Determinación de Nivel de Felicidad a Nivel Mundial

Arturo Marca Rivera, email: arturo.marca@pucp.edu.pe
Deyvis Raúl Atencio Velasquez, email: deyvis.atencio.v@gmail.com
Freddy Alvarado, email: freddy.alvarado.bazan@gmail.com, falvarado@cmiconsulting.pe
Jorge Sepúlveda Sepúlveda , email: buzondepitagoras@gmail.com
Javier José Quispe Varillas, email: jjqv1972@gmail.com

Abstract: El Machine Learning ha revolucionado en los últimos años las técnicas y metodologías de procesamiento de datos. Los datos estructurados son aquellos que nos informan de características o atributos de un tema específico. Uno de estos temas es el Nivel de Felicidad a Nivel Mundial. Nosotros proponemos un estudio basado en algoritmos de machine learning con la finalidad de clasificar el nivel de felicidad a nivel mundial. Finalmente, proponemos el algoritmo de mejor performance.

Keywords: Machine Learning, Felicidad

I. INTRODUCCIÓN

El informe de felicidad mundial se encuentra basado en encuestas históricas a nivel mundial. Es importante mencionar que es gestionado por las Naciones Unidas. Dicho informe va ganando más aceptación en los gobiernos ya que utilizan con mayor frecuencia indicadores de felicidad para informar sobre su toma de decisiones. Diversos expertos manifiestan que dichas mediciones se pueden utilizar para valorar el progreso de cada nación.

II. ÁREA DE ESTUDIOS Y DATASET

El dataset de felicidad mundial [1] utilizado en la presente investigación se encuentra basado en una serie de encuestas a nivel mundial desde el año 2015 a 2019.

Los puntajes y clasificaciones de felicidad utilizan datos de la Encuesta Mundial de Gallup. Los puntajes se basan en las respuestas a la pregunta principal de evaluación de la vida que se hizo en la encuesta. Esta pregunta, conocida como la escalera de Cantril, pide a los encuestados que piensen en una escalera en la que la mejor vida posible para ellos sea un 10 y la peor vida posible sea un 0 y que califiquen sus propias vidas actuales en esa escala. Los puntajes provienen de muestras representativas a nivel nacional para los años 2015-2019 y utilizan las ponderaciones de Gallup para hacer las estimaciones representativas. Las columnas que siguen al puntaje de felicidad estiman en qué medida cada uno de los seis factores (producción económica, apoyo social, esperanza de vida, libertad, ausencia de corrupción y generosidad) contribuyen a hacer que las evaluaciones de vida sean más altas en cada país.

Los datos tienen una serie de indicadores o atributos que se evalúan para determinar el ranking de felicidad a nivel mundial, podemos mencionar: PBI per Cápita, Familia, Expectativa de Vida, Libertad, Generosidad, Confianza en Gobierno, entre otros. Finalmente, en base a estos se obtiene la puntuación de felicidad.

1. PROBLEMA

¿Cuales son los indicadores que afectan la felicidad en el mundo?

2. HIPÓTESIS

H0: Los indicadores: PBI per cápita, Expectativa de Vida Saludable, Apoyo Social y Libertad no determinan los niveles de felicidad en el mundo.

H1: Los indicadores: PBI per cápita, Expectativa de Vida Saludable, Apoyo Social y Libertad determinan los niveles de felicidad en el mundo.

III. METODOLOGÍA

Los datos del reporte de felicidad a nivel mundial son tomados desde el 2015 a 2019. Se aprecia en el análisis exploratorio de los datos, que las encuestas anuales muestran ligera variación en cuanto a los atributos o indicadores para el cálculo del ranking.

En base a este análisis, se decidió utilizar en los siguientes atributos o indicadores de las encuestas:

- Ranking general (Overall rank)
- País o región (Country or región)
- Puntuación (Score)
- PBI per capita (GDP per capita)
- Apoyo social (Social support)
- Esperanza de vida saludable (Health life expectancy)
- Libertad para toma de decisiones (Freedom to make life choices)
- Generosidad (Generosity)
- Percepciones de corrupción (Perceptions of corruption)

Nosotros proponemos analizar utilizando diversos algoritmos de machine learning en base a los indicadores de felicidad encuestados para clasificar el nivel de felicidad a nivel mundial. Para ello, luego del proceso de ranking por puntaje obtenido en las encuestas, delimitamos los siguientes niveles:

- Nivel de felicidad Alto: puntuación mayor a 6.1
- Nivel de felicidad Normal: puntuación mayor a 4.5 y menor a 6.1
- Nivel de felicidad Bajo: puntuación menor a 4.5

Utilizaremos en el estudio los siguientes algoritmos:

