
# Fill null values by combining columns
data['Job/Study Satisfaction'] = data['Job Satisfaction'].fillna(data['Study Satisfaction'])
data['Work/Academic Pressure'] = data['Academic Pressure'].fillna(data['Work Pressure'])

# Drop unnecessary columns
data.drop(columns=['Job Satisfaction', 'Study Satisfaction', 'Academic Pressure', 'Work Pressure'], inplace=True)
data.head()
```

- Por ultimo, para la columna Profession se rellenaron los valores nulos de la columna *Profession* con la moda, que es *Teacher*. Luego, para los registros donde se detectó que el participante es estudiante (según

la columna *Working Professional or Student),* se cambió el valor de *Profession* a *Student*, asegurando así que cada participante tiene una profesión asignada.

```
# Null data cleaning: Imputation by mode
from sklearn.impute import SimpleImputer

var_categoricas = ['Profession']

# #Imputation of categorical variables: mode
ImpCategorias = SimpleImputer(missing_values=np.nan, strategy='most_frequent')
data[var_categoricas] = ImpCategorias.fit_transform(data[var_categoricas])

print(ImpCategorias.statistics_)
```

```
['Teacher']
```

```
# Ensure that 'Student' is a valid category in 'Profession'
if data['Profession'].dtype.name == 'category':
    if 'Student' not in data['Profession'].cat.categories:
        data['Profession'] = data['Profession'].cat.add_categories(['Student'])

# Set 'Profession' to 'Student' where 'Working Professional or Student' is 'Student'
data.loc[data['Working Professional or Student'] == 'Student', 'Profession'] = 'Student'
```

```
data['Profession'].value_counts()
```

```
Profession
Student                502
Teacher                493
Content Writer         116
HR Manager              84
Pharmacist              75
```

Con estas transformaciones, se obtuvo un conjunto de datos más limpio, eliminando los valores nulos en las columnas.

```
Data columns (total 15 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   Gender                                 2556 non-null   category
 1   Age                                    2556 non-null   int64
 2   City                                   2556 non-null   category
 3   Working Professional or Student        2556 non-null   category
 4   Profession                             2556 non-null   object
 5   Sleep Duration                         2556 non-null   category
 6   Dietary Habits                         2556 non-null   category
 7   Degree                                 2556 non-null   category
 8   Have you ever had suicidal thoughts ?  2556 non-null   category
 9   Work/Study Hours                       2556 non-null   int64
 10  Financial Stress                       2556 non-null   category
 11  Family History of Mental Illness      2556 non-null   category
 12  Depression                             2556 non-null   category
 13  Job/Study Satisfaction                 2556 non-null   category
 14  Work/Academic Pressure                 2556 non-null   category
```

Luego de la limpieza de nulos se obtuvieron **15** columnas sin nulos.

## 3.6 CREACIÓN DE NUEVAS VARIABLES

No se crearon nuevas variables para efectos de este ejercicio de predicción.

## 3.7 ANÁLISIS DE CORRELACIONES PARA REDUNDANCIA

Se analizaron las correlaciones entre las variables para identificar aquellas que sean redundantes. Las variables con una alta correlación entre sí (superiores a 0.8) indicaron redundancia. En tales casos, se eliminaron estas variables.

Estos son los pasos que se siguieron para identificar correlaciones:

Se utilizaron dummies para transformar las variables categóricas en variables numéricas. Luego, la variable objetivo Depression fue transformada en valores numéricos mediante LabelEncoder, asignando 0 para "No" y 1 para "Yes".

### Matriz de Correlación

Se calculó una matriz de correlación que incluye tanto variables numéricas como categóricas utilizando el paquete dython, lo cual permite analizar asociaciones entre variables categóricas y numéricas y se generó un heatmap para visualizar las correlaciones, identificando correlaciones altas (0.8 - 1.0).

Se observó una correlación alta entre las columnas *Working Professional or Student* y *Profession,* lo que indica que ambas variables representan información similar.

Para reducir la redundancia, se decidió eliminar la columna *Profession*, conservando *Working Professional or Student* como la variable representativa.

## 3.8 ANÁLISIS DE CORRELACIONES PARA IRRELEVANCIA

Se evaluó la relevancia de cada variable para la predicción del riesgo de depresión. Las variables que no tenían ninguna correlación significativa con la variable objetivo ("Depression") fueron consideradas irrelevantes y eliminadas, ya que no aportan al modelo predictivo.

