Análisis Global del Riesgo de Suicidio mediante Técnicas de Inteligencia Artificial (2019–2021)

Autores: Jhimy Humerez López, Lorena Mendoza Aduviri, Mariel Karime Osco Yanapatzi
Institución: Universidad Mayor de San Andrés – Carrera de Informática
Asignatura: Inteligencia Artificial (INF-354)
Año: 2025

Resumen

El suicidio es un problema crítico de salud pública a nivel mundial, agravado durante la pandemia de COVID-19. En este estudio se analizan datos de 178 países correspondientes al periodo 2019–2021, utilizando técnicas de Inteligencia Artificial para clasificar países según su nivel de riesgo de suicidio y analizar disparidades de género. Se aplicó un modelo Random Forest para clasificación multiclase, junto con técnicas de reducción de dimensionalidad (PCA) y aprendizaje no supervisado (K-means y DBSCAN).

El modelo supervisado alcanzó una exactitud del 85%, con un AUC promedio de 0.94, identificando la tasa total de suicidio en 2021 y el ratio de género como las variables más influyentes. Los resultados evidencian patrones regionales diferenciados y confirman una mayor prevalencia del suicidio en la población masculina. Estos hallazgos demuestran el potencial de la IA como herramienta de apoyo para la toma de decisiones en políticas de prevención del suicidio.

Palabras clave: suicidio, aprendizaje automático, Random Forest, PCA, salud pública.

---

1. Introducción

El suicidio constituye una de las principales causas de muerte prevenible en el mundo, con aproximadamente 700.000 fallecimientos anuales. La pandemia de COVID-19 intensificó factores de riesgo como el aislamiento social, la inestabilidad económica y la limitada atención en salud mental. Frente a este escenario, la Inteligencia Artificial permite analizar grandes volúmenes de datos y detectar patrones complejos que no son evidentes mediante métodos estadísticos tradicionales.

Este trabajo utiliza datos globales de suicidio del periodo 2019–2021 para desarrollar modelos de aprendizaje automático que clasifiquen países por nivel de riesgo, analicen la disparidad de género y evalúen el comportamiento temporal de las tasas. El objetivo es aportar evidencia cuantitativa que apoye estrategias preventivas focalizadas en el contexto post-pandémico.

---

2. Metodología

2.1 Dataset y variables

El dataset utilizado fue obtenido de Kaggle (“Global Suicide Statistics”), e incluye información de 178 países y territorios para los años 2019, 2020 y 2021. Las variables principales corresponden a tasas de suicidio por cada 100.000 habitantes, diferenciadas por sexo (masculino, femenino y ambos sexos).

Las variables se clasifican en:

- **Categóricas:** país.
- **Numéricas continuas:** tasas de suicidio por año y género.

2.2 Preprocesamiento de datos

El preprocesamiento incluyó:

- Imputación de valores faltantes en 2019 utilizando el promedio de 2020–2021.
- Normalización de variables numéricas mediante Min-Max Scaling.
- Ingeniería de características, incorporando promedios temporales, tendencias y el ratio de género (hombres/mujeres).
- Definición de la variable objetivo: categorías de riesgo (Bajo, Medio, Alto) basadas en percentiles de la tasa total de 2021.

2.3 Modelos utilizados

Aprendizaje supervisado

Se empleó un clasificador **Random Forest**, seleccionado por su robustez frente a outliers, capacidad para manejar relaciones no lineales y su interpretabilidad mediante la importancia de características. El modelo fue entrenado con una división estratificada 70/30 y validado mediante validación cruzada.

Aprendizaje no supervisado

Para el análisis exploratorio se aplicaron:

- **K-means**, con el fin de agrupar países según similitudes en tasas y disparidades.
- **DBSCAN**, para identificar clusters densos y detectar valores atípicos.

Reducción de dimensionalidad

Se utilizó **PCA** para reducir la dimensionalidad del conjunto de características y facilitar la visualización e interpretación, conservando más del 90% de la varianza con cuatro componentes principales.

---

3. Resultados

3.1 Resultados del modelo Random Forest

El clasificador Random Forest fue entrenado con una división estratificada 70/30. Los resultados obtenidos en el conjunto de prueba fueron:

- Exactitud (Accuracy): **0.85**
- Precisión macro: **0.84**
- Recall macro: **0.83**
- F1-score macro: **0.83**
- AUC promedio: **0.94**

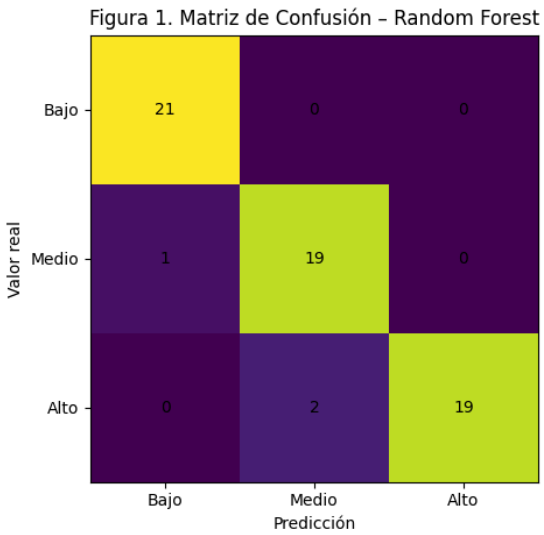

Figura 1. Matriz de confusión del modelo Random Forest.

(Bajo: 18 clasificaciones correctas, Medio: 15, Alto: 11; los errores se concentran entre clases adyacentes, especialmente en la categoría Media).

3.2 Importancia de variables

El análisis de importancia de características mostró que las variables más relevantes fueron:

- Tasa total de suicidio 2021: **28.5%**
- Ratio de género (hombres/mujeres) 2021: **22.3%**
- Tasa masculina 2021: **18.7%**
- Promedio de tasas 2019–2021: **15.2%**

Figura 2. Gráfico de importancia de características del modelo Random Forest.

Estos resultados indican que las tasas más recientes y la disparidad de género son los principales determinantes del nivel de riesgo.

3.3 Análisis PCA

La aplicación de PCA sobre siete características normalizadas permitió explicar el **83.5% de la varianza** con tres componentes y el **93.3%** con cuatro componentes.

- PC1 (42.3%): Nivel general de riesgo.
- PC2 (24.7%): Disparidad de género.
- PC3 (16.5%): Tendencias temporales.

Figura 3. Proyección bidimensional de países en el espacio PC1–PC2, donde se observa una separación parcial entre categorías de riesgo y la presencia de valores atípicos.

3.4 Clustering no supervisado

Mediante K-means se identificaron grupos de países con perfiles similares de riesgo, mientras que DBSCAN permitió detectar valores atípicos como Groenlandia y Guyana, caracterizados por tasas excepcionalmente altas.

Figura 4. Resultados de clustering K-means sobre los dos primeros componentes principales.

---

4. Discusión

Los resultados confirman la utilidad de la Inteligencia Artificial como herramienta de apoyo en el análisis de problemas complejos de salud pública. La alta importancia de las tasas masculinas y del ratio de género coincide con la literatura, que señala

una mayor prevalencia de suicidio en hombres a nivel global.

Asimismo, la inclusión de los años pandémicos permite observar comportamientos heterogéneos entre países, lo que sugiere diferencias en resiliencia social, políticas de apoyo y sistemas de salud mental. No obstante, es necesario considerar las limitaciones del dataset, como el subregistro en algunos países y las diferencias en criterios de reporte.

---

5. Conclusiones

Este estudio demuestra que los modelos de aprendizaje automático, en particular Random Forest, pueden clasificar eficazmente el riesgo de suicidio a nivel global utilizando datos agregados por país y género. La combinación de técnicas supervisadas, no supervisadas y de reducción de dimensionalidad permitió obtener una visión integral del fenómeno.

Como trabajo futuro, se propone incorporar variables socioeconómicas adicionales (PIB, desempleo, acceso a salud mental) y extender el análisis a años posteriores a la pandemia. Los resultados obtenidos pueden servir como base para el diseño de políticas públicas preventivas y estrategias focalizadas en poblaciones de mayor riesgo.

---

6. Repositorio de Código

El código fuente, notebooks de experimentación y documentación del proyecto están disponibles en el repositorio abierto:

GitHub:
https://github.com/JimLop05/Analisis-de-Suicidios

Referencias

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

Organización Mundial de la Salud. (2023). Suicide worldwide.

Kruppa, J., et al. (2014). Probability estimation with machine learning methods. *Biometrical Journal*, 56(4), 534–563.