- Regresión logística
- KNN
- SVM
- Árboles de Decisión
- Random Forest
- GBoosting

1. ANÁLISIS EXPLORATORIO DE LOS DATOS

   ❑ En nivel de felicidad de los países se mide mediante un score, el cual tiende a presentar una distribución normal según la figura 1:

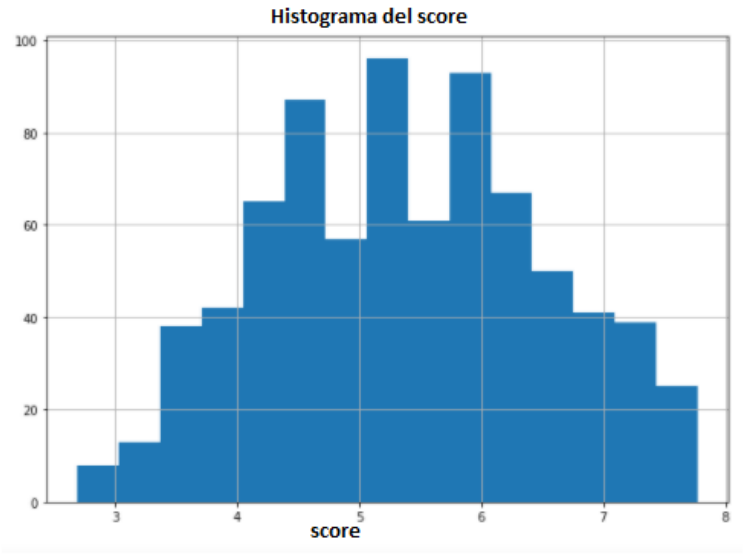

Figura 1: Histograma del score

- Las variables GDP per capita, Freedom to make life choicesa, Social support, Healthy life expectancy presentan relación lineal muy fuerte con el score de felicidad.

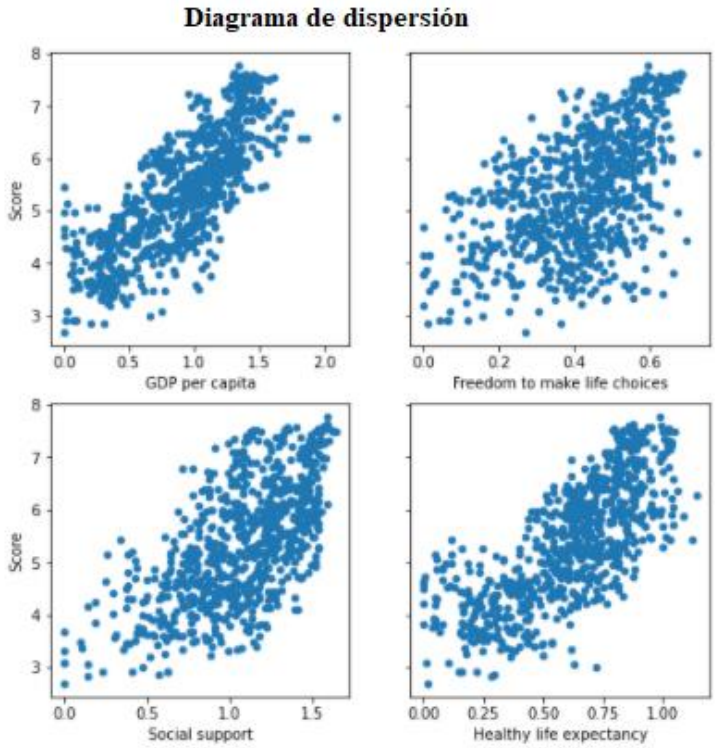

Figura 2: Diagrama de dispersión

- En la figura 3, podemos corroborar el nivel de correlación de las cuatro variables mencionadas anteriormente.

