INTELIGENCIA ARTIFICIAL

PROYECTO FINAL

Andrea Pardo Gispert

New Technology School

Tokio.

New Technology School

ANDREA PARDO GISPERT

RESPONSABLE DE PROYECTO:

Desarrollo de un Modelo Predictivo para Identificar el Riesgo de Suicidio

Fecha: 13 de agosto del 2024

INTELIGENCIA ARTIFICIAL
PROYECTO FINAL
Andrea Pardo Gispert

Tokio.

DESARROLLO

1. Define el problema y sus objetivos.
2. Define que datos vas a obtener y cárgalos.
3. Realiza la limpieza de los datos oportuna explicando el porqué de cada acción.
4. Explora los datos y comparte tus observaciones.
5. Escoge un modelo y justifícalo.
6. Realiza una representación final de los resultados que obtengas.

DESCRIPCIÓN DEL PROBLEMA Y SUS OBJETIVOS

1.1 Problema:

El objetivo de este proyecto es desarrollar un modelo predictivo para identificar individuos con alto riesgo de suicidio. Esto se basa en un análisis detallado de los factores demográficos asociados con el riesgo de suicidio utilizando el conjunto de datos proporcionado por la OMS.

1.2 Objetivos Específicos:

- Desarrollar un modelo predictivo eficiente.

- Identificar patrones clave en los datos que contribuyan al riesgo de suicidio.

- Asistir en la implementación de intervenciones preventivas efectivas.

DEFINICION DE DATOS QUE SE VAN A OBTENER Y CARGA

2.1 Datos a Obtener:

Se utilizarán los datos del conjunto "master.csv" (renombrado como "suicide_rates_df") obtenidos desde Kaggle, que incluye estadísticas de suicidios desde 1985 hasta 2016, cubriendo variables como tasas de suicidio, características demográficas y económicas, y datos de población.

2.2 Carga de los Datos:

Los datos se cargaron en el entorno de análisis utilizando bibliotecas de Python como Pandas para la manipulación de datos.

LIMPIEZA DE DATOS Y EXPLICACIÓN DE CADA ACCIÓN

3.1 Identificación de Problemas en los Datos:

- Valores nulos en la columna "HDI for year".

- Inconsistencias en los nombres de columnas y formatos.

3.2 Acciones de Limpieza:

- Manejo de Valores Nulos: Se exploraron métodos para imputar los valores faltantes en "HDI for year".

- Normalización de Nombres y Formatos: Se renombraron columnas para eliminar espacios y ajustar el formato de los nombres.

- Eliminación de Columnas Redundantes: Se consideró la eliminación de la columna "Country-year" al estar compuesta por otras dos columnas.

EXPLORACION DE DATOS Y OBSERVACIONES

4.1 Análisis Descriptivo:

Se realizó un análisis descriptivo de las variables, como la distribución de las tasas de suicidio por edad, género y país.

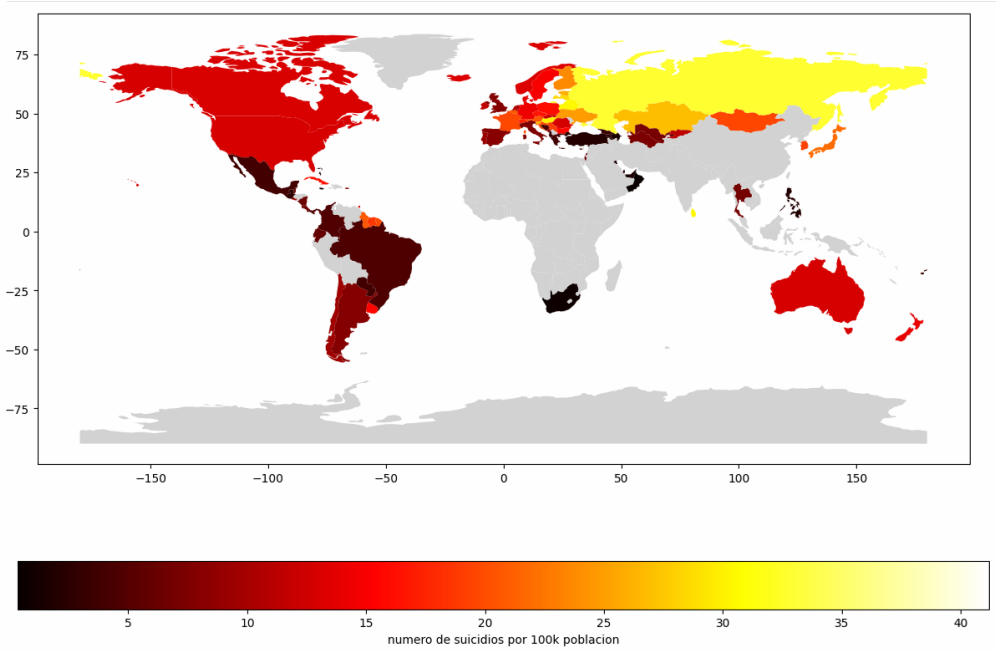
- **Descripción del Conjunto de Datos**

El conjunto de datos contiene 12 columnas en total, divididas en 6 columnas numéricas y 6 columnas categóricas. A continuación, describimos cada una de las columnas y sus características:

1. **Country:** Indica el país en el que se registró el suicidio.
2. **Year:** Año en que ocurrió el suicidio.
3. **Sex:** Género del individuo, categorizado como masculino o femenino.
4. **Age:** Rango de edad del individuo. Los rangos están definidos como 5-14, 15-24, 25-34, 35-54, 55-74, y 75+.
5. **Population:** Número total de personas en el grupo de edad específico para cada país y año.
6. **Suicides_no:** Número total de suicidios registrados en el grupo de edad especificado.
7. **HDI for year:** Índice de Desarrollo Humano (IDH) del año. Este índice compuesto mide logros medios en tres dimensiones básicas del desarrollo humano: vida larga y saludable, conocimientos, y nivel de vida decente.
8. **gdp_for_year:** Producto Interno Bruto (PIB) total del país para el año especificado. El PIB representa el valor monetario total de todos los bienes y servicios finales producidos dentro de un país.
9. **gdp_per_capita:** PIB dividido por la población total del país, proporcionando una medida del PIB per cápita.
10. **Suicides/100k pop:** Número de suicidios por cada 100,000 habitantes, ajustado por población para cada país y año.
11. **Country-year**: Combinación del país y el año. Esta columna puede ser redundante ya que la información está presente en las columnas "Country" y "Year" y podría eliminarse en el futuro.
12. **Generation:** Nombre de la generación asociada con el año de nacimiento del individuo

4.2 Visualización de Datos:

Gráficos de barras, histogramas, y mapas de calor se utilizaron para identificar patrones y correlaciones significativas.



