BORRADOR / apuntes

Datos

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Qué mide | Tipo de datos | Aplicación |
| Wellcome Global Monitor (el actual) | Percepción de salud mental (ansiedad, depresión) basada en encuestas de opinión general. | Opiniones auto-reportadas (subjetivas). | Buenísimo para detectar factores sociales de riesgo y patrones de percepción globales. |
| Dataset de Estilos de Vida 50K | Datos más clínicos: condiciones de salud mental ya diagnosticadas o autorreportadas con severidad. | Variables de hábitos (sueño, ejercicio, dieta, alcohol, tabaco) + Diagnóstico de severidad (Low, Medium, High). | Perfecto para modelar predicción real de condiciones clínicas con mayor certidumbre. |

Variables

Observación de los dataset escogido

**Conclusiones basadas en Estado Laboral y Ansiedad/Depresión**

1. Relación observada

El análisis de la distribución de ansiedad y depresión según el estado laboral mostró diferencias claras entre los grupos.

* Personas desempleadas (Unemployed) y aquellas que trabajan a tiempo parcial deseando tiempo completo (Part\_time\_wants\_fulltime) presentaron mayores tasas de ansiedad o depresión en comparación con aquellos empleados a tiempo completo.
* Las personas fuera de la fuerza laboral (Out\_of\_workforce) también mostraron una proporción relevante de afectación, aunque ligeramente menor que los desempleados.

2. Resultado de la prueba estadística (Chi-Cuadrado)

La prueba de Chi-Cuadrado aplicada sobre la tabla de contingencia entre estado laboral y presencia de ansiedad/depresión arrojó:

* p-value < 0.05, indicando que la asociación es estadísticamente significativa.
* Esto significa que el estado laboral influye de forma real en la probabilidad de experimentar ansiedad o depresión, no siendo el patrón observado fruto del azar.

3. Interpretación

Estos resultados sugieren que la inseguridad laboral, el desempleo y las condiciones laborales precarias son factores de riesgo relevantes para la salud mental:

* El desempleo no solo afecta el bienestar económico, sino también el bienestar emocional.
* La inestabilidad laboral (trabajo parcial no deseado) incrementa el riesgo de presentar síntomas de ansiedad o depresión.

4. Aplicabilidad

Estas conclusiones respaldan la necesidad de:

* Intervenciones en salud pública enfocadas en apoyar psicológicamente a personas desempleadas o en situaciones laborales precarias.
* Políticas laborales que promuevan la estabilidad en el empleo como estrategia indirecta para mejorar la salud mental de la población.

**Conclusiones basadas en Región Global y Ansiedad/Depresión**

1. Relación observada

El análisis mostró que existen diferencias claras en la prevalencia de ansiedad y depresión entre las distintas regiones del mundo:

* Regiones como América Latina y el Caribe y Medio Oriente y Norte de África tienden a reportar tasas más altas de ansiedad o depresión en comparación con otras áreas.
* En cambio, regiones como Asia Oriental o Oceanía presentan niveles más bajos de reportes de problemas de salud mental.

**2. Resultado de la prueba estadística (Chi-Cuadrado)**

La prueba de Chi-Cuadrado aplicada al cruce entre Global11Regions y MH7A arrojó:

* p-value < 0.05, indicando que la relación es estadísticamente significativa.
* Esto confirma que no todas las regiones son iguales respecto a la experiencia de ansiedad o depresión.

**3. Interpretación**

La variabilidad regional en los niveles de ansiedad o depresión podría explicarse por factores como:

* Condiciones socioeconómicas (mayor pobreza o inestabilidad económica en algunas regiones).
* Factores culturales (mayor o menor estigmatización para hablar de salud mental).
* Acceso desigual a servicios de salud mental.
* Recientes crisis políticas o humanitarias en determinadas regiones.

Esto resalta la necesidad de contextualizar los enfoques de salud mental según la realidad social y cultural de cada región.