Se identificaron variables con correlaciones muy bajas (en el rango de 0.0 a 0.1) con la variable objetivo Depression. Las variables *City*, *Family History of Mental Illness, Sleep Duration* y *Gender* presentaron correlaciones muy bajas y, por lo tanto fueron eliminadas del conjunto de datos para simplificar el modelo.

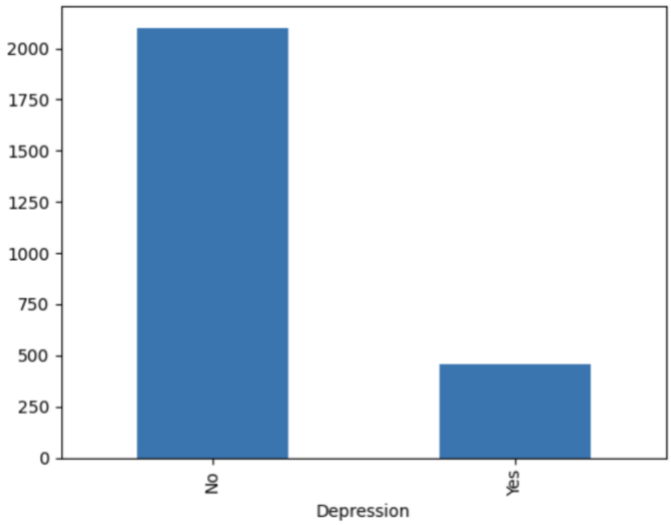
Con estos pasos, se logró reducir la dimensionalidad del conjunto de datos mediante la eliminación de variables redundantes e irrelevantes. Estas son las variables luego del proceso de análisis de correlaciones, quedan 10 variables.

```
Data columns (total 10 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Age                                     2556 non-null   int64
 1   Working Professional or Student         2556 non-null   category
 2   Dietary Habits                          2556 non-null   category
 3   Degree                                  2556 non-null   category
 4   Have you ever had suicidal thoughts ?   2556 non-null   category
 5   Work/Study Hours                        2556 non-null   int64
 6   Financial Stress                        2556 non-null   category
 7   Depression                              2556 non-null   category
 8   Job/Study Satisfaction                  2556 non-null   category
 9   Work/Academic Pressure                  2556 non-null   category
dtypes: category(8), int64(2)
memory usage: 62.4 KB
```

## 3.9 REDUCCIÓN DE DIMENSIÓN (OPCIONAL EN PREDICCIONES)

No se realizó PCA para la reducción de dimensionalidad.

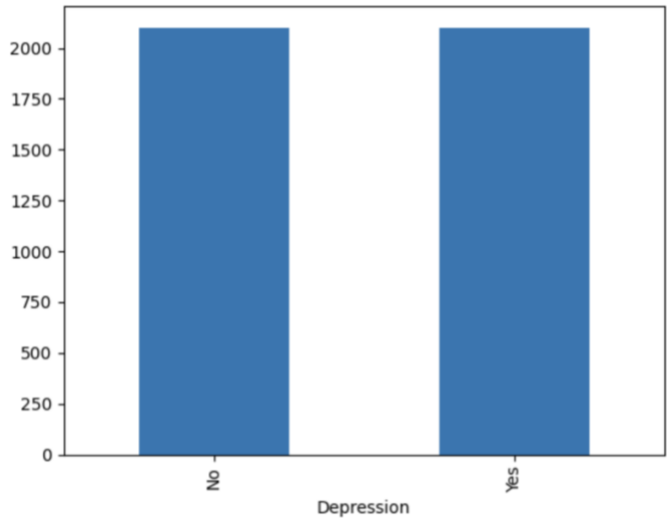
### 3.10 BALANCEO




Dado que la variable objetivo *Depression* está desbalanceada (la mayoría de los registros no presentan riesgo de depresión), es importante aplicar técnicas de balanceo como SMOTE. El objetivo es lograr que el modelo pueda aprender de ambas clases de manera equilibrada y mejorar la precisión de la clasificación para ambas categorías ("Yes" y "No").

Se aplicó el método SMOTE para balancear la variable objetivo, dado que el conjunto de datos incluye tanto variables categóricas como numéricas, se utilizó SMOTENC para el balanceo. Se generaron muestras sintéticas de la clase minoritaria (Yes), logrando un conjunto de datos balanceado.