Disponibilidad de Datos y Análisis Regional

- **Falta de Datos en África y Asia:** En muchas regiones de África y Asia, los datos sobre suicidio son escasos o inexistentes. Esta falta de información puede deberse a diversos factores, como limitaciones en la infraestructura de recolección de datos, estigmatización del suicidio, y la falta de sistemas de registro adecuados en algunas áreas.
- **Análisis Regional con Mapas:** Utilizando un mapa, podemos identificar los países con las tasas más altas de suicidio ajustadas por población. Este análisis geográfico permite visualizar las regiones con mayores incidencias de suicidio en relación con su tamaño poblacional. Aunque la falta de datos en algunas áreas limita la cobertura completa, los países con datos disponibles pueden revelar patrones significativos y ayudar a dirigir los esfuerzos de prevención hacia regiones con tasas preocupantes.

Análisis de Suicidio Representado en Gráficas

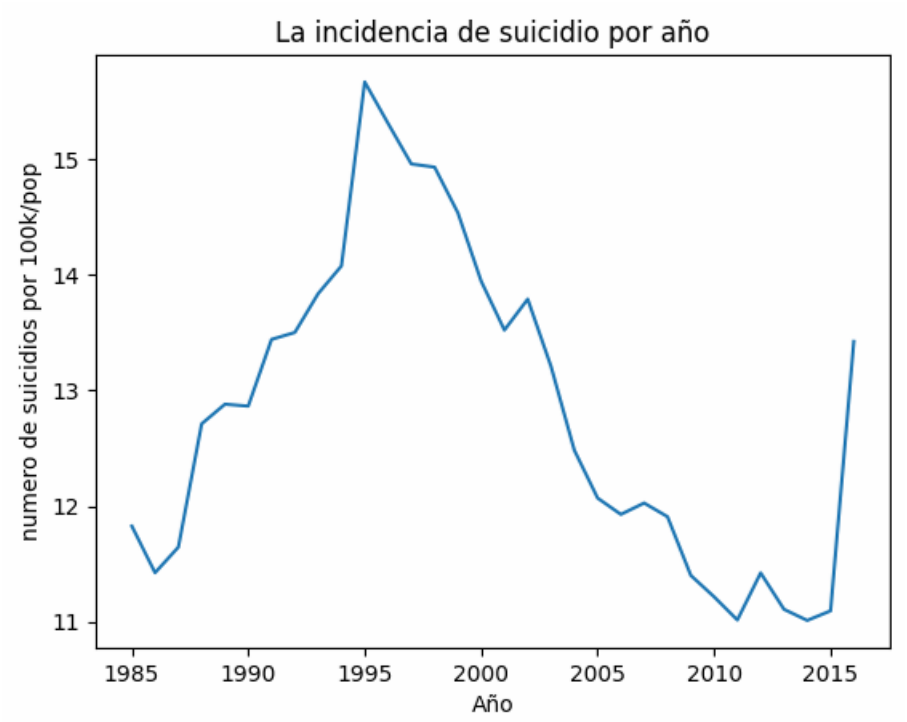
- **Suicidio a lo Largo del Tiempo:** Examinaremos cómo las tasas de suicidio han cambiado a lo largo de los años. Esta gráfica mostrará las tendencias temporales, permitiéndonos identificar aumentos o disminuciones en las tasas de suicidio a lo largo de los años. También puede resaltar eventos o periodos específicos que han influido en estas tendencias.

- **Suicidio por Edad y Género:** Esta gráfica desglosa las tasas de suicidio según grupos de edad y género.

Permite observar las diferencias en las tasas de suicidio entre hombres y mujeres en diversos rangos de edad. Es útil para identificar patrones de riesgo específicos para cada grupo etario y de género, ayudando a focalizar las estrategias de prevención y apoyo.

- **Correlaciones entre Variables Categóricas:** Analizaremos las relaciones entre diferentes variables categóricas relacionadas con el suicidio. Esto incluye la identificación de correlaciones entre factores como ubicación geográfica, estado socioeconómico, y otras variables relevantes.

Esta gráfica ayuda a comprender cómo estas variables se interrelacionan y cómo pueden influir en las tasas de suicidio, proporcionando una visión más completa de los factores contribuyentes.

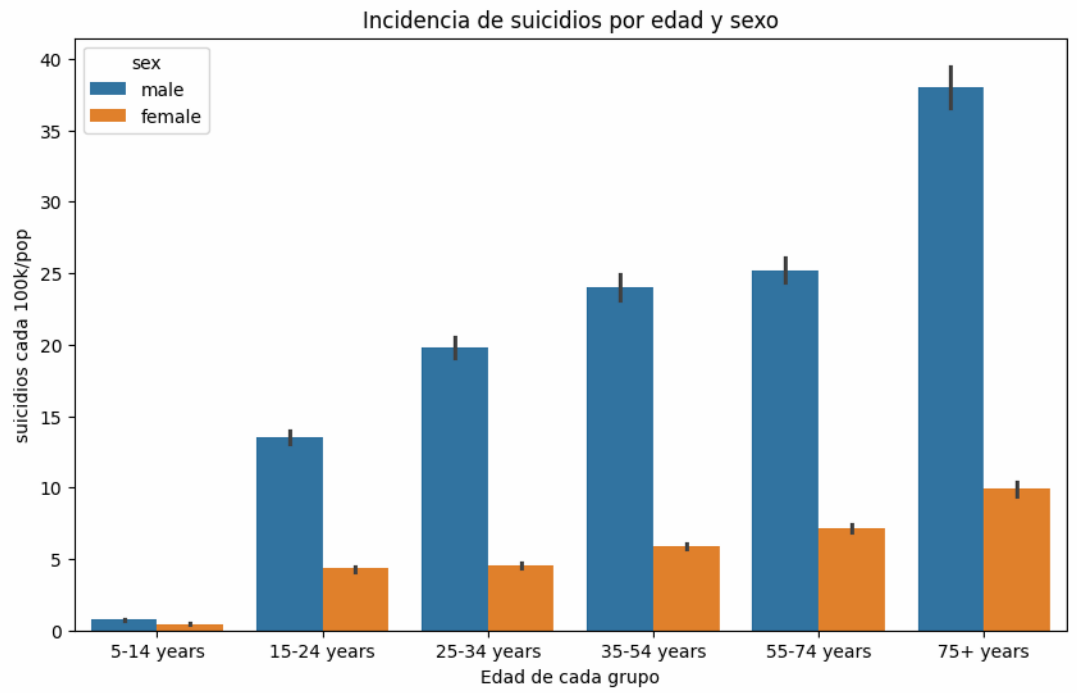


Análisis de la Tendencia de las Tasas de Suicidio

La gráfica ilustra la evolución de las tasas de suicidio desde 1985:

- **Incremento Inicial (1985-1995):** Las tasas de suicidio aumentaron notablemente desde 1985, alcanzando su punto máximo en 1995. Este aumento puede estar relacionado con eventos globales significativos de la época, como la recesión económica global a principios de los años 90 y las crisis económicas en varias regiones, que podrían haber exacerbado el estrés y la inseguridad económica.