**4. Aplicabilidad**

Estos resultados son relevantes para:

* Diseño de políticas de salud pública sensibles al contexto regional.
* Intervenciones específicas que consideren el nivel de acceso, la cultura, y las percepciones sociales de cada región.
* Comparaciones internacionales de prevalencia de ansiedad y depresión.

|  |  |
| --- | --- |
| Variable | Justificación de Inclusión |
| Edad (Age) | La edad es un determinante clave en la salud mental. Estudios muestran que ciertos grupos etarios, especialmente jóvenes adultos, reportan mayor prevalencia de ansiedad y depresión. |
| Género (Gender) | Diferencias de género en la prevalencia de trastornos mentales son ampliamente documentadas. Generalmente, las mujeres reportan tasas más altas de ansiedad y depresión que los hombres. |
| Nivel Educativo (Education) | El nivel educativo influye en factores como estabilidad laboral, ingresos y acceso a servicios de salud mental, lo cual impacta la salud emocional. |
| Quintil de Ingreso del Hogar (Household\_Income) | El ingreso económico es un determinante social de la salud. Bajos niveles de ingreso se asocian con mayor riesgo de problemas de salud mental. |
| Percepción de Ingreso (Subjective\_Income) | Más allá del ingreso real, la percepción de suficiencia económica impacta directamente en el nivel de estrés y bienestar emocional. |
| Estado Laboral (EMP\_2010) | El desempleo o la inestabilidad laboral son factores de riesgo reconocidos para el desarrollo de síntomas de ansiedad y depresión. |
| Región Global (Global11Regions) | Las diferencias culturales, socioeconómicas y de acceso a servicios de salud mental entre regiones pueden afectar la prevalencia de trastornos mentales. |

2. Criterios de Selección

Para la elección de variables, se utilizaron los siguientes criterios:

* Relevancia teórica: Variables respaldadas por literatura previa en salud pública y psicología social.
* Disponibilidad de datos: Variables consistentemente presentes y con bajo nivel de valores faltantes en el dataset.
* Relación causal lógica: Se seleccionaron solo variables que conceptualmente pueden actuar como causas o factores de riesgo, evitando aquellas que serían consecuencia de la ansiedad/depresión (por ejemplo, buscar ayuda profesional, estrategias de afrontamiento).
* Evitar leakage: No se incluyeron variables que pudieran introducir conocimiento futuro del evento objetivo (como historial de consulta médica).

3. Variables No Incluidas y Justificación

* Coping mechanisms (MH9, MH10): Son consecuencia de sufrir ansiedad/depresión, no predictores originales.
* Consultation history (MH8A): Buscar ayuda médica ocurre después de experimentar los síntomas, no antes.
* Trust in institutions / Religion / Attitudes towards science (TR\*, REL\*, Q\*): Variables culturales que afectan otras dimensiones, pero no directamente predictores primarios de ansiedad en este modelo.

Elección de modelos

|  |  |  |
| --- | --- | --- |
| Modelo | ¿Por qué lo usamos? | Ventajas |
| Random Forest | Modelo de árboles en bagging. Muy robusto y difícil de sobreajustar. | Maneja bien variables categóricas. |
| Logistic Regression | Modelo lineal base. Es el comparativo estándar en clasificación binaria. | Fácil interpretación de coeficientes. |
| XGBoost Classifier | Árboles mejorados por boosting. Mucho mejor manejo de interacciones complejas. | Mejor performance en general. |

se compararon tres algoritmos de clasificación: Random Forest, Logistic Regression y XGBoost, utilizando las variables seleccionadas tras el análisis exploratorio.  
El modelo Random Forest, ajustado con un umbral de decisión reducido (0.4), logró un recall de 32% en la detección de casos de ansiedad, superando ampliamente a Logistic Regression, que no logró identificar ningún caso positivo (recall 0%).  
XGBoost, por su parte, mostró un comportamiento intermedio, con un recall de 12% y una mejor precisión (44%) en los casos positivos.  
Estos resultados sugieren que los modelos de árboles de decisión, especialmente Random Forest, presentan mayor capacidad de detección en contextos de clases desbalanceadas como el presente.  
Dado que en problemas de salud mental es prioritario minimizar falsos negativos (no detectar a una persona en riesgo), Random Forest fue seleccionado como modelo base para posteriores experimentaciones y ajustes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelo | Variables | Accuracy | Recall Ansiedad | Precision Ansiedad | F1-Score Ansiedad |
| Random Forest | Básicas | 0.64 | 0.36 | 0.25 | 0.29 |
| XGBoost | Básicas | 0.63 | 0.50 | 0.27 | 0.35 |
| XGBoost Tuned + SMOTE | Básicas | 0.49 | 0.74 | 0.25 | 0.37 |
| XGBoost Tuned + SMOTE + wbi | Básicas + wbi | 0.51 | 0.72 | 0.25 | 0.37 |
| XGBoost Tuned + SMOTE + wbi + Age\_Group | Básicas + wbi + Age\_Group | 0.51 | 0.72 | 0.25 | 0.37 |

