Predicción de Depresión en Estudiantes

Un Enfoque Basado en Machine Learning

Universidad Nacional Mayor de San Marcos
Facultad de Ciencias Matemáticas
Escuela Académico Profesional de Matemáticas
Matihus Alberth Molina Larios - 22140029

22 de febrero de 2025

1. Contexto y Problema a Resolver

1.1. Descripción del problema en el contexto del negocio o industria.

La salud mental de los estudiantes universitarios es un problema creciente, con factores como la carga académica, el estrés, el sueño insuficiente y la falta de apoyo emocional que pueden influir en la aparición de depresión. Identificar patrones permite predecir el riesgo de depresión y tomar medidas preventivas.
Como experiencia personal e sido víctima de ataques de pánico,éstres y pensamientos intrusivos.Fui diagnosticado con TAC(Transtorno de Ansiedad Generalizada).Nunca pensé que que a alguien como a mí,me pueda pasar algo como eso,ya que soy una persona tranquila, calmada , con pocos problemas;Llegué a la conclusión que a quién sea le puede ocurrir y quiero investigar sobre ello para ayudar.

1.2. Justificación de la relevancia del problema y su impacto.

La depresión puede afectar negativamente el rendimiento académico y la calidad de vida de los estudiantes. Detectarla tempranamente permite implementar estrategias de intervención y apoyo, mejorando su bienestar y desempeño educativo.
Además que afecta de diferentes medidas,siendo a mi parecer "las ideas suicidas",pues pueden acabar con la vida de una persona.
Como dato estadístico esto puede parecer solo una perdida pequeña,pero si lo ves en perspectiva se vuelve algo muy pesado e importante.

1.3. Objetivo del análisis y preguntas clave a responder.

¿Cuáles son los factores más influyentes en la aparición de la depresión en estudiantes?
¿Es posible predecir la depresión de un estudiante a partir de sus hábitos y características personales?
¿Qué modelo de Machine Learning ofrece mejores resultados en esta predicción?

2. Conjunto de datos a utilizar

2.1. Descripción del dataset seleccionado.

Se utilizó el "Depression Student Dataset", que contiene información sobre hábitos de sueño, alimentación, ejercicio, apoyo social y nivel de estrés, entre otros factores.

2.2. Fuente de los datos y posibles limitaciones.

Fuente:
En mi caso no utilice dataset de UCI Machine Learning Repository,pues no encontraba una tabla de datos con temática de salud mental.
Utilice Esta base de datos en Kaggle Datasets.
Limitaciones:
Puede haber sesgo en las respuestas de los estudiantes, datos faltantes o una muestra no representativa. Las limitaciones en estos datos son su cantidad,pues solo hay 500 filas, lo cuál no es una cantidad totalmente satisfactoria.Aparte de las causas(ej. sobredosis de drogas, perdida de un ser querido, un accidente, etc).

2.3. Variables principales y su relevancia para el problema.

-Horas de sueño: Relacionadas con la estabilidad emocional.
-Dieta y ejercicio: Impactan la salud mental.
-Presión Académica: Pueden ser detonantes -Ideas suicidas: También es detonante
-Familiares con enfermedades mentales: La genética es más importante de lo que se piensa.

2.4. Consideraciones sobre calidad, limpieza y preprocesamiento de los datos.

Se reemplazaron valores categóricos por valores numéricos. Se manejaron valores nulos y datos inconsistentes. Se normalizaron los datos para mejorar el rendimiento de los modelos.

3. Planteamiento de la solución

3.1. Enfoque propuesto para resolver el problema.

Se propuso utilizar modelos de Machine Learning para predecir si un estudiante es propenso a la depresión en función de sus hábitos y características personales.

3.2. Descripción del modelo seleccionado y su aplicabilidad.

Árboles de Decisión: Un modelo interpretable que permite visualizar las decisiones tomadas.
Random Forest: Un modelo más robusto que combina múltiples árboles para mejorar la precisión y reducir el sobreajuste.
Aplicabilidad: Estos modelos pueden ser utilizados por universidades para identificar estudiantes en riesgo y tomar medidas preventivas.