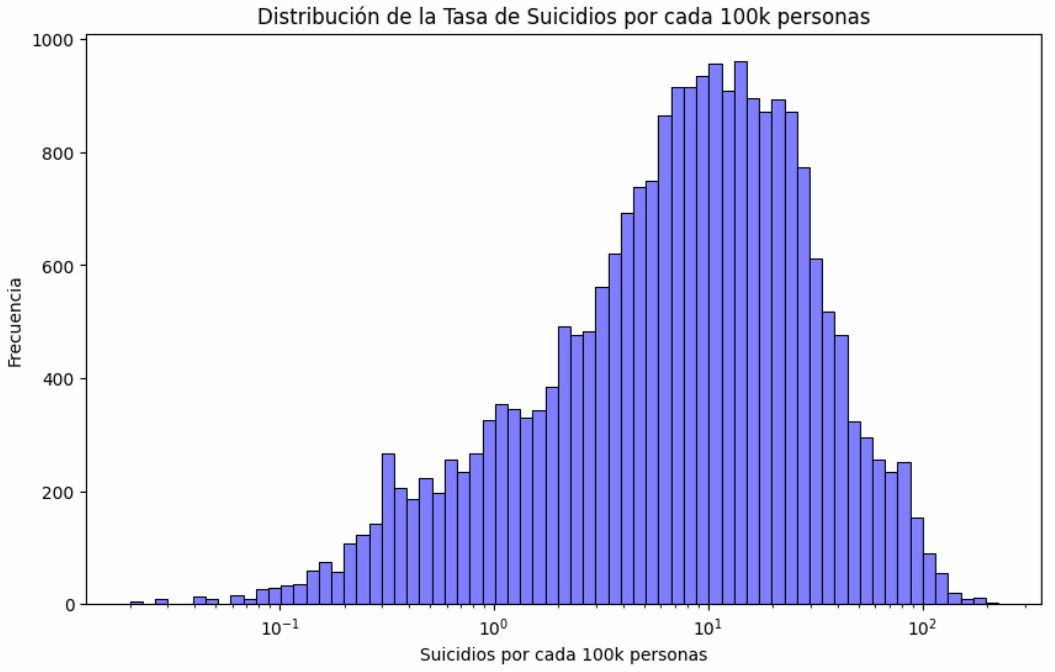
- **Descenso Posterior (1995-2011):** Después del pico en 1995, las tasas de suicidio comenzaron a disminuir, alcanzando su nivel más bajo en 2011. Este período vio importantes mejoras económicas y sociales en muchas regiones, así como el incremento de las inversiones en salud mental y programas de prevención del suicidio, lo que podría haber contribuido a la reducción de las tasas.
- **Tendencia Reciente (2015 en adelante):** A partir de 2015, las tasas de suicidio empiezan a subir nuevamente. Esta tendencia podría estar relacionada con varios factores contemporáneos, como el aumento de la incertidumbre económica global, el impacto de la pandemia de COVID-19 que comenzó en 2020, y el creciente estrés social y económico asociado con la crisis sanitaria global. Sin embargo, no se dispone de datos más allá de 2015 para confirmar la continuidad de esta tenmporáneas.



Envejecimiento, Masculinidad y Salud:

- **Aislamiento Social:** A medida que las personas envejecen, es más probable que enfrenten aislamiento social debido a la jubilación, la pérdida de amigos y seres queridos, o la movilidad reducida. El aislamiento social es un factor de riesgo conocido para la depresión y el suicidio.
- **Problemas de Salud:** Las enfermedades crónicas, el dolor físico y la disminución de la calidad de vida también aumentan con la edad, lo que puede llevar a sentimientos de desesperanza o inutilidad. La generación G.I., en particular, habría enfrentado estos problemas en una era en la que los avances médicos eran menos eficaces en el manejo del dolor y las enfermedades crónicas.

- **Dependencia y Pérdida de Autonomía:** La pérdida de independencia, ya sea por discapacidad, enfermedades o la necesidad de cuidados a largo plazo, puede ser devastadora para personas que valoraban su autonomía y capacidad de cuidar de sí mismos.
- **Métodos Letales y Presión Masculina:** Los hombres tienen tasas de suicidio 3-4 veces mayores que las mujeres, en parte debido al uso de métodos más violentos y letales, y a la presión social para no mostrar vulnerabilidad, lo que dificulta que pidan ayuda.



El histograma muestra la distribución de la tasa de suicidios por cada 100,000 personas.

Observaciones:

1. Distribución:

La distribución de la tasa de suicidios tiene una forma asimétrica, con un sesgo hacia la derecha. La mayor parte de los datos se concentra en un rango específico, mientras que hay una larga cola hacia la derecha, indicando la presencia de algunas tasas de suicidio relativamente altas.

La tasa de suicidios más común se encuentra en el rango aproximado de (10^1) (10 a 20 suicidios por cada 100,000 personas), lo que es consistente con la moda de la distribución.

2. Frecuencia Alta:

El pico del histograma se encuentra alrededor de (10^1) (entre 10 y 20 suicidios por cada 100,000 personas), con una frecuencia que alcanza cerca de 1000 casos.

Este rango representa la tasa de suicidios más común en la muestra, lo que sugiere que es una tasa prevalente en muchas poblaciones.

3. Tasa Baja de Suicidios:

A medida que la tasa de suicidios disminuye hacia el rango de (10^{-1}) (menos de 1 suicidio por cada 100,000 personas), la frecuencia disminuye considerablemente, indicando que es poco común encontrar poblaciones con tasas de suicidio extremadamente bajas.

4. Tasa Alta de Suicidios:

Hay una caída gradual en la frecuencia a medida que la tasa de suicidios aumenta por encima de (10^1). Sin embargo, hay algunos casos (outliers) con tasas extremadamente altas, hasta el rango de más de 100 suicidios por cada 100,000 personas, aunque son mucho menos frecuentes.