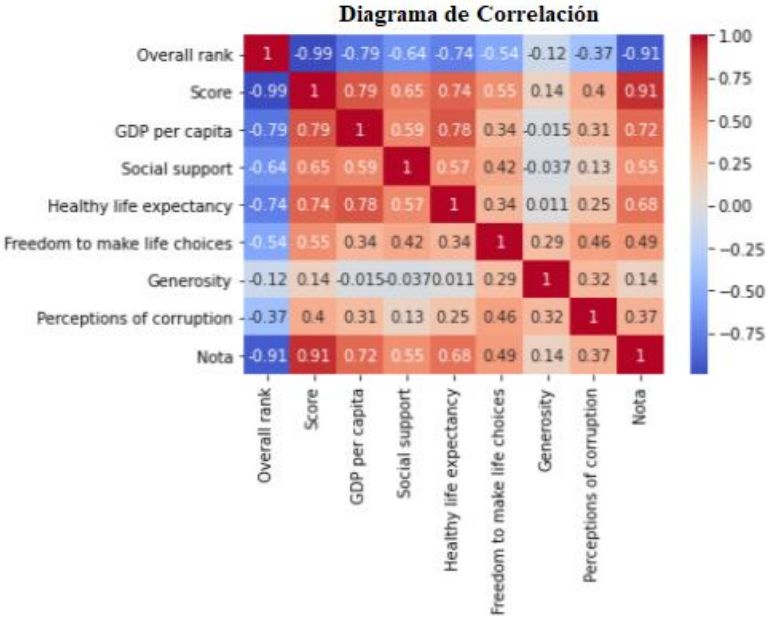

Figura 3: Diagrama de correlación

2. PREPROCESAMIENTO

- ❑ Renombrar las columnas o variables de los dataset desde el 2015 a 2019, ya que que en nombre de las columnas eran diferentes en algunos años. Al final nos quedamos con las siguientes variables:

    - Overall rank
    - Country or region
    - Score
    - GDP per capita
    - Social support
    - Healthy life expectancy
    - Freedom to make life choices
    - Generosity
    - Perceptions of corruption

- ❑ Unir la data de cada año, y con ello obtenemos el dataset final.

	Overall rank	Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	1	Finland	7.769	1.340	1.587	0.986	0.596	0.153	0.393
1	2	Denmark	7.600	1.383	1.573	0.996	0.592	0.252	0.410
2	3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341
3	4	Iceland	7.494	1.380	1.624	1.026	0.591	0.354	0.118
4	5	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298

Figura 4: Cinco primeros registros de la data final.

- ❑ Categorizar la variable score, de la Figura 1, podemos segmentar aproximadamente los valores menores a 4, entre 4 y 6, y mayores a 6.

    Se ha categorizado el score mediante los percentiles:

    P25=4.5

    P75=6.1

    - (3) Nivel de Felicidad Alto: Score > 6.1
    - (2) Nivel de Felicidad Medio : 4.5 < Score < 6.1
    - (1) Nivel de Felicidad Bajo: Score < 4.5

- ❑ Se han escalado las variables numéricas independientes, se ha aplicado la técnica de Min-Max normalization

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- ❑ Para entrenar y validar los modelos ha dividido la data en muestra de entrenamiento y prueba, en este caso, para entrenar el modelo se han usado el 80% de la data y el resto para validar.

3. MODELO : REGRESIÓN LOGÍSTICA

La Regresión Logística es una técnica estadística multivariante que nos permite estimar la relación existente entre una variable dependiente no métrica, en particular dicotómica y un conjunto de variables independientes métricas o no métricas.

Tabla de métricas:

	precision	recall	f1-score	support
1	0.64	0.68	0.66	34
2	0.78	0.72	0.75	86
3	0.74	0.84	0.78	37
accuracy			0.74	157
macro avg	0.72	0.75	0.73	157
weighted avg	0.74	0.74	0.74	157

4. MODELO: KNN - K-NEAREST NEIGHBORS

En la *clasificación k-NN* , la salida es una pertenencia a una clase. Un objeto se clasifica por una pluralidad de votos de sus vecinos, y el objeto se asigna a la clase más común entre sus *k* vecinos más cercanos (*k* es un número entero positivo , típicamente pequeño). Si *k* = 1, entonces el objeto simplemente se asigna a la clase de ese único vecino más cercano.

Tabla de métricas:

	precision	recall	f1-score	support
1	0.63	0.79	0.70	34
2	0.82	0.70	0.75	86
3	0.76	0.84	0.79	37
accuracy			0.75	157
macro avg	0.74	0.78	0.75	157
weighted avg	0.76	0.75	0.75	157

5. MODELO: SVM - SUPPORT VECTOR MACHINE

Estos métodos están relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama **vector soporte**. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase.

Tabla de métricas:

	precision	recall	f1-score	support
1	0.64	0.79	0.71	34
2	0.83	0.64	0.72	86
3	0.67	0.89	0.77	37
accuracy			0.73	157
macro avg	0.72	0.78	0.73	157
weighted avg	0.75	0.73	0.73	157

6. MODELO: DECISION TREE - ÁRBOLES DE DECISIÓN

Un árbol de decisión es un modelo de predicción utilizado en diversos ámbitos que van desde la inteligencia artificial hasta la Economía. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Tabla de métricas:

	precision	recall	f1-score	support
1	0.66	0.68	0.67	34
2	0.79	0.69	0.73	86
3	0.68	0.86	0.76	37
accuracy			0.73	157
macro avg	0.71	0.74	0.72	157
weighted avg	0.73	0.73	0.73	157

7. MODELO: RANDOM FOREST

Random Forest se considera como la "panacea" en todos los problemas de ciencia de datos. Útil para regresión y clasificación. Un grupo de **modelos** "débiles", se combinan en un **modelo** robusto. ... Para regresión, se toma el promedio de las salidas (predicciones) de todos los árboles.

Tabla de métricas:

	precision	recall	f1-score	support
1	0.68	0.74	0.70	34
2	0.81	0.74	0.78	86
3	0.76	0.84	0.79	37
accuracy			0.76	157
macro avg	0.75	0.77	0.76	157
weighted avg	0.77	0.76	0.76	157

8. MODELO: GBOOSTING

Los **modelos** Gradient **Boosting** están formados por un conjunto de árboles de decisión individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de mejorar los errores de los árboles anteriores.

Tabla de métricas:

	precision	recall	f1-score	support
1	0.61	0.65	0.63	34
2	0.79	0.69	0.73	86
3	0.72	0.89	0.80	37
accuracy			0.73	157
macro avg	0.71	0.74	0.72	157
weighted avg	0.73	0.73	0.72	157

IV. RESULTADOS Y DISCUSIONES

Evaluamos diferentes modelos de clasificación para los indicadores seleccionados:

- GDP per capita
- Social support

- Healthy life expectancy
- Freedom to make life choices

Utilizando el modelo de mejor accuracy, Random Forest, sobre el grupo de validación se obtuvieron resultados satisfactorios por lo que se acepta la hipótesis alternativa H1.

V. CONCLUSIONES

Se observa que las variables analizadas: GDP per capita, Social support, Healthy life expectancy y Freedom to make life choices; explican adecuadamente los niveles de felicidad que ostentan los países en los que se aplicó la encuesta desde el año 2015 hasta el año 2019.

Así mismo se identificó el mejor modelo de predicción, usando los algoritmos de Clasificación de Machine Learning, es el algoritmo Random Forest que obtuvo el mayor valor de accuracy: **0.764** tal como se puede apreciar en la Tabla 1:

	Accuracy
RANDOM FOREST	0.764
KNN - K-NEAREST NEIGHBORS	0.751
REGRESIÓN LOGÍSTICA	0.738
SVM - SUPPORT VECTOR MACHINE	0.732
DECISION TREE - ÁRBOLES DE DECISIÓN	0.726
GBOOSTING	0.726

Tabla 1: Accuracy para cada modelo

VI. RECOMENDACIONES

Es importante considerar, los resultados del presente trabajo, como punto de partida para futuros trabajos de investigación que evalúen la posibilidad de identificar la combinación óptima de los indicadores: GDP, SocialS, Healthy y Freedom a los que debe aspirar un Estado en su intento por mejorar los niveles de felicidad de su población.

Por ejemplo para el caso de **South Sudan** quien se encuentra en el puesto 156 del ranking del año 2019 con un **Score de 2.853**, según el modelo propuesto le corresponde un índice de felicidad **BAJA**:

Score = f(GDP,SocialS,Healthy, Freedom)
2.853 = f(0.306,0.575,0.295,0.01)

Se recomienda evaluar el desarrollo de un modelo de optimización que determine, según las restricciones propias del país en estudio, cual es el valor óptimo que deben alcanzar los indicadores antes mencionados a fin de maximizar su Score de Felicidad. De esta manera se podría orientar de manera efectiva las políticas de gobierno que buscan mejorar la calidad de vida de sus ciudadanos.

Asi mismo y como parte de los nuevos trabajos de investigación, se recomienda tener como referencia a **Finlandia** que tiene el Score más alto: 7.769 y se ubica en el primer puesto del ranking 2019 con un Score de felicidad **ALTO**:

Score = f(GDP,SocialS,Healthy, Freedom)
7.769 = f(1.34,1.587,0.986,0.596)

o seleccionar como referente otro país con un mayor Score de Felicidad y con una realidad geopolítica, demográfica y cultural muy similar.

REFERENCIAS

[1] World Happiness Report 2019. Disponible en línea: https://www.kaggle.com/unsdsn/world-happiness. [Consulta: 25/01/2021]

[2] Enrique M.G. (2008). Manual de Uso de SPSS. Disponible en: http://e-spacio.uned.es/fez/eserv/bibliuned:500727/Guia_SPSS.pdf [Consulta: 03/02/2021]

[3] Miller, Vlad (2019) Algoritmos de Aprendizaje Automático Supervisado. Disponible en: https://www.toptal.com/machine-learning/explorando-algoritmos-de-aprendizaje-automatico-supervisado [Consulta: 03/02/2021]