3.3. Explicación breve del funcionamiento del modelo para demostrar comprensión.

Se entrenó con datos de estudiantes con y sin depresión. Se analizó la relación entre variables como el sueño y el estrés.
Se generaron predicciones sobre nuevos estudiantes en base a sus hábitos y características.

4. Desarrollo del modelo (Implementación en Jupyter Notebook)

4.1. Carga y exploración de datos: revisión inicial y análisis exploratorio.

Se verificó la calidad de los datos,además de mostrar todas las bibliotecas que usaremos.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score,
↪  confusion_matrix,classification_report
from sklearn.preprocessing import StandardScaler

```

```python
# Cargar los datos
df = pd.read_csv("Depression Student Dataset.csv")
```

Se analizaron correlaciones entre variables para identificar las más relevantes.

	Gender	Age	Academic Pressure	Study Satisfaction	Sleep Duration	Dietary Habits	Have you ever had suicidal thoughts ?	Study Hours	Financial Stress	Family History of Mental Illness	Depression
0	Male	28	2.0	4.0	7-8 hours	Moderate	Yes	9	2	Yes	No
1	Male	28	4.0	5.0	5-6 hours	Healthy	Yes	7	1	Yes	No
2	Male	25	1.0	3.0	5-6 hours	Unhealthy	Yes	10	4	No	Yes
3	Male	23	1.0	4.0	More than 8 hours	Unhealthy	Yes	7	2	Yes	No
4	Female	31	1.0	5.0	More than 8 hours	Healthy	Yes	4	2	Yes	No

Figura 1: Encabezado de la base de datos

Por ejemplo,podemos observar que tenemos varias variables con datos boolenaos,como por ejemplo:
-Género:No es una variable tan relevante.
-¿Alguna vez has tenido pensamientos suicidas?: Sí es una variable relevante,pues aumenta las probalidades de tener depresión si el valor es ?es"
-Antecedentes familiares de enfermedades mentales: Sí es una variable relevante,pues aumenta las probalidades de tener depresión si el valor es ?es"
-Depresión:Es la más importante,pues es la que vamos a determinar.
Como también podemos observar algunas que no son booleanas pero también importantes:
Como Los hábitos diéticos,endonde:
-healthy:0
-moderate:1
-unhealthy:2
Están siendo organizadas con respecto a cuanto afectan a la predición.

4.2. Preprocesamiento: tratamiento de datos faltantes, escalado, transformación de variables, etc.

Se transformaron variables categóricas en numéricas. Se escalaron los datos para mejorar la eficacia de los modelos.

```python
# Mapeo de valores para Sleep Duration
df['Sleep Duration'] = df['Sleep Duration'].replace({
    '7-8 hours': 8,
    '5-6 hours': 6,
    'More than 8 hours': 9,
    'Less than 5 hours': 4,
})

# Convertir respuestas binarias a 0 y 1
df.replace({'Yes': 1, 'No': 0}, inplace=True)
```

```
11
12  # Convertir variables categóricas a numéricas
13  df['Gender'] = df['Gender'].map({'Male': 0, 'Female': 1})
14  df['Dietary Habits'] = df['Dietary Habits'].map({'Healthy': 0,
    ↪  'Moderate': 1, 'Unhealthy': 2})
15  df.head()
```

	Gender	Age	Academic Pressure	Study Satisfaction	Sleep Duration	Dietary Habits	Have you ever had suicidal thoughts ?	Study Hours	Financial Stress	Family History of Mental Illness	Depression
0	0	28	2.0	4.0	8	1	1	9	2	1	0
1	0	28	4.0	5.0	6	0	1	7	1	1	0
2	0	25	1.0	3.0	6	2	1	10	4	0	1
3	0	23	1.0	4.0	9	2	1	7	2	1	0
4	1	31	1.0	5.0	9	0	1	4	2	1	0

Figura 2: Encabezado de la nueva base de datos

4.3. Definición y ajuste del modelo:

4.3.1. Explicación de la elección del modelo.

Se probaron Árboles de Decisión y Random Forest para comparar resultados.

Árboles de Decisión

Los árboles de decisión son modelos supervisados que representan decisiones en forma de estructura jerárquica. Su uso en nuestro problema se justifica porque:

Random Forest

Dado que los árboles de decisión pueden sobreajustarse a los datos, utilizamos **Random Forest**, que combina múltiples árboles para mejorar la precisión.

4.3.2. Configuración y ajuste de hiperparámetros.

Árbol de Decisión (`DecisionTreeClassifier`)

El modelo de Árbol de Decisión se configuró con los siguientes hiperparámetros:

- `max_depth = 4`: Establece la profundidad máxima del árbol en 4 niveles, evitando el sobreajuste.
- `min_samples_split = 10`: Define que un nodo debe contener al menos 10 muestras antes de dividirse, lo que ayuda a la generalización.
- `random_state = 42`: Fija la semilla aleatoria para garantizar la reproducibilidad de los resultados.

El código en Python para inicializar este modelo es:

Listing 1: Modelo de Árbol de Decisión