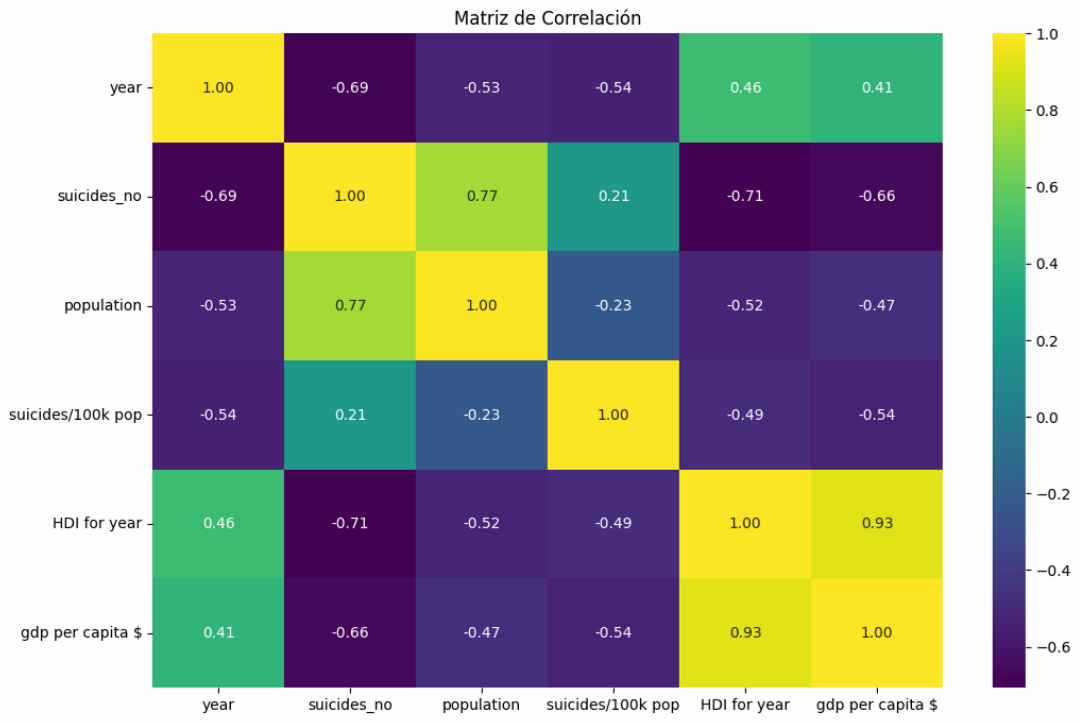
Conclusiones:

- **Tendencia General:** La mayoría de las poblaciones estudiadas tienen una tasa de suicidio entre 10 y 20 suicidios por cada 100,000 personas. Esta es la tasa más común y representa el centro de la distribución.
- **Extremos:** Existen poblaciones con tasas de suicidio extremadamente altas y bajas, pero estas son menos comunes. La presencia de outliers con tasas de suicidio muy altas sugiere que ciertos factores específicos pueden estar influyendo fuertemente en algunas poblaciones.
- **Forma de la Distribución:** La asimetría de la distribución indica que aunque la mayoría de las poblaciones tienen tasas de suicidio moderadas, un número reducido de poblaciones experimenta tasas mucho más altas, lo que podría ser un punto de interés para estudios adicionales.

Este histograma proporciona una visión general de cómo se distribuyen las tasas de suicidio entre diferentes poblaciones, destacando tanto las tendencias comunes como los extremos que podrían merecer una atención más detallada.

4.3 Análisis de Correlación:

Se destacaron las correlaciones más relevantes entre las variables demográficas y las tasas de suicidio, identificando factores de alto riesgo.



Observaciones:

1. Correlación entre suicides_no (número de suicidios) y population:

Hay una correlación positiva alta (0.77) entre el número de suicidios y la población. Esto es esperado, ya que en poblaciones más grandes es más probable que haya un mayor número absoluto de suicidios.

2. Correlación entre suicides_no y suicides/100k pop (tasa de suicidios por cada 100k personas):

Existe una correlación positiva baja (0.21) entre el número de suicidios y la tasa de suicidios por cada 100k personas, lo que indica que el número absoluto de suicidios no siempre se traduce directamente en una alta tasa de suicidios.

3. Correlación entre gdp per capita $ (PIB per cápita) y HDI for year (IDH por año):

La correlación entre el PIB per cápita y el IDH es muy alta (0.93), lo que sugiere que las economías más ricas suelen tener un Índice de Desarrollo Humano más alto.

4. Correlación entre suicides/100k pop y otras variables:

La tasa de suicidios por cada 100k personas tiene una correlación negativa moderada con el PIB per cápita (-0.54), el IDH (-0.49), la población (-0.23), y el año (-0.54). Esto podría indicar que a medida que las condiciones socioeconómicas mejoran, la tasa de suicidios tiende a disminuir, aunque la relación no es extremadamente fuerte.

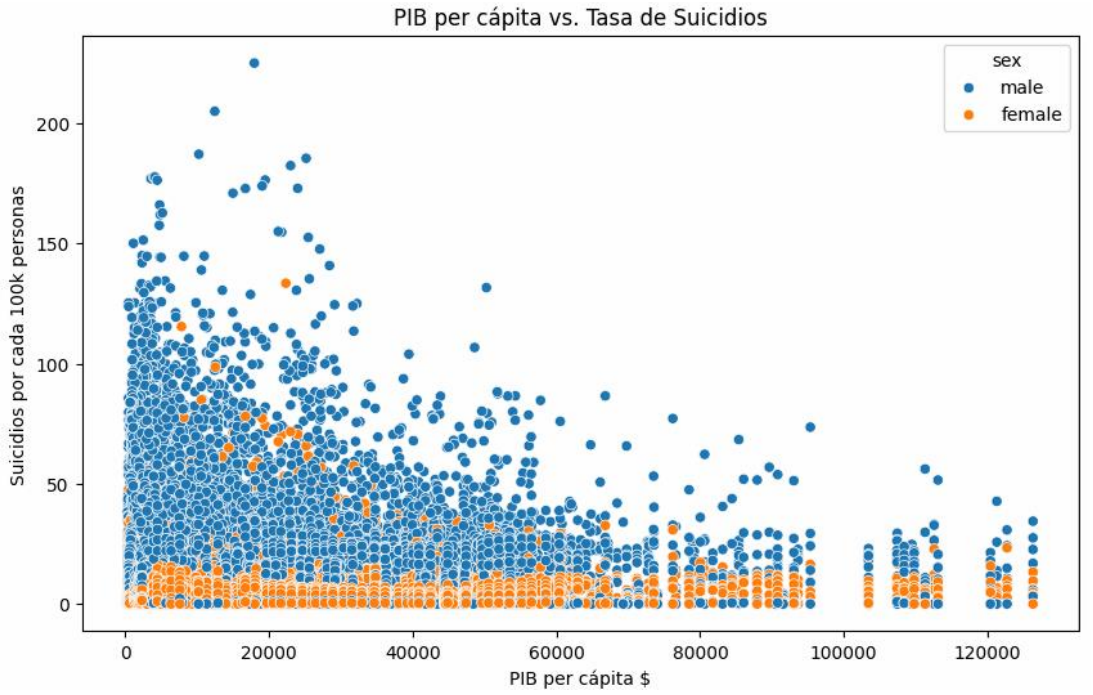
5. Correlación entre suicides_no y gdp per capita $:

Hay una correlación negativa moderada (-0.66) entre el número de suicidios y el PIB per cápita, lo que sugiere que en países con un mayor PIB per cápita, el número total de suicidios tiende a ser menor.