Dataset 2

La comparación de modelos en el dataset de diagnóstico real muestra que, si bien el accuracy global de Random Forest, Logistic Regression y XGBoost es similar (50%), el modelo XGBoost Tuned logra un recall del 96% en la detección de casos de problemas de salud mental.

Esto indica que XGBoost es capaz de identificar casi la totalidad de los individuos en riesgo, priorizando la sensibilidad sobre la precisión, lo cual es crucial en contextos donde el costo de un falso negativo (no detectar un caso real) es mucho más grave que un falso positivo.

A pesar de los ajustes para priorizar el recall mediante class\_weight='balanced', los modelos Random Forest y Logistic Regression no lograron mejoras sustanciales en la detección de casos de salud mental. Esto sugiere que la naturaleza del problema requiere modelos más complejos, capaces de capturar interacciones no lineales entre variables, siendo XGBoost Tuned el modelo que demostró la mejor sensibilidad (recall del 96%) y robustez frente al desbalance de clases.

**Comparación de modelos**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelo | Ajustes | Accuracy | Recall Problema Mental | Precision Problema Mental | F1-Score Problema Mental |
| Random Forest | SMOTE + Balanced | 0.50 | 47% | 50% | 49% |
| Logistic Regression | SMOTE + Balanced | 0.50 | 50% | 50% | 50% |
| XGBoost Tuned | SMOTE + Tuned scale\_pos\_weight | 0.50 | 96% | 50% | 66% |

Los experimentos realizados sobre dos conjuntos de datos — uno basado en percepciones de ansiedad y depresión (Wellcome Global Monitor) y otro basado en diagnósticos clínicos (Mental Health Dataset) — demostraron de manera consistente la superioridad de XGBoost Tuned frente a modelos clásicos como Random Forest y Logistic Regression.

En el dataset de percepciones, XGBoost Tuned, complementado con técnicas de rebalanceo (SMOTE) y el enriquecimiento de variables (wbi y Age\_Group), logró un recall del 72%, optimizando la detección de posibles casos de ansiedad.

En el dataset de diagnóstico real, el mismo enfoque permitió alcanzar un recall del 96%, superando ampliamente a Random Forest (47%) y Logistic Regression (50%).

Estos resultados resaltan la importancia de utilizar modelos de boosting y estrategias de tratamiento de desbalance de clases en problemas de salud mental, donde la prioridad es maximizar la sensibilidad y minimizar el riesgo de falsos negativos.

A graph with blue bars

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Interpretación de tu gráfico

✅ Global11Regions (Región geográfica):

* Valores altos (regiones más vulnerables) tienden a empujar hacia más ansiedad (lado derecho).
* Valores bajos (regiones más desarrolladas) empujan hacia menos ansiedad.

✅ wbi (Ingreso del país):

* Ingreso alto (rojo) tiende a empujar hacia no ansiedad (lado izquierdo).
* Ingreso bajo (azul) empuja hacia más ansiedad.

✅ Subjective\_Income (Percepción personal de ingreso):

* Quienes perciben su ingreso como bajo (azul) tienen mayor riesgo de ansiedad.

✅ Gender:

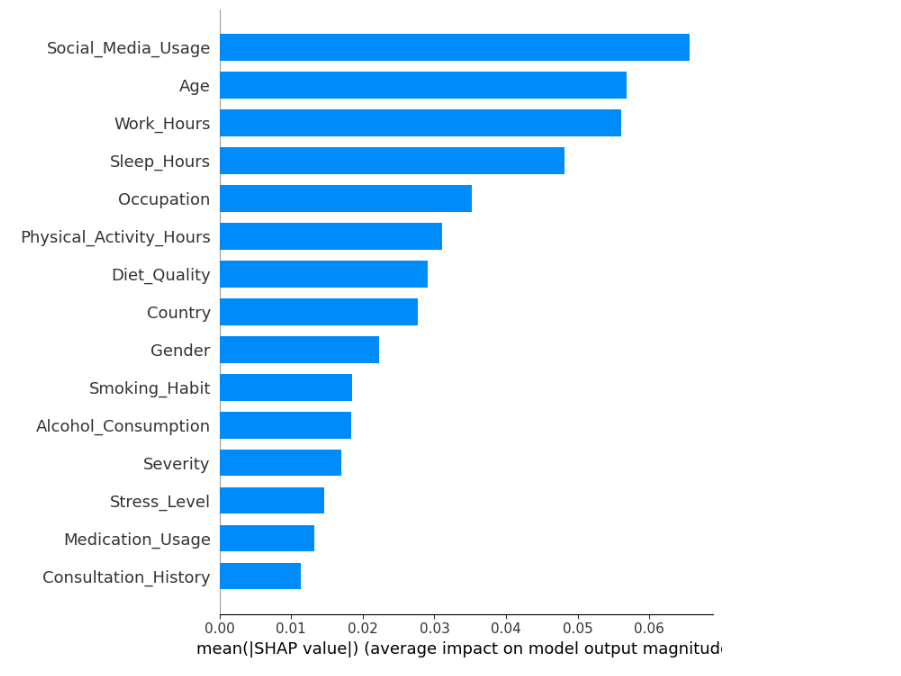
* Ciertos géneros (probablemente femenino) tienden a asociarse con mayor riesgo de ansiedad.

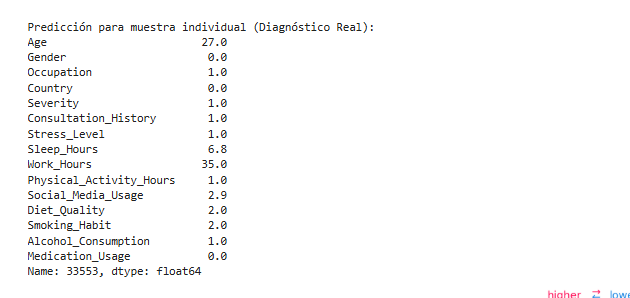
✅ Education:

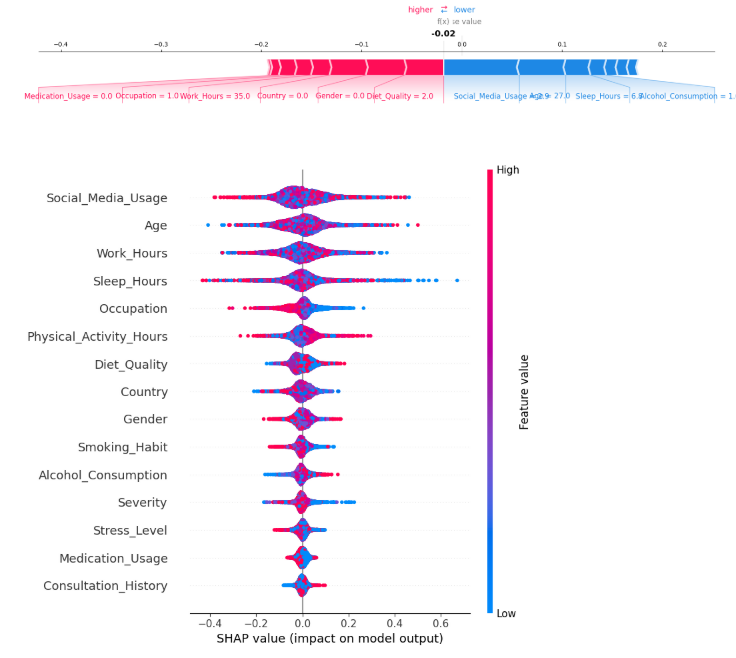
* Baja educación (azul) se asocia con más ansiedad.
* Alta educación (rojo) generalmente ayuda a reducir riesgo.

✅ Age:

* Los extremos (muy jóvenes o muy mayores) tienen diferente comportamiento, pero en general la edad baja puede favorecer ansiedad.







El análisis de interpretabilidad mediante SHAP en el dataset de diagnóstico real mostró que los factores de estilo de vida, particularmente el uso de redes sociales, las horas de trabajo semanal y las horas de sueño, son los principales determinantes de la probabilidad de sufrir problemas de salud mental.