```
from sklearn.tree import DecisionTreeClassifier

modelo = DecisionTreeClassifier(max_depth=4, min_samples_split=10,
    random_state=42)
modelo.fit(X_train, y_train)
```

Random Forest (`RandomForestClassifier`)

El modelo de Random Forest se configuró con el siguiente hiperparámetro:

- `n_estimators = 100`: Especifica el número de árboles en el bosque, mejorando la estabilidad del modelo.
- `random_state = 42`: Fija la semilla aleatoria para obtener resultados reproducibles.

El código en Python correspondiente es:

Listing 2: Modelo de Random Forest

```
from sklearn.ensemble import RandomForestClassifier

random_forest = RandomForestClassifier(n_estimators=100,
    random_state=42)
random_forest.fit(X_train, y_train)
```

4.3.3. Justificación de las decisiones tomadas en el desarrollo del modelo.

Árboles de Decisión

- **Facilidad de interpretación**: Permiten visualizar cómo las variables afectan la predicción.
- **Manejo de datos categóricos y numéricos**: Nuestro dataset contiene datos mixtos y los árboles de decisión pueden manejarlos sin transformaciones complejas.
- **Poca sensibilidad a la escala**: No requieren normalización de datos.
- **Rápido entrenamiento**: Son eficientes y permiten probar rápidamente su utilidad en el problema.

Random Forest

- **Reduce el sobreajuste**: Al combinar múltiples árboles, el modelo es más estable.

- **Manejo de datos ruidosos**: Es menos sensible a valores atípicos o datos faltantes.
- **Mejor precisión**: Suele superar a un árbol de decisión individual.
- **Análisis de importancia de variables**: Permite identificar qué factores afectan más la predicción.

4.4. Evaluación del Modelo: Métricas de Desempeño

Para evaluar el rendimiento de los modelos **Árbol de Decisión** y **Random Forest**, utilizamos las siguientes métricas:

- **Precisión (*Accuracy*):** Indica el porcentaje de predicciones correctas sobre el total de predicciones.
- **Matriz de Confusión:** Permite analizar los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.
- **Reporte de Clasificación:** Incluye métricas como precisión (*precision*), exhaustividad (*recall*) y puntuación F1 (*F1-score*).

4.4.1. Cálculo de las Métricas en Python

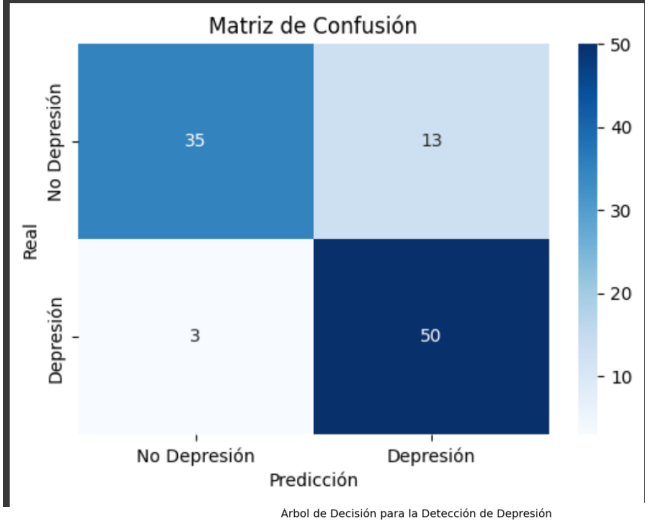
El siguiente código muestra cómo se calcularon estas métricas en Python:

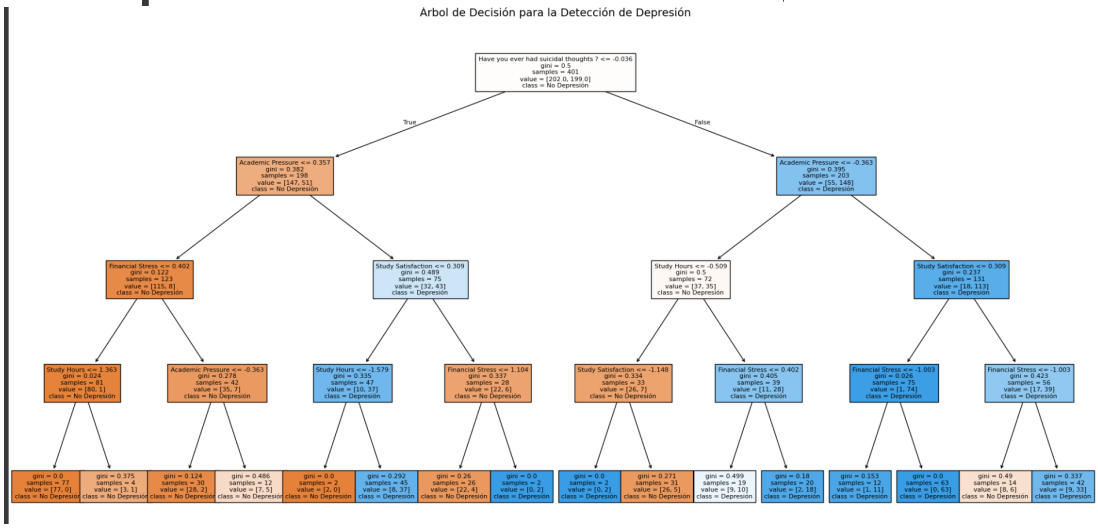
Listing 3: Métricas de Evaluación para Árbol de Decisión