Conclusiones:

- **Impacto de las Condiciones Socioeconómicas:** El PIB per cápita y el IDH tienen una correlación negativa moderada con la tasa de suicidios, lo que sugiere que en países con mejor desarrollo económico y humano, las tasas de suicidio tienden a ser más bajas. Sin embargo, esta correlación no es lo suficientemente fuerte como para ser determinante por sí sola.
- **Relación Población-Suicidios:** La alta correlación entre la población y el número de suicidios indica que las cifras absolutas de suicidios están fuertemente influenciadas por el tamaño de la población, pero esta relación no se refleja de la misma manera en la tasa de suicidios por cada 100k personas.
- **Año y Condiciones de Vida:** La correlación negativa del año con variables como el número de suicidios, la población y la tasa de suicidios sugiere que con el paso del tiempo, podría haber una tendencia general a la mejora de las condiciones de vida y una posible reducción en los suicidios, aunque esto requiere un análisis más detallado.

En resumen, el análisis muestra que existen varias relaciones entre las condiciones socioeconómicas y las tasas de suicidios, pero estas relaciones no son extremadamente fuertes, lo que sugiere que otros factores también pueden estar influyendo en las tasas de suicidios en diferentes países y contextos.



El gráfico muestra la relación entre el PIB per cápita y la tasa de suicidios por cada 100,000 personas, diferenciada por sexo (masculino y femenino).

1. Distribución General:

Hay una tendencia general a la baja en la tasa de suicidios a medida que aumenta el PIB per cápita. Esto sugiere que los países o regiones con mayor riqueza por persona tienden a tener menores tasas de suicidio.

Sin embargo, incluso en los niveles más altos de PIB per cápita, todavía hay algunas tasas significativas de suicidios, lo que indica que la riqueza no es el único factor que influye en esta tasa.

2. Diferencias por Sexo:

Los puntos azules (hombres) son predominantemente más numerosos y se encuentran a niveles más altos de tasa de suicidio en comparación con los puntos naranjas (mujeres). Esto sugiere que, en general, la tasa de suicidios es más alta en hombres que en mujeres.

La mayoría de los puntos naranjas (mujeres) se concentran en tasas de suicidio más bajas, independientemente del nivel de PIB per cápita.

3. Tendencias Específicas:

En niveles muy bajos de PIB per cápita, menos de $ 20.000, las tasas de suicidio son muy variadas, pero parecen concentrarse en niveles más altos.

A medida que el PIB per cápita aumenta, especialmente por encima de $ 40.000, las tasas de suicidio disminuyen y se vuelven menos dispersas.

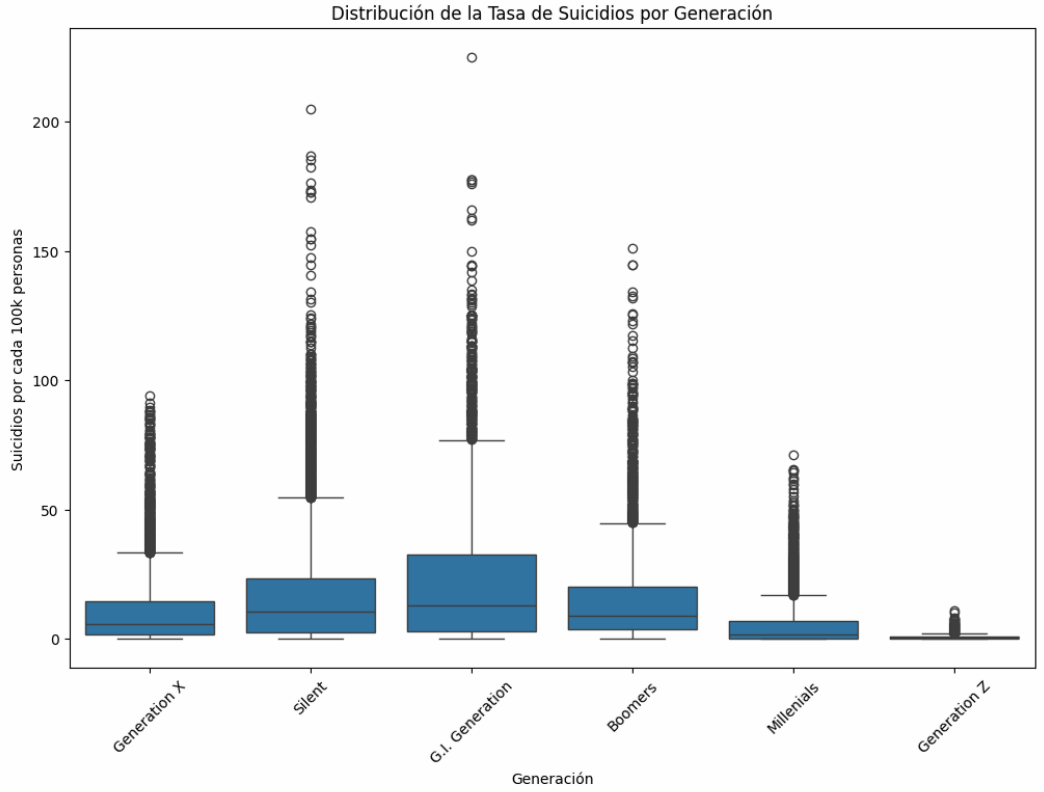
Hay menos datos para niveles extremadamente altos de PIB per cápita, por encima de $ 100.000, pero aún así, las tasas de suicidio tienden a ser bajas.

4. Outliers:

Existen algunos puntos que se apartan significativamente de la tendencia general, con tasas de suicidio extremadamente altas incluso a niveles de PIB per cápita más altos.

Conclusión:

El gráfico sugiere que hay una correlación inversa entre el PIB per cápita y la tasa de suicidios: a mayor PIB per cápita, generalmente hay menores tasas de suicidio. Además, hay una notable diferencia en las tasas de suicidio entre hombres y mujeres, siendo los hombres los que tienen una tasa significativamente mayor. Sin embargo, la relación no es perfecta, ya que hay excepciones y otros factores que también pueden influir en las tasas de suicidio.



1. Generación Silenciosa (Silent Generation)

Años de nacimiento: 1928-1945

Características: Crecieron durante la Gran Depresión y la Segunda Guerra Mundial. Valoran la estabilidad, la disciplina, el trabajo duro y la conformidad social. Tienden a ser conservadores y respetuosos de la autoridad. Son conocidos por su lealtad y dedicación a sus empleadores.

2. Generación G.I. (G.I. Generation)

Años de nacimiento: 1901-1927

Características: También conocida como la "Generación de la Segunda Guerra Mundial". Vivieron durante la Primera Guerra Mundial, la Gran Depresión y la Segunda Guerra Mundial. Se les atribuye la construcción de instituciones fuertes y un fuerte sentido de deber cívico y patriotismo. Fueron testigos del auge industrial y del crecimiento económico después de la guerra.