Tras aplicar SMOTENC, el número total de registros aumentó de 2556 a 4202, logrando una distribución equitativa entre las clases de la variable Depression.

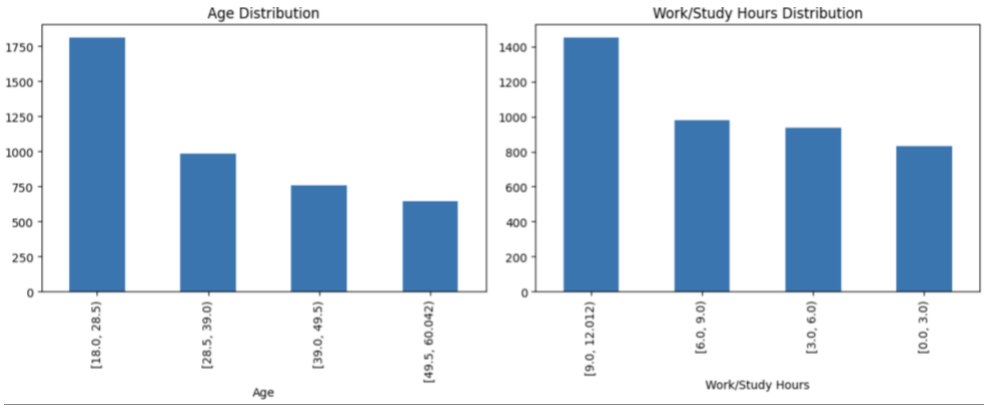


## 3.11 TRANSFORMACIONES

Para adaptarse a los diferentes requerimientos de los modelos de Machine Learning, se generaron dos versiones del conjunto de datos: uno categórico y otro numérico. Esto permite que cada modelo trabaje con el tipo de datos que necesita:

**Dataset categórico:** Diseñado para métodos de ML que requieren variables discretizadas o categóricas (árboles de decisión, Naive Bayes).

- **Discretización:** Variables numéricas como *Age* y *Work/Study Hours* se agruparon en intervalos (bins) para modelos que funcionan mejor con datos categóricos, como los árboles de decisión.




Asi se ve el conjunto de datos categórico.

| | Age | Working Professional or Student | Dietary Habits | Degree | Have you ever had suicidal thoughts ? | Work/Study Hours | Financial Stress | Job/Study Satisfaction | Work/Academic Pressure | Depression |
|---|---|---|---|---|---|---|---|---|---|---|
| 4084 | [18.0, 28.5) | Student | Healthy | M.Pharm | Yes | [3.0, 6.0) | 4 | 4.0 | 5.0 | Yes |
| 199 | [49.5, 60.042) | Working Professional | Moderate | M.Pharm | No | [9.0, 12.012) | 5 | 4.0 | 1.0 | No |
| 2520 | [18.0, 28.5) | Working Professional | Moderate | Class 12 | No | [9.0, 12.012) | 3 | 5.0 | 1.0 | No |
| 2330 | [39.0, 49.5) | Working Professional | Moderate | BSc | Yes | [3.0, 6.0) | 2 | 1.0 | 4.0 | No |
| 3523 | [18.0, | Student | Moderate | M.Pharm | Yes | [6.0, 9.0) | 4 | 4.0 | 5.0 | Yes |

**Conjunto numérico:** Preparado para modelos de ML que requieren datos numéricos y normalizados (Redes neuronales, SVM, regresión logística, KNN).

- **Normalización:** Se aplicó la normalización Min-Max a variables numéricas para modelos que requieren datos continuos en un rango de 0 a 1, como redes neuronales y SVM.

- **Dummies:** Las variables categóricas fueron convertidas en variables numéricas para que todos los modelos pudieran interpretarlas correctamente.

- **Codificación de la Variable Objetivo:** La variable Depression fue codificada en 0 y 1 con LabelEncoder.

| | Age | Work/Study Hours | Depression | Working Professional or Student_Working Professional | Have you ever had suicidal thoughts ?_Yes | Dietary Habits_Healthy | Dietary Habits_Moderate | Dietary Habits_Unhealthy | Degree_B.Arch | Degree_B.Com |
|---|---|---|---|---|---|---|---|---|---|---|
| 13 | 0.785714 | 0.666667 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 3749 | 0.023810 | 0.416667 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 1094 | 0.357143 | 0.583333 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2650 | 0.095238 | 0.500000 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 2823 | 0.309524 | 0.250000 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |

Así se ve el conjunto de datos numérico.

Con estas transformaciones, cada conjunto de datos está listo para ajustarse a los requerimientos específicos de cada modelo de ML, optimizando el rendimiento y la precisión.

## 4. MODELAMIENTO, EVALUACIÓN E INTERPRETACIÓN

### 4.1 CONFIGURACIÓN MÉTODOS DE MACHINE LEARNING

Para el modelado, se seleccionaron nueve métodos de Machine Learning en total: cinco modelos base y cuatro modelos de ensamble como se mencionó en la propuesta de solución. A continuación, se detallan los aspectos de configuración y el proceso de entrenamiento para cada uno.

**Modelos Base**

**Support Vector Classifier (SVC):** Este modelo utiliza el algoritmo de clasificación de máquinas de vectores de soporte, configurado con un kernel lineal y habilitado para calcular probabilidades (probability=True). Esto permite generar la curva ROC y el AUC. La regularización se controla a través del parámetro C=1.

**Artificial Neural Network (ANN):** Se emplea una red neuronal multicapa con una capa oculta de 26 neuronas y activación relu. La red utiliza un aprendizaje constante con una tasa de aprendizaje inicial de 0.02, momento de 0.3 y un máximo de 500 iteraciones. Esta configuración busca equilibrar un modelo simple y efectivo para problemas de clasificación.

**K-Nearest Neighbors (KNN):** Este modelo se configura con 4 vecinos más cercanos y una métrica de distancia euclidiana. La simplicidad del modelo hace que sea adecuado para comparar con otros métodos más complejos.

**Logistic Regression:** Se utiliza como modelo de clasificación lineal, configurado con el algoritmo lbfgs para la optimización. Este modelo es útil debido a su simplicidad y capacidad interpretativa en la clasificación binaria.

**Decision Tree:** Un árbol de decisión con profundidad máxima de 2 y criterio gini, configurado para reducir el sobreajuste, al limitar el número de muestras mínimas por hoja a 20.

**Modelos de Ensamble**

**Random Forest:** Un RandomForest compuesto por 300 árboles, con un muestreo máximo de 70% en cada árbol y mínimo de 2 muestras por hoja. Este modelo busca aprovechar el consenso de múltiples árboles para reducir el riesgo de sobreajuste y mejorar la precisión general.

**XGBoost:** Configurado con una profundidad máxima de 10, tasa de aprendizaje de 0.1, y 100 estimadores. XGBoost es conocido por su rendimiento en problemas de clasificación, optimizando iterativamente las predicciones a través de árboles de decisión.

**CatBoost:** Este modelo se configura con 100 iteraciones, una profundidad de 10 y una tasa de aprendizaje de 0.1. CatBoost se destaca en su capacidad para manejar datos categóricos sin necesidad de codificación adicional.

**Voting Hard:** Es un clasificador combinado que toma decisiones mediante un voto mayoritario de los modelos Random Forest, XGBoost y CatBoost. Este ensamble se utiliza para capturar la robustez de los tres modelos al unificarlos en una sola predicción.

**Configuración y Entrenamiento de los Modelos**

**Validación Cruzada:** Para evaluar de forma consistente el rendimiento de cada modelo, se aplicó una validación cruzada estratificada de 10 particiones. Este proceso permite medir el desempeño promedio de cada modelo en múltiples métricas (exactitud, precisión, recall, F1-score y tiempo de ajuste).

**Entrenamiento:** Cada modelo fue entrenado utilizando el conjunto de entrenamiento balanceado. Se aplicó el método de sobremuestreo SMOTE al

conjunto de entrenamiento para igualar las clases y mitigar el sesgo en las predicciones.

**Evaluación de Desempeño:** Se utilizaron múltiples métricas de rendimiento, incluyendo exactitud, precisión, recall, F1-score, y AUC (cuando el modelo permitía predict_proba). Además, se generaron gráficos de matrices de confusión y curvas ROC para visualizar la efectividad y la capacidad de discriminación de cada modelo.

Cada modelo fue configurado cuidadosamente de acuerdo con sus características propias y se ajustaron los parámetros relevantes para maximizar su rendimiento en la tarea de clasificación. Estos métodos permitieron evaluar diferentes enfoques y niveles de complejidad, desde modelos más simples como KNN hasta modelos complejos de ensamble como XGBoost y Voting Hard.

## 4.2 ANALISIS DE MEDIDAS DE CALIDAD

En el análisis de este problema de predicción, se utilizaron múltiples modelos de clasificación, incluyendo modelos base y modelos de ensamble. Cada modelo se evaluó usando varias métricas de rendimiento:

- Exactitud (accuracy)
- Precisión (precision)
- Sensibilidad (recall)
- F1 Score
- Área bajo la curva (AUC).

Estas métricas son esenciales en un problema de clasificación binaria, ya que permiten evaluar tanto la habilidad del modelo para predecir correctamente ambas clases (riesgo de depresión y no riesgo de depresión) como su capacidad para mantener un equilibrio entre las métricas de precisión y sensibilidad.

### Métricas de Validación Cruzada para Modelos base y de ensamble

- Support Vector Classifier y Logistic Regression mostraron un rendimiento alto y consistente, con una exactitud promedio en el conjunto de prueba superior al 97%, lo que indica que estos modelos generalizan bien.

- Artificial Neural Network también mostró buenos resultados, con una exactitud del 96.9% en prueba, lo que sugiere que captura correctamente las relaciones no lineales en los datos.

- K-Nearest Neighbors y Decision Tree tuvieron un rendimiento relativamente menor, con exactitudes de 88.7% y 82.8%, respectivamente. Esto sugiere que estos modelos pueden estar sobreajustados o subajustados en comparación con otros.

- XGBoost, CatBoost y Voting Hard mostraron un rendimiento superior en el conjunto de prueba, con precisiones de validación cruzada alrededor del 96.8%.

- Random Forest tuvo una precisión de prueba ligeramente inferior (96.0%), aunque sigue siendo competitiva.

**Resultados Después del Entrenamiento de Modelos Base**

Después de entrenar los modelos con el conjunto balanceado, los resultados en el conjunto de prueba fueron los siguientes:

- Support Vector Classifier y Logistic Regression mantuvieron una alta precisión y sensibilidad, ambas alrededor del 96.2%. Estas métricas sugieren que ambos modelos son robustos y confiables para este tipo de predicción.

- Artificial Neural Network alcanzó una precisión del 94.6% y una F1 de 94.5%, siendo una opción competitiva.

- K-Nearest Neighbors y Decision Tree continuaron mostrando un rendimiento más bajo en términos de precisión y sensibilidad, indicando que pueden no ser tan adecuados para este problema.

- Voting Hard y CatBoost se destacaron con una exactitud de 95.5% y 95.3%, respectivamente. Estos modelos también mostraron un buen equilibrio entre precisión y sensibilidad, haciendo que sean altamente confiables.

- XGBoost y Random Forest tuvieron una exactitud y F1 de alrededor de 94.7%, siendo modelos efectivos aunque ligeramente por debajo de los otros ensambles.

En cuanto al AUC, tanto Support Vector Classifier como Logistic Regression alcanzaron valores altos (0.992), indicando una gran capacidad para diferenciar entre las clases de riesgo de depresión y no riesgo. Otros modelos, como Artificial Neural Network y K-Nearest Neighbors, presentaron valores de AUC entre 0.856 y 0.985, lo cual es aceptable pero menor en comparación.

**Análisis de las Matrices de Confusión**

Aquí se observan los siguientes puntos clave:

Support Vector Classifier y Logistic Regression: Estos modelos muestran un bajo número de falsos negativos y falsos positivos, lo que es indicativo de su precisión en ambas clases.

Artificial Neural Network: Este modelo tiene un número ligeramente mayor de falsos negativos en comparación con Logistic Regression y Support Vector Classifier, lo que puede indicar una ligera disminución en su sensibilidad.

K-Nearest Neighbors y Decision Tree: Estos modelos tienen un número considerable de falsos positivos, lo que puede indicar problemas en la detección precisa de casos sin riesgo de depresión.

| Modelo | Exactitud | Precisión | Recall | F1 Score | AUC |
|---|---|---|---|---|---|
| Voting Hard | 95.6% | 95.7% | 95.6% | 95.6% | N/A |
| CatBoost | 95.3% | 95.6% | 95.3% | 95.4% | 0.9903 |
| Logistic Regression | 96.2% | 96.3% | 96.2% | 96.3% | 0.9929 |
| Support Vector Classifier | 96.2% | 96.1% | 96.2% | 96.2% | 0.9922 |
| XGBoost | 94.8% | 95.1% | 94.8% | 94.9% | 0.9865 |
| Random Forest | 94.6% | 94.8% | 94.6% | 94.7% | 0.9832 |
| Artificial Neural Network | 94.7% | 94.5% | 94.7% | 94.5% | 0.9854 |
| Decision Tree | 82.8% | 87.0% | 82.8% | 84.1% | 0.8565 |
| K-Nearest Neighbors | 82.3% | 87.1% | 82.3% | 83.7% | 0.8878 |

Entre los modelos base, Logistic Regression y Support Vector Classifier son los mejores debido a sus altas métricas de evaluación en todas las categorías. Entre los modelos de ensamble, Voting Hard y CatBoost son los mejores debido a su excelente capacidad predictiva y balance entre precisión y sensibilidad. Voting Hard tiene una ligera ventaja por su método de votación que combina los puntos fuertes de múltiples modelos. En conjunto, estos modelos son los más adecuados para la predicción de riesgo de

depresión, con una buena capacidad para generalizar y una baja tasa de errores de clasificación.

## 4.3 SELECCIÓN DEL MEJOR MODELO

Para la selección del modelo óptimo, se consideraron tanto las métricas de rendimiento (precisión, exactitud, recall, F1 y AUC) como la complejidad computacional (tiempo de entrenamiento y predicción). Se llevaron a cabo los siguientes pasos:

### Análisis de Varianza y Prueba de Tukey

Se realizó un ANOVA para determinar si había diferencias estadísticamente significativas en las métricas de F1-score entre los modelos.

La prueba de Tukey mostró que algunos pares de modelos no presentaban diferencias significativas, como es el caso de Logistic Regression, Support Vector Classifier, XGBoost, entre otros.

Esto permitió reducir la selección a un conjunto de modelos donde las diferencias de rendimiento no eran significativas.

### Ranking de Modelos Sin Diferencias Significativas

De acuerdo con los resultados de la ANOVA y Tukey, se identificaron los modelos con diferencias no significativas:

- Logistic Regression
- Artificial Neural Network
- CatBoost
- Support Vector Classifier
- Voting Hard
- XGBoost
- Random Forest.

Estos modelos se ordenaron según su F1-score promedio en validación cruzada, donde Logistic Regression lideró con el mejor desempeño (F1 promedio = 0.9725), seguido de otros modelos como Artificial Neural Network y CatBoost.

**Evaluación de Complejidad Computacional:**

Los modelos se evaluaron según su tiempo promedio de entrenamiento y predicción. Los tres modelos con menor tiempo total fueron:

- Logistic Regression (13.399 ms)
- Support Vector Classifier (167.178 ms)
- XGBoost (183.313 ms)
-

Esto llevó a seleccionar estos tres modelos para la etapa final, equilibrando el rendimiento y la eficiencia computacional.

**Hiperparametrización:**

Los modelos seleccionados fueron optimizados usando GridSearch, BayesSearch y Algoritmos Genéticos para encontrar los mejores hiperparámetros.

Al final, los mejores parámetros obtenidos para cada modelo fueron:

- Logistic Regression: {'C': 10, 'solver': 'liblinear'}
- Support Vector Classifier: {'C': 10, 'gamma': 'scale', 'kernel': 'linear'}
- XGBoost: {'learning_rate': 0.08, 'max_depth': 3, 'n_estimators': 500}

**Evaluación Final:**
Los modelos optimizados fueron evaluados en el conjunto de prueba y se obtuvo el siguiente rendimiento:

| Model | Accuracy | Precision | Recall | F1 Score | AUC | Train Time (s) | Prediction Time (s) |
|---|---|---|---|---|---|---|---|
| **Logistic Regression** | 96.6% | 96.7% | 96.6% | 96.6% | 0.9931 | 0.0069 | 0.0016 |
| **Support Vector Classifier** | 95.8% | 95.8% | 95.8% | 95.8% | N/A | 0.0608 | 0.0034 |
| **XGBoost** | 95.0% | 95.3% | 95.0% | 95.1% | 0.9894 | 0.5143 | 0.0053 |

**Conclusión:**

Logistic Regression fue elegido como el mejor modelo, ya que ofrece el rendimiento más alto en todas las métricas principales y es el modelo más eficiente en términos de tiempo de entrenamiento y predicción.

Support Vector Classifier y XGBoost también muestran un rendimiento sólido, pero son menos eficientes computacionalmente.

Este análisis concluye que la Regresión Logística es la opción óptima para este problema de predicción de riesgo de depresión, logrando un balance ideal entre precisión, desempeño general y eficiencia computacional.

## 5. DESPLIEGUE

La fase de despliegue del modelo se centró en la creación de un pipeline que permitiera procesar y predecir el riesgo de depresión a partir de nuevos datos. Este pipeline fue integrado en una interfaz gráfica desarrollada en Streamlit, lo cual permite al usuario cargar datos, ejecutar el modelo y visualizar las predicciones de manera interactiva. La aplicación fue desplegada de forma local y accesible mediante LocalTunnel, generando un enlace público para facilitar el acceso remoto.

### 5.1 PREDICCIÓN DE DATOS FUTUROS

Para este proyecto, el despliegue se realiza a través de una interfaz gráfica que permite una predicción accesible para los usuarios finales. La interfaz, desarrollada en Streamlit, incluye la siguiente funcionalidad:

**Carga y procesamiento de datos:**

Los usuarios pueden cargar un archivo CSV con nuevos datos de personas, incluyendo sus hábitos y condiciones demográficas. Los datos son procesados por el pipeline de transformación y limpieza, el cual maneja cualquier valor atípico o faltante en las variables relevantes.
Este preprocesamiento garantiza que los datos estén en un formato adecuado para el modelo, aplicando transformaciones necesarias para una predicción precisa.

**Pipeline de Preparación de Datos:**

Se creó un pipeline utilizando la biblioteca scikit-learn para estandarizar el proceso de preparación y predicción de datos. El pipeline sigue estos pasos:

- **Carga y conversión de categorías:** Las columnas categóricas son transformadas en variables de tipo category.

- **Creación de nuevas características:** Algunas columnas como Job Satisfaction y Study Satisfaction se combinan para formar nuevas variables Job/Study Satisfaction y Work/Academic Pressure, permitiendo al modelo tener información consolidada.

- **Eliminación de columnas irrelevantes:** Se eliminan columnas que no aportan valor al modelo final, como Name, City, Family History of Mental Illness, entre otras.

- **Transformación y escalado de variables:** Las columnas numéricas se escalan con MinMaxScaler y las variables categóricas se codifican mediante OneHotEncoder.

**Pipeline completo con modelo:**

Al final del pipeline, se incorpora el modelo de Regresión Logística, que fue elegido como el mejor modelo tras el proceso de selección.

Este pipeline permite que los nuevos datos pasen por todas las etapas de preprocesamiento antes de ser evaluados por el modelo, generando predicciones de riesgo de depresión.

**Despliegue en Streamlit:**

El pipeline y el modelo final fueron integrados en una aplicación Streamlit que permite al usuario cargar los datos futuros y recibir predicciones de forma sencilla.

La aplicación fue desplegada utilizando LocalTunnel, lo cual permite que la aplicación sea accesible a través de un enlace público, facilitando el acceso remoto para otros usuarios sin necesidad de configuraciones adicionales.
Pipeline de Predicción para Datos Futuros:

## 5.2 CRONOGRAMA DE MANTENIMIENTO

El modelo de predicción y su interfaz desplegada están sujetos a un cronograma de mantenimiento, que se define para asegurar que el modelo continúe ofreciendo resultados precisos y relevantes. El mantenimiento incluye:

Reentrenamiento del modelo: Para mantener la precisión del modelo, se recomienda realizar un reentrenamiento cada seis meses, utilizando nuevos datos de usuarios si están disponibles. Esto permite que el modelo se adapte a posibles cambios en la población o patrones emergentes en los datos.

Revisión del pipeline de preprocesamiento: Se revisará el pipeline de preprocesamiento de datos cada tres meses para asegurar que se manejen adecuadamente posibles cambios en los valores categóricos, rangos de variables numéricas y otras transformaciones relevantes. Cualquier ajuste necesario será implementado para mantener la consistencia en las predicciones.

La implementación de este pipeline, junto con la interfaz gráfica en Streamlit y el acceso remoto mediante LocalTunnel, ofrece una solución completa y funcional para la predicción del riesgo de depresión. La estructura del pipeline asegura que el procesamiento de datos y las predicciones se realicen de manera consistente y eficiente. El mantenimiento planificado garantiza que el modelo y su interfaz mantengan un alto nivel de precisión y utilidad para los usuarios finales en el tiempo.