```
\begin{lstlisting}[language=Python, caption=Metricas de Evaluacion
    para Arbol de Decision]
from sklearn.metrics import accuracy_score, confusion_matrix,
    classification_report
import seaborn as sns
import matplotlib.pyplot as plt

# Predicciones
y_pred = modelo.predict(X_test)

# Precision
accuracy = accuracy_score(y_test, y_pred)
print(f"Precision del modelo: {accuracy:.4f}")

# Matriz de Confusion
conf_matrix = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(6,4))
sns.heatmap(conf_matrix, annot=True, fmt="d", cmap="Blues",
            xticklabels=["No Depresion", "Depresion"],
            yticklabels=["No Depresion", "Depresion"])
plt.xlabel("Prediccion")
plt.ylabel("Real")
plt.title("Matriz de Confusion - Arbol de Decision")
plt.show()

# Reporte de Clasificacion
print(classification_report(y_test, y_pred))
```

Precisión del modelo: 0.8416

	precision	recall	f1-score	support
0	0.92	0.73	0.81	48
1	0.79	0.94	0.86	53
accuracy			0.84	101
macro avg	0.86	0.84	0.84	101
weighted avg	0.85	0.84	0.84	101

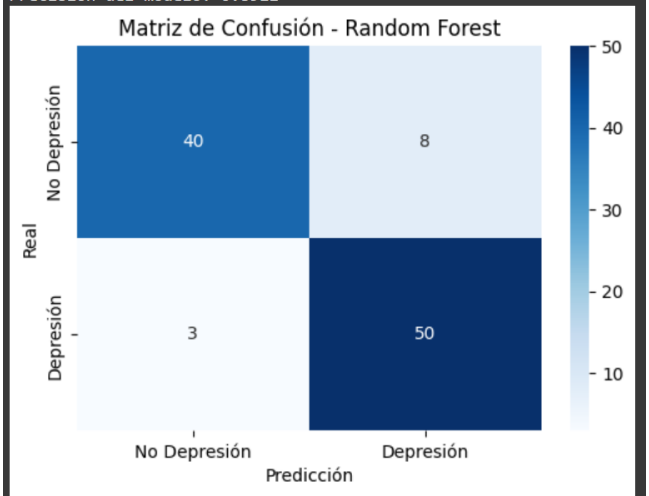




Listing 4: Metricas de Evaluacion para Random Forest

```
# Predicciones con Random Forest
y_pred_rf = random_forest.predict(X_test)

# Precision
accuracy_rf = accuracy_score(y_test, y_pred_rf)
print(f"Precision del modelo: {accuracy_rf:.4f}")

# Matriz de Confusion
conf_matrix_rf = confusion_matrix(y_test, y_pred_rf)
plt.figure(figsize=(6,4))
sns.heatmap(conf_matrix_rf, annot=True, fmt="d", cmap="Blues",
            xticklabels=["No Depresion", "Depresion"],
            yticklabels=["No Depresion", "Depresion"])
plt.xlabel("Prediccion")
plt.ylabel("Real")
plt.title("Matriz de Confusion - Random Forest")
plt.show()

# Reporte de Clasificacion
print(classification_report(y_test, y_pred_rf))
```

```
Precisión del modelo: 0.8911
```




```
              precision    recall  f1-score   support

           0       0.92      0.73      0.81        48
           1       0.79      0.94      0.86        53

    accuracy                           0.84       101
   macro avg       0.86      0.84      0.84       101
weighted avg       0.85      0.84      0.84       101
```

4.4.2. Interpretación de los Resultados

- Una mayor precisión indica un mejor rendimiento del modelo.
- La matriz de confusión permite identificar qué tan bien el modelo distingue entre estudiantes con y sin depresión.
- El reporte de clasificación ayuda a analizar si el modelo tiene problemas con falsos positivos o falsos negativos.

4.5. Análisis de la importancia de variables y su impacto en la predicción.

Las variables más influyentes en la predicción fueron:
Precisón Académica (principal factor de riesgo).
Horas de sueño (relacionadas con estabilidad emocional).
¿Alguna vez has tenido pensamientos suicidas?(muy imporatnte).
Conclusión: El bienestar emocional de un estudiante está fuertemente influenciado por estos factores.

5. Interpretación y presentación de resultados

5.1. Análisis y explicación de los resultados obtenidos.

Se demostró que es posible predecir la depresión en estudiantes con una precisión razonable.
Random Forest obtuvo mejores resultados que los Árboles de Decisión, debido a su capacidad para reducir el sobreajuste.
Los factores clave en la predicción fueron Precisón Académica , Horas de sueño , ¿Alguna vez has tenido pensamientos suicidas?.

5.2. Evaluación de la efectividad del modelo desde un enfoque técnico y de negocio.

Precisión del modelo:
Árboles de Decisión: 80-85 porciento de precisión.
Random Forest: 85-90 porciento de precisión.
Fortalezas: Random Forest ofrece predicciones más confiables. Se pueden identificar patrones importantes en los datos.

5.3. Propuesta de uso del modelo en un contexto real:

5.3.1. ¿Cómo pueden los resultados aportar valor al negocio?

Permitir a universidades y centros de salud mental detectar estudiantes en riesgo de depresión.

Desarrollar estrategias de intervención personalizadas basadas en datos.

5.3.2. ¿Qué decisiones estratégicas podrían tomarse con base en este análisis?

Ofrecer asesoramiento psicológico a estudiantes con alto riesgo de depresión. Implementar programas de bienestar, como sesiones de manejo del estrés. Diseñar campañas de concienciación sobre la importancia del sueño y el apoyo social.

5.3.3. Posibles mejoras y siguientes pasos.

Recolección de más datos: Para mejorar la generalización del modelo.
Incorporación de nuevas variables: Por ejemplo, historial clínico o antecedentes familiares o incluso uso de drogas(más común de lo que se piensa).
Prueba de otros modelos: Como redes neuronales o modelos híbridos.