3. Baby Boomers (Boomers)

Años de nacimiento: 1946-1964

Características: Nacidos en la época de prosperidad económica posterior a la Segunda Guerra Mundial. Experimentaron grandes cambios sociales, como los movimientos de derechos civiles y la contracultura de los años 60. Son conocidos por su enfoque en la autosuficiencia, el éxito profesional y la revolución social.

4. Generación X

Años de nacimiento: 1965-1980

Características: Crecieron en un período de incertidumbre económica, aumento del divorcio y avances tecnológicos. Son independientes, adaptables y valoran el equilibrio entre la vida laboral y personal. A menudo se les ve como escépticos y desilusionados, pero también son conocidos por su creatividad y pragmatismo.

5. Millennials (Generación Y)

Años de nacimiento: 1981-1996

Características: Primeros nativos digitales, crecieron con el auge de Internet y las redes sociales. Son diversos, educados y tienen una mentalidad global. Valoran la autenticidad, la responsabilidad social y buscan un propósito en su trabajo. Son más abiertos a cambios sociales y son flexibles en sus enfoques de la vida y la carrera.

6. Generación Z

Años de nacimiento: 1997-2012

Características: Completamente inmersos en la tecnología desde su nacimiento, con acceso a información constante a través de dispositivos móviles y redes sociales. Son pragmáticos, diversos y valoran la individualidad. Se preocupan por la estabilidad financiera y el bienestar mental, y tienen un enfoque más consciente hacia temas bservados en tus gráficos.

Existen varias razones que podrían explicar por qué las tasas de suicidio son más altas en la Generación G.I., seguida por la Generación Silenciosa y los Baby Boomers:

1. Eventos Traumáticos y Experiencias de Vida:

Generación G.I.: Esta generación vivió durante la Primera Guerra Mundial, la Gran Depresión y la Segunda Guerra Mundial. Estos eventos traumáticos y la exposición a violencia, pobreza extrema y desarraigo social pueden haber tenido un impacto duradero en su salud mental. Además, la experiencia de la guerra y la pérdida de compañeros podrían haber aumentado el riesgo de trastornos de salud mental, como el trastorno de estrés postraumático (TEPT), que a menudo se asocia con tasas más altas de suicidio.

Generación Silenciosa: Crecieron en la sombra de la Gran Depresión y experimentaron la Segunda Guerra Mundial en su juventud. También vivieron la Guerra de Corea y el inicio de la Guerra Fría. Las expectativas de conformidad social, junto con la supresión de emociones y el estigma asociado a los problemas de salud mental, pueden haber contribuido al aumento de las tasas de suicidio en esta generación.

Boomers: Aunque crecieron en una época de relativa prosperidad, también enfrentaron cambios sociales y económicos importantes, como la Guerra de Vietnam, los movimientos de derechos civiles, y la crisis económica en la década de 1970. Estos factores, junto con el envejecimiento y la pérdida

de un propósito o rol social a medida que envejecen, podrían contribuir a un mayor riesgo de suicid mismos.

3. Estigma y Falta de Atención a la Salud Mental:

Generación G.I. y Silenciosa: En estas generaciones, buscar ayuda para problemas de salud mental era a menudo estigmatizado, y el suicidio era un tema tabú. La falta de acceso a servicios de salud mental adecuados y la reticencia a buscar ayuda podrían haber contribuido a tasas más altas de suicidio.

Boomers: Aunque los Baby Boomers crecieron en una época de mayor conciencia sobre la salud mental, muchos aún pueden haber internalizado actitudes estigmatizantes hacia el tratamiento de la salud mental, lo que podría llevar a un menor uso de los servicios disponibles.

4. Cambio de Roles y Desarraigo:

A medida que estas generaciones envejecen, experimentan cambios significativos en sus roles sociales, como la jubilación, la pérdida del rol de proveedor, o la separación de los hijos. Estos cambios pueden llevar a una crisis de identidad y sentimientos de inutilidad o pérdida de propósito, factores que pueden incrementar el riesgo de suicidio.

5. Acceso a Métodos Letales:

En generaciones mayores, particularmente en áreas rurales o en poblaciones que fueron militares, puede haber un mayor acceso a armas de fuego, que es un método de suicidio altamente letal. Esto podría contribuir a las tasas más altas observadas en estas generaciones.

Estas razones, combinadas, podrían explicar por qué las generaciones mayores como la G.I., la Silenciosa y los Baby Boomers muestran tasas más altas de suicidio en comparación con gene medida que envejecen.

SELECCIÓN DEL MODELO Y JUSTIFICACIÓN

5.1 Modelos Considerados:

Se evaluaron varios modelos de clasificación, como:

- **Regresión Logística**

```
> Accuracy: 0.8907156673114119
> Precisión promedio: 0.8642045454545455
> Recall promedio: 0.8805437495783579
> F1 Score promedio: 0.8715958481463368
> Matriz de confusión:
[[1566  264]
[ 414 3960]]
> Reporte de clasificación:
              precision    recall  f1-score   support

        alto       0.79      0.86      0.82      1830
        bajo       0.94      0.91      0.92      4374

    accuracy                           0.89      6204
   macro avg       0.86      0.88      0.87      6204
weighted avg       0.89      0.89      0.89      6204

>>> Métrica final Logistic Regression: 87.16
```

- **Máquinas de Vectores de Soporte (SVM)**

```
> Accuracy: 0.9028046421663443
> Precisión promedio: 0.8786697979967424
> Recall promedio: 0.8934077672086171
> F1 Score promedio: 0.8854466971963559
> Matriz de confusión:
[[1593  237]
[ 366 4008]]
> Reporte de clasificación:
              precision    recall  f1-score   support

        alto       0.81      0.87      0.84      1830
        bajo       0.94      0.92      0.93      4374

    accuracy                           0.90      6204
   macro avg       0.88      0.89      0.89      6204
weighted avg       0.91      0.90      0.90      6204

>>> Métrica final Logistic Regression: 88.54
```

- **Bosques Aleatorios**

```
> Accuracy: 0.9024822695035462
> Precisión promedio: 0.8754673360287558
> Recall promedio: 0.9039851731918115
> F1 Score promedio: 0.8873012239915421
> Matriz de confusión:
[[1661  169]
[ 436 3938]]
> Reporte de clasificación:
              precision    recall  f1-score   support

        alto       0.79      0.91      0.85      1830
        bajo       0.96      0.90      0.93      4374

    accuracy                           0.90      6204
   macro avg       0.88      0.90      0.89      6204
weighted avg       0.91      0.90      0.90      6204

>>> Métrica final Logistic Regression: 88.73
```

- **K-Vecinos Más Cercanos (KNN)**

```
> Accuracy: 0.9160219213410703
> Precisión promedio: 0.8932319527640954
> Recall promedio: 0.9112036849640573
> F1 Score promedio: 0.9013727722144642
> Matriz de confusión:
[[1646  184]
[ 337 4037]]
> Reporte de clasificación:
              precision    recall  f1-score   support

        alto       0.83      0.90      0.86      1830
        bajo       0.96      0.92      0.94      4374

    accuracy                           0.92      6204
   macro avg       0.89      0.91      0.90      6204
weighted avg       0.92      0.92      0.92      6204

>>> Métrica final Logistic Regression: 90.14
```

- **Naive Bayes**

```
> Accuracy: 0.634107027724049
> Precisión promedio: 0.7033206442560105
> Recall promedio: 0.7270045799695668
> F1 Score promedio: 0.632223350381067
> Matriz de confusión:
[[1745   85]
[2185 2189]]
> Reporte de clasificación:
              precision    recall  f1-score   support

       alto       0.44      0.95      0.61      1830
       bajo       0.96      0.50      0.66      4374

   accuracy                           0.63      6204
  macro avg       0.70      0.73      0.63      6204
weighted avg       0.81      0.63      0.64      6204

>>> Métrica final Logistic Regression: 63.22
```

- **Perceptrón**

```
> Accuracy: 0.8480012894906512
> Precisión promedio: 0.8830444315952544
> Recall promedio: 0.7528379320425465
> F1 Score promedio: 0.7851780912770556
> Matriz de confusión:
[[ 953  877]
[  66 4308]]
> Reporte de clasificación:
              precision    recall  f1-score   support

       alto       0.94      0.52      0.67      1830
       bajo       0.83      0.98      0.90      4374

   accuracy                           0.85      6204
  macro avg       0.88      0.75      0.79      6204
weighted avg       0.86      0.85      0.83      6204

>>> Métrica final Logistic Regression: 78.52
```

- **Linear SVC**

```
> Accuracy: 0.8934558349451966
> Precisión promedio: 0.8701591364241834
> Recall promedio: 0.8758127384619998
> F1 Score promedio: 0.872899551089924
> Matriz de confusión:
[[1524  306]
 [ 355 4019]]
> Reporte de clasificación:
              precision    recall  f1-score   support

        alto       0.81      0.83      0.82      1830
        bajo       0.93      0.92      0.92      4374

    accuracy                           0.89      6204
   macro avg       0.87      0.88      0.87      6204
weighted avg       0.89      0.89      0.89      6204

>>> Métrica final Logistic Regression: 87.29
```

- **Clasificador de Descenso de Gradiente Estocástico**

```
> Accuracy: 0.8891038039974211
> Precisión promedio: 0.8626112704115709
> Recall promedio: 0.8776525969401906
> F1 Score promedio: 0.8694724712988497
> Matriz de confusión:
[[1555  275]
 [ 413 3961]]
> Reporte de clasificación:
              precision    recall  f1-score   support

        alto       0.79      0.85      0.82      1830
        bajo       0.94      0.91      0.92      4374

    accuracy                           0.89      6204
   macro avg       0.86      0.88      0.87      6204
weighted avg       0.89      0.89      0.89      6204

>>> Métrica final Logistic Regression: 86.95
```

- **Árboles de Decisión**

```
> Accuracy: 0.8662153449387492
> Precisión promedio: 0.8348337329111839
> Recall promedio: 0.8657109447030515
> F1 Score promedio: 0.8467715835697925
> Matriz de confusión:
[[1582  248]
[ 582 3792]]
> Reporte de clasificación:
              precision    recall  f1-score   support

        alto       0.73      0.86      0.79      1830
        bajo       0.94      0.87      0.90      4374

    accuracy                           0.87      6204
   macro avg       0.83      0.87      0.85      6204
weighted avg       0.88      0.87      0.87      6204

 >>> Métrica final Logistic Regression: 84.68
```

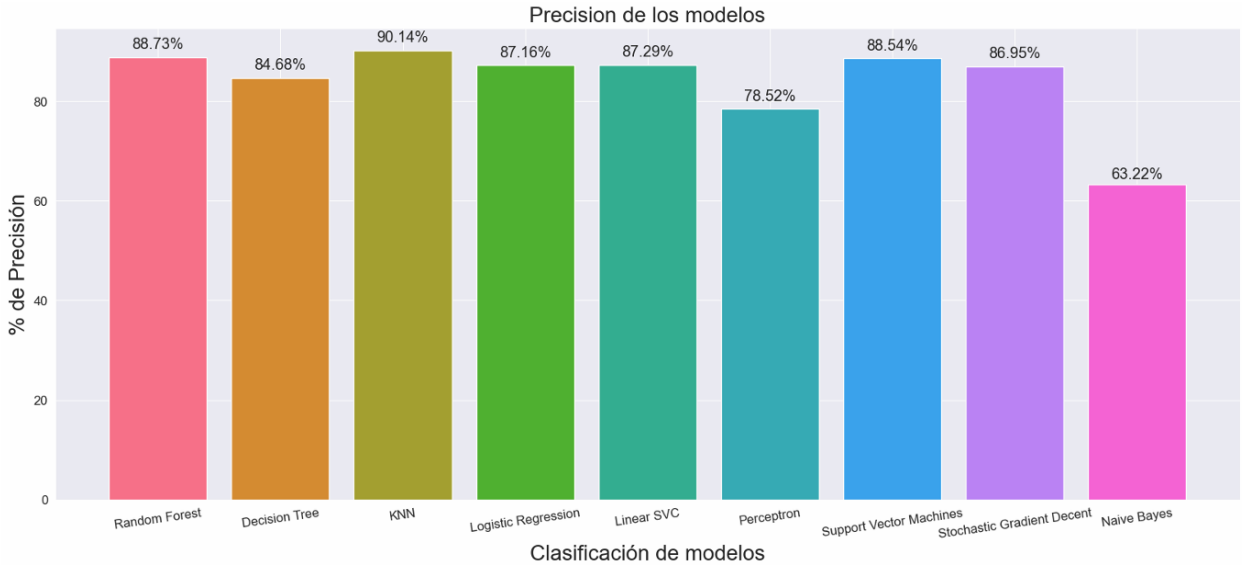
5.2 Justificación del Modelo Escogido:

Los modelos fueron seleccionados por su capacidad de manejar tareas de clasificación binaria y su robustez en la evaluación del riesgo de suicidio. Se escogió el modelo final basado en su desempeño en métricas clave como precisión, recall, F1-score y ROC-AUC.

REPRESENTACIÓN FINAL DE LOS RESULTADOS OBTENIDOS

6.1 Dashboard :

Se desarrolló un dashboard para visualizar las predicciones y probabilidades de riesgo de suicidio, permitiendo una interpretación clara y accesible.



6.2 Gráficos de Importancia:

Se mostraron los gráficos de importancia de características para identificar los factores que más impactan en la predicción.

6.3 Tasas de Predicción por Grupos Demográficos:

Se presentaron las tasas de predicción diferenciadas por grupos demográficos, destacando áreas de mayor riesgo.

CONCLUSIONES DEL PROYECTO

1. Resumen de Modelos y Rendimiento

En el proceso de evaluación de modelos de machine learning para la predicción del riesgo de suicidio, se utilizaron diversos algoritmos para determinar su eficacia. Los modelos evaluados y sus respectivos scores son los siguientes:

1. **KNN (K-Nearest Neighbors)**: 90.14
2. **Random Forest**: 88.98
3. **Support Vector Machines**: 88.54
4. **Linear SVC**: 87.29
5. **Logistic Regression**: 87.16
6. **Stochastic Gradient Descent**: 86.46
7. **Decision Tree**: 84.55
8. **Perceptron**: 78.52
9. **Naive Bayes**: 63.22

El **KNN** se destacó como el mejor modelo con un score de 90.14, seguido por **Random Forest** y **Support Vector Machines**. Estos resultados sugieren que los modelos basados en vecinos más cercanos y en bosques aleatorios son los más efectivos para esta tarea en particular.

2. Análisis de Predicciones

Para evaluar la precisión de las predicciones finales, se compararon las predicciones del modelo con los valores reales del conjunto de validación. Se generó un DataFrame con las siguientes columnas:

- **Suicide Risk Predicción**: Resultado de las predicciones del modelo.
- **Suicide Risk Real**: Valores reales del conjunto de validación.
- **Diferencia**: Indicador booleano que marca si la predicción difiere del valor real.

Se guardaron los resultados en dos formatos:

- **CSV**: Se utilizó para una revisión rápida y almacenamiento.
- **Excel**: Se aplicó un estilo para resaltar las diferencias en las predicciones.

El DataFrame muestra que la mayoría de las predicciones coinciden con los valores reales. De los 62,204 registros, 591 predicciones fueron incorrectas. Esto representa aproximadamente el 0.95% de las predicciones totales.

3. Observaciones Detalladas

La tabla de diferencias revela que las predicciones incorrectas tienden a agruparse, con algunos registros importantes donde el modelo predice "bajo" cuando el valor real es "alto", y viceversa. Estos errores deben ser analizados más a fondo para entender las posibles causas, tales como:

- **Datos Desbalanceados**: Puede ser que el modelo haya tenido dificultades para aprender patrones en clases menos representadas.
- **Características del Modelo**: Algunos modelos pueden no haber capturado adecuadamente las características del conjunto de datos.

4. Recomendaciones

1. **Optimización del Modelo**: Considerar ajustes en los parámetros del modelo y pruebas con técnicas de regularización para mejorar la precisión de las predicciones.
2. **Análisis de Errores**: Realizar un análisis detallado de las predicciones incorrectas para identificar patrones o características específicas que el modelo podría estar pasando por alto.
3. **Ampliación de Datos**: Investigar si el conjunto de datos puede ser enriquecido con más características o ejemplos para mejorar la capacidad predictiva del modelo.
4. **Validación Cruzada**: Implementar técnicas de validación cruzada para asegurar que el rendimiento del modelo sea consistente en diferentes particiones del conjunto de datos.

En conclusión, aunque el **KNN** ha demostrado ser el mejor modelo, siempre hay espacio para la mejora. Las recomendaciones proporcionadas buscan mejorar la precisión y robustez del sistema de predicción, con el objetivo de minimizar los errores y asegurar que las predicciones sean lo más precisas posible.

CONCLUSIONES SOBRE LOS DATOS DEL DATAFRAME

1. Resumen del Conjunto de Datos y Modelos Entrenados

El análisis y modelado del riesgo de suicidio se han llevado a cabo utilizando un conjunto de datos que incluye factores como el grupo de edad, el sexo, el Índice de Desarrollo Humano (IDH) anual, el PIB per cápita, y el país. A partir de estos datos, se han entrenado varios modelos de machine learning con el objetivo de predecir el riesgo de suicidio.

2. Limitaciones del Conjunto de Datos

Una limitación significativa de este proyecto es la falta de datos representativos para la mayoría de los países de Asia y África. Esta carencia implica que los modelos entrenados tienen una capacidad limitada para generalizar sus predicciones más allá de los países incluidos en el conjunto de datos. Los resultados obtenidos son probablemente más precisos para los países representados, pero es posible que no sean aplicables a regiones con contextos socioeconómicos y culturales diferentes.

3. Impacto de la Falta de Datos Globales

La ausencia de datos de regiones importantes como Asia y África afecta negativamente la robustez y aplicabilidad del modelo. Dado que las tasas de suicidio pueden estar influenciadas por factores culturales, económicos y sociales específicos de cada región, la falta de representación en el conjunto de datos puede llevar a predicciones sesgadas o inexactas cuando se aplican fuera del contexto en el que el modelo fue entrenado.

4. Próximos Pasos para Mejorar el Modelo

Para mejorar la capacidad predictiva y la generalización del modelo, se sugieren los siguientes pasos:

- **Ampliación del Conjunto de Datos**: Es fundamental recopilar datos adicionales de países de Asia y África para mejorar la representación global en el conjunto de datos. Esto permitiría al modelo aprender de una variedad más amplia de contextos y, por ende, realizar predicciones más generalizables.
- **Incorporación de Nuevas Variables**: Agregar características adicionales que puedan influir en el riesgo de suicidio, como la situación laboral, la situación financiera (deudas), problemas de salud crónicos y la infraestructura de salud mental de cada país, podría mejorar la precisión del modelo. Estas variables adicionales pueden capturar aspectos más específicos y relevantes del riesgo de suicidio.
- **Análisis Profundo en Países Específicos**: En lugar de tratar de abarcar todos los países del mundo, una estrategia alternativa podría ser centrarse en un país específico para recopilar y analizar datos de manera más detallada. Esto podría permitir el desarrollo de modelos más precisos y estrategias de prevención específicas para el grupo de mayor riesgo dentro de ese país.

5. Implicaciones para la Prevención del Suicidio

Un enfoque más centrado en datos y características específicas podría conducir a la creación de estrategias de prevención más efectivas. Por ejemplo, en función de los datos adicionales recopilados, se podrían diseñar intervenciones específicas como la inclusión de educación sobre salud mental en la escuela secundaria, o la organización de eventos comunitarios para personas mayores, lo que podría abordar directamente las causas subyacentes del riesgo de suicidio en poblaciones específicas.

6. Conclusión General

Aunque el modelo actual proporciona una base útil para la predicción del riesgo de suicidio, su aplicabilidad está limitada por la falta de datos globales y de características adicionales relevantes. Abordar estas limitaciones mediante la ampliación del conjunto de datos y la inclusión de nuevas variables podría mejorar significativamente tanto la precisión como la generalización del modelo, lo que, a su vez, podría contribuir a la creación de estrategias de prevención del suicidio más efectivas y adaptadas a las necesidades de poblaciones específicas.