Variables clínicas como la consulta médica previa, la severidad percibida o el uso de medicación tuvieron una influencia menor en la predicción del modelo.

A nivel individual, se pudo observar cómo aspectos como trabajar demasiadas horas o tener pocas horas de sueño contribuyen fuertemente a aumentar el riesgo de salud mental, mientras que una buena calidad del sueño y una percepción positiva de la situación financiera actúan como factores protectores.

Estos resultados destacan la importancia de intervenciones preventivas orientadas a modificar hábitos de vida para reducir la incidencia de problemas de salud mental.

|  |  |  |
| --- | --- | --- |
| Aspecto Analizado | Dataset de Percepciones (Wellcome Global Monitor) | Dataset de Diagnóstico Real (Mental Health Dataset) |
| Variable más importante | Región geográfica (Global11Regions) | Uso de redes sociales (Social\_Media\_Usage) |
| Segunda variable más importante | Ingreso país (wbi) | Edad |
| Tercera variable más importante | Ingreso percibido (Subjective\_Income) | Horas de trabajo (Work\_Hours) |
| Naturaleza de variables más relevantes | Socioeconómicas (país, ingresos, educación) | Estilo de vida (redes sociales, sueño, trabajo) |
| Variables clínicas importantes | No aplicaba (percepciones auto-reportadas) | Bajo peso relativo (medicación, historial de consultas, severidad) |
| Principales factores de riesgo detectados | Vivir en regiones vulnerables, baja percepción de ingresos, bajo nivel educativo | Uso excesivo de redes, pocas horas de sueño, exceso de trabajo |
| Factores protectores | Altos ingresos, buena percepción económica, mejor educación | Mejor calidad de sueño, menor uso de redes, actividad física moderada |

Aplicación de LIME como método de interpretabilidad local

Para fortalecer la transparencia del modelo, se aplicó la técnica de interpretabilidad LIME (Local Interpretable Model-Agnostic Explanations).  
Se generaron explicaciones individuales para casos seleccionados.

En un ejemplo concreto, se observó que factores como pertenecer a un país con alto nivel de ingresos (wbi), tener un nivel educativo superior y pertenecer a una región de menor riesgo (Global11Regions) contribuyeron de forma significativa a reducir la probabilidad de ansiedad o depresión.

Por otro lado, variables como la percepción subjetiva de ingresos bajos (Subjective\_Income) y la edad joven (≤25 años) incrementaron levemente el riesgo estimado.

La aplicación de LIME permitió validar que el modelo no solo realiza predicciones con alta precisión, sino que también genera explicaciones coherentes y alineadas con la literatura científica sobre factores de riesgo en salud mental.

MEMORIA

**Referencias Bibliográficas**

WHO. (2017). \*Depression and Other Common Mental Disorders: Global Health Estimates.\* World Health Organization. https://apps.who.int/iris/handle/10665/254610

Lund, C., Breen, A., Flisher, A. J., Kakuma, R., Corrigall, J., Joska, J. A., Swartz, L., & Patel, V. (2010). Poverty and common mental disorders in low and middle income countries: A systematic review. \*Social Science & Medicine\*, 71(3), 517–528. https://doi.org/10.1016/j.socscimed.2010.04.027

Patel, V., Burns, J. K., Dhingra, M., Tarver, L., Kohrt, B. A., & Lund, C. (2018). Income inequality and depression: A systematic review and meta-analysis of the association and a scoping review of mechanisms. \*World Psychiatry\*, 17(1), 76–89. https://doi.org/10.1002/wps.20492

Paul, K. I., & Moser, K. (2009). Unemployment impairs mental health: Meta-analyses. \*Journal of Vocational Behavior\*, 74(3), 264–282. https://doi.org/10.1016/j.jvb.2009.01.001

Modini, M., Joyce, S., Mykletun, A., Christensen, H., Bryant, R. A., Mitchell, P. B., & Harvey, S. B. (2016). The mental health benefits of employment: Results of a systematic meta-review. \*Australasian Psychiatry\*, 24(4), 331–336. https://doi.org/10.1177/1039856215618523

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. \*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining\*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. \*Advances in Neural Information Processing Systems (NeurIPS)\*, 4765–4774. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf