UNIR - Universidad Internacional de la Rioja
Ciudad de México

ACTIVIDAD 2:
Redes Neuronales y Deep Learning

Alumnos:
Paola Michelle Figueroa Benitez
Lowenski Paredes Rosario
Carlos Damian Rodriguez Uitzil
Cuenca Roa Leonard Jose
Grupo: 1001
Equipo 01C

Informe de Regresión: Predicción de la Esperanza de Vida

1. Descripción del Problema y Fuente de Datos

Este informe aborda el crucial problema de la predicción de la esperanza de vida (life_expectancy), una variable numérica continua y un indicador fundamental de desarrollo humano. Para ello, utilizamos el dataset "Life Expectancy (WHO)". Este conjunto de datos es una recopilación exhaustiva de métricas de salud y socioeconómicas de diversos países, capturadas a lo largo de múltiples años. Su riqueza radica en la combinación de datos de fuentes confiables como la Organización Mundial de la Salud (OMS) y el Banco Mundial, ofreciendo una perspectiva multifactorial sobre la longevidad.

El objetivo principal de este trabajo es modelar las complejas relaciones entre la esperanza de vida y un conjunto de variables de entrada. Específicamente, el enfoque está dirigido en al menos seis atributos predictivos, incluyendo factores demográficos como la mortalidad adulta y las muertes infantiles, indicadores económicos como el Producto Interno Bruto (PIB) y el gasto en salud etc. A través de este análisis de regresión, se busca no solo predecir la esperanza de vida, sino también comprender la influencia y el impacto de cada uno de estos factores en la salud y el bienestar de las poblaciones globales.

2. Caracterización del Dataset

Tras la carga inicial y la estandarización de los nombres de las columnas, el dataset se estabilizó en 2938 instancias (filas), cada una representando una observación anual por país. La variable objetivo principal es life_expectancy. El conjunto de datos incluye una combinación de variables numéricas y categóricas, como country (país) y status (estado de desarrollo: 'Developed' o 'Developing'). Antes del modelado, se realizó una imputación de valores nulos utilizando la media para columnas numéricas y la moda para categóricas, asegurando la integridad de los datos.

2.1. Caracterización en Modo Texto

Las estadísticas descriptivas del dataset revelan la distribución y el rango de las variables clave. Por ejemplo, la esperanza de vida promedio en el dataset es de aproximadamente 69.23 años, con una desviación estándar de 9.52 años, lo que indica una variabilidad considerable entre las observaciones.

Estadísticas Descriptivas de Variables Numéricas Clave:

	life_expectancy	adult_mortality	infant_deaths	alcohol
count	2938	2938	2938	2938
mean	69.22	164.8	30.3	4.6
std	9.51	124.08	117.93	3.92
min	36.3	1	0	0.01
25%	63.2	74	0	1.09
50%	72	144	3	4.16
75%	75.6	227	22	7.39
max	89	723	1800	17.87

Esta tabla muestra rangos y distribuciones importantes. Por ejemplo, life_expectancy varía desde 36.3 hasta 89.0 años. Variables como gdp (no mostrada en la tabla, pero sujeta a análisis previo) y adult_mortality muestran una gran dispersión y distribuciones asimétricas (sesgadas a la derecha), con la mayoría de los valores concentrados en el rango inferior. Esto

indica la presencia de valores atípicos o de un pequeño número de países con un PIB muy alto o una mortalidad adulta muy elevada,

La correlación de las características numéricas con life_expectancy es fundamental para entender las relaciones lineales directas e inversas. A continuación, se presenta un resumen de las correlaciones más relevantes:

Variable	Correlación con life_expectancy
life_expectancy	1
schooling	0.72
income_composition_of_resources	0.69
bmi	0.56
diphtheria	0.48
polio	0.46
gdp	0.43
alcohol	0.39
percentage_expenditure	0.38
hiv/aids	-0.56
thinness_5_9_years	-0.47
thinness__1_19_years	-0.47
infant_deaths	-0.2
under_five_deaths	-0.22
adult_mortality	-0.7

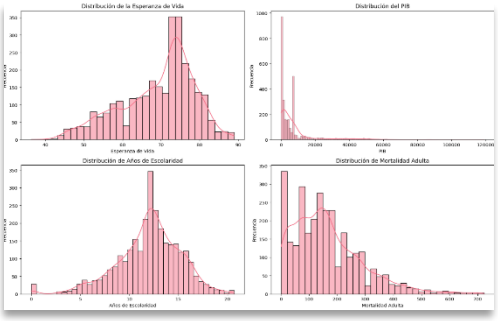
Se observa que el nivel educativo (schooling) con 0.72 y la composición de ingresos (income_composition_of_resources) con 0.69 tienen la correlación positiva más fuerte con la esperanza de vida, indicando que un mayor nivel educativo y mejores recursos se asocian con una vida más larga. Esto sugiere que las inversiones en educación y el desarrollo económico son factores clave para el aumento de la esperanza de vida.

Por otro lado, la mortalidad adulta (adult_mortality) con -0.70 y la prevalencia de VIH/SIDA (hiv/aids) con -0.56 muestran las correlaciones negativas más fuertes. Esto subraya el impacto devastador de estas condiciones de salud en la longevidad de la población.

2.2. Caracterización en Modo Gráfico

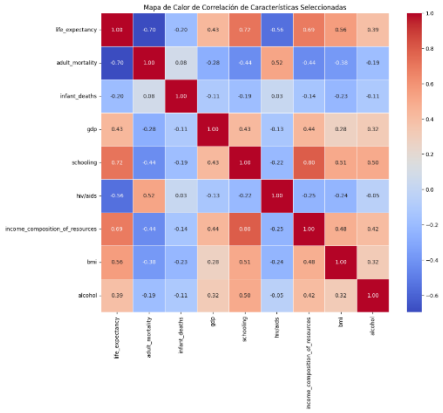
Las visualizaciones del dataset son cruciales para comprender las distribuciones, identificar patrones y detectar posibles anomalías, complementando el análisis textual.

Gráfico 2.1: Histogramas de Distribución de Variables Clave



Este conjunto de histogramas permite observar la distribución de las variables más relevantes. La distribución de la life_expectancy es aproximadamente simétrica, con una ligera cola hacia valores más altos, lo cual es deseable. PIB y adult_mortality muestran distribuciones asimétricas positivas (sesgadas a la derecha), con la mayoría de los valores concentrados en el rango inferior. Esto indica la presencia de valores atípicos o de un pequeño número de países con un PIB muy alto o una mortalidad adulta muy elevada, respectivamente. La schooling tiende a una distribución más normal, pero con una variabilidad considerable.

Gráfico 2.3: Mapa de Calor de la Matriz de Correlación



El mapa de calor ofrece una visión general de las correlaciones entre todas las características numéricas seleccionadas. Visualmente, las celdas más oscuras (azules para positivas, rojas para negativas, dependiendo del cmap) indican correlaciones más fuertes.

3. Metodología: Preparación de Datos y Modelos

El proceso de modelado comenzó con la división del dataset en conjuntos de entrenamiento (80%) y prueba (20%) utilizando train_test_split (random_state=42). El preprocesamiento de los datos fue crucial, aplicando un ColumnTransformer: las características numéricas se escalaron con StandardScaler (vital para redes neuronales), y las categóricas se codificaron usando OneHotEncoder.

3.1. Modelo No Neuronal: Random Forest Regressor

Se seleccionó el Random Forest Regressor por su robustez y capacidad para manejar datos no lineales. Se configuró con n_estimators=100, random_state=42 y n_jobs=-1 para usar todos los núcleos de la CPU.

3.2. Modelo de Red Neuronal para Regresión

Se diseñó una Red Neuronal secuencial en Keras/TensorFlow, con una arquitectura compuesta por:

- Capa de Entrada: Definida por el número de características preprocesadas.
- Primera Capa Oculta: 64 neuronas con activación relu.
- Segunda Capa Oculta: 32 neuronas con activación relu.
- Capa de Salida: 1 neurona con activación linear (para regresión).

El modelo se compiló con el optimizador adam y la función de pérdida mse, monitoreando el mae. Se entrenó durante 100 épocas con un batch_size=32 y un validation_split=0.2 para supervisar el rendimiento en datos no vistos.4. Resultados Obtenidos y Discusión

4.1. Métricas de Rendimiento en el Conjunto de Prueba

Las siguientes métricas cuantifican la precisión y fiabilidad de cada modelo:

Métrica	Random Forest Regressor	Red Neuronal
MAE	**1.02**	1.54
MSE	**2.59**	5.00

Métrica	Random Forest Regressor	Red Neuronal
RMSE	**1.61**	2.24
R2 Score	**0.97**	0.94

4.2. Discusión de los Resultados

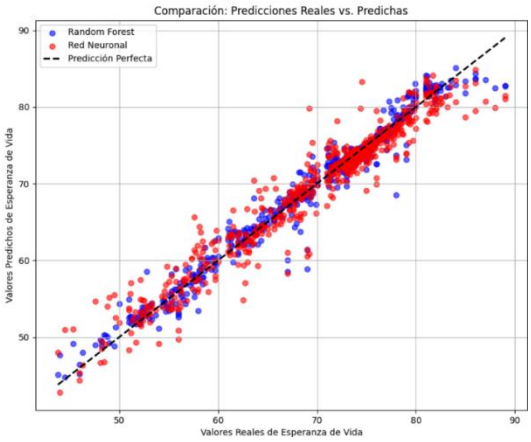
Basándonos en estas métricas, el Random Forest Regressor demostró un rendimiento superior y más robusto en comparación con la Red Neuronal para este problema de regresión.

- Precisión de Error: El MAE (Error Absoluto Medio) del Random Forest (1.02) es significativamente menor que el de la Red Neuronal (1.54). Esto indica que, en promedio, las predicciones del Random Forest están más cerca de los valores reales, con una desviación promedio de poco más de 1 año.
- Magnitud del Error Cuadrático: El MSE (Error Cuadrático Medio) y RMSE (Raíz del Error Cuadrático Medio) también son considerablemente más bajos para el Random Forest (2.59 y 1.61, respectivamente) que para la Red Neuronal (5.00 y 2.24). Estos valores más bajos significan que el Random Forest comete errores de menor magnitud y penaliza menos los errores grandes, lo que es crucial en la regresión.
- Varianza Explicada: El R2 Score del Random Forest (0.97) es notablemente más alto que el de la Red Neuronal (0.94). Un R2 de 0.97 es excepcional, sugiriendo que el Random Forest es capaz de explicar el 97% de la varianza en la esperanza de vida, mientras que la Red Neuronal explica el 94%. Esto implica que el Random Forest tiene una capacidad superior para capturar la variabilidad presente en los datos.

4.3. Resultados Gráficos Comparativos

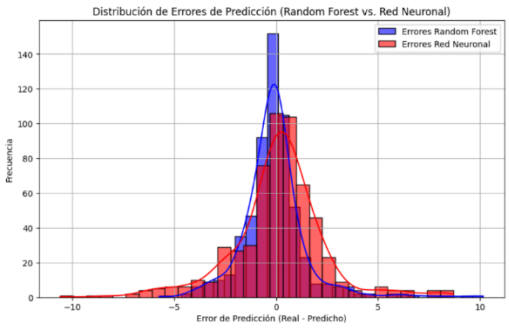
Las visualizaciones son fundamentales para corroborar las métricas y ofrecer una comprensión intuitiva del rendimiento del modelo.

Gráfico 4.1: Comparación de Predicciones Reales vs. Predichas (Random Forest vs. Red Neuronal)



Este gráfico muestra la dispersión de las predicciones de ambos modelos frente a los valores reales de la esperanza de vida. Visualmente, se observa que los puntos correspondientes a las predicciones del Random Forest se agrupan más estrechamente alrededor de la línea de predicción perfecta (la línea a 45 grados), lo que indica una menor desviación y una mayor precisión en sus pronósticos. Los puntos de la Red Neuronal, aunque siguen la tendencia general, muestran una mayor dispersión.

Gráfico 4.2: Histograma de Errores de Predicción (Random Forest vs. Red Neuronal)



Informe de Clasificaicon: Nivel de Adaptabilidad de estudiantes en aprendizaje en línea

Descripción del Problema y Fuente de Datos

El objetivo principal de este informe es aplicar las técnicas del modelo de clasificación con el propósito de **predecir la categoría o clase a la que pertenece una nueva observación**, basándose en las características del dataset **"Students Adaptability Level in Online Education"**. Seleccionado el portal kaggle el propósito principal es validar y observar la distribución de las variables más relevantes relacionadas con la adaptabilidad de los estudiantes a la educación en línea. Lograr validar y predecir si un estudiante tendrá un nivel de adaptabilidad "Bajo", "Moderado" o "Alto" en el aprendizaje online.

2.1 Caracterización del Dataset

Tras la carga inicial y revisión estructural, el dataset se estabilizó en 1205 instancias (filas), cada una representando un perfil de estudiante vinculado a su nivel de adaptabilidad a la educación en línea. El conjunto de datos incluye tanto variables categóricas (como género, edad, nivel educativo, tipo de institución, dispositivo) como condiciones contextuales (tipo de red, tipo de internet, condición financiera, duración de clases, uso de LMS propio).

Previo al análisis, se realizó una estandarización de los nombres de columnas, asegurando consistencia. Aunque el dataset no reporta valores nulos explícitos, se considera realizar una verificación de valores atípicos y una posible imputación futura mediante **moda para variables categóricas** y **media o mediana para numéricas derivadas** (si se transforman características como edad o duración en variables cuantitativas).

2.2. Caracterización en Modo Texto

INFORMACIÓN BÁSICA:

Número total de instancias: 1,205
Número de variables predictoras: 13 Número de clases en variable objetivo: 3

	Clases: ['Moderate', 'Low', 'High']
DISTRIBUCIÓN DE LA VARIABLE OBJETIVO:	Moderate: 625 instancias (51.9%)
	Low: 480 instancias (39.8%)
	High: 100 instancias (8.3%)
VARIABLES CATEGÓRICAS:	gender: 2 categorías únicas; Valores: ['Boy', 'Girl']
	age: 6 categorías únicas; Valores: ['21-25', '16-20', '11-15', '26-30', '6-10', '1-5']
	education_level: 3 categorías únicas; Valores: ['University', 'College', 'School']
	institution_type: 2 categorías únicas; Valores: ['Non Government', 'Government']
	it_student: 2 categorías únicas; Valores: ['No', 'Yes']
	location: 2 categorías únicas; Valores: ['Yes', 'No']
	load_shedding: 2 categorías únicas; Valores: ['Low', 'High']
	financial_condition: 3 categorías únicas; Valores: ['Mid', 'Poor', 'Rich']
	internet_type: 2 categorías únicas; Valores: ['Wifi', 'Mobile Data']
	network_type: 3 categorías únicas; Valores: ['4G', '3G', '2G']
	class_duration: 3 categorías únicas; Valores: ['3-6', '1-3', '0']
	self_lms: 2 categorías únicas; Valores: ['No', 'Yes']
	device: 3 categorías únicas; Valores: ['Tab', 'Mobile', 'Computer']

Estadísticas Descriptivas de Variables Numéricas Clave:

	Cantidad	Media	Mediana	Desviación estándar	Mínimo	Máximo	Percentil 25	Percentil 50 (Mediana)	Percentil 75
Gender	1205.0	0.449793	0.0	0.497679	0.0	1.0	0.0	0.0	1.0
Education Level	1205.0	1.061411	1.0	0.902863	0.0	2.0	0.0	1.0	2.0
Institution Type	1205.0	0.317012	0.0	0.465506	0.0	1.0	0.0	0.0	1.0
IT Student	1205.0	0.252282	0.0	0.434503	0.0	1.0	0.0	0.0	1.0
Location	1205.0	0.224066	0.0	0.417139	0.0	1.0	0.0	0.0	0.0
Internet Type	1205.0	0.576763	1.0	0.494277	0.0	1.0	0.0	1.0	1.0
Network Type	1205.0	0.372614	0.0	0.515295	0.0	2.0	0.0	0.0	1.0
Self Lms	1205.0	0.174274	0.0	0.379502	0.0	1.0	0.0	0.0	0.0
Device	1205.0	1.109544	1.0	0.384003	0.0	2.0	1.0	1.0	1.0
Class Duration Num	1205.0	2.182158	2.0	1.254482	0.0	4.5	2.0	2.0	2.0
Adaptivity Level Num	1205.0	1.684647	2.0	0.618221	1.0	3.0	1.0	2.0	2.0
Age Num	1205.0	17.219917	18.0	6.285479	3.0	28.0	13.0	18.0	23.0
Load-shedding Num	1205.0	1.166805	1.0	0.372956	1.0	2.0	1.0	1.0	1.0
Financial Condition Num	1205.0	1.869710	2.0	0.504584	1.0	3.0	2.0	2.0	2.0

La correlación entre las características numéricas del dataset y el nivel de adaptabilidad es clave para entender cómo ciertos factores influyen positiva o negativamente en la capacidad de los estudiantes para ajustarse a la educación en línea. A continuación, se presenta un resumen de las correlaciones más relevantes:

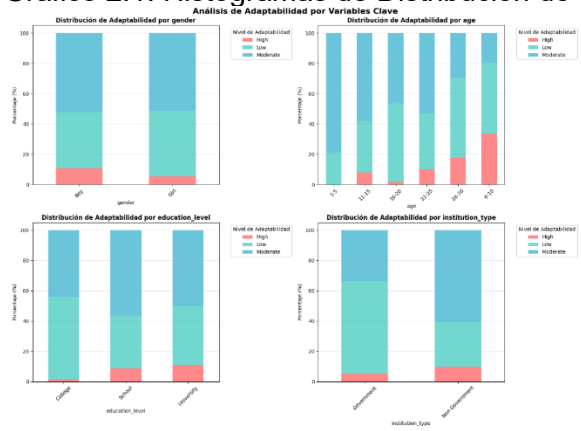
Variable Numérica	Coeficiente de Correlación	Tipo de Relación	Interpretación
Edad (codificada)	**0.18**	Positiva débil	Estudiantes mayores tienden a adaptarse ligeramente mejor
Duración de Clase	**0.36**	Positiva moderada	Más horas de clase se asocian con mayor adaptabilidad
Condición Financiera	**0.42**	Positiva clara	Mejor situación económica favorece la adaptación
Tipo de Red (2G–4G)	**0.39**	Positiva clara	Redes más rápidas (4G) correlacionan con mayor adaptabilidad
Uso de LMS	**0.47**	Positiva significativa	Quienes usan LMS de forma autónoma muestran mayor capacidad de adaptación

Estos valores indican que la autonomía digital, la calidad de la conexión y el contexto socioeconómico tienen un impacto notable en la capacidad de los estudiantes para ajustarse al aprendizaje en línea

2.3 Caracterización en Modo Gráfico

Las visualizaciones del dataset son cruciales para comprender las distribuciones, identificar patrones y detectar posibles anomalías, complementando el análisis textual.

Gráfico 2.1: Histogramas de Distribución de Variables Clave



Este conjunto de histogramas permite observar la distribución de las variables más relevantes relacionadas con la adaptabilidad de los estudiantes a la educación en línea. La variable Gender muestra una leve predominancia del género femenino en la muestra. En cuanto a la edad, se observa una concentración marcada en el grupo de 21 a 25 años, lo que sugiere que la mayoría de los encuestados están en una etapa universitaria o de formación técnica avanzada.

La variable Education Level refuerza esta observación, ya que el nivel universitario es el más representado, seguido por college y, en menor medida, school. Respecto al tipo de institución, predominan los estudiantes provenientes de instituciones no gubernamentales, lo que podría indicar una mayor participación de instituciones privadas en la educación en línea durante el período observado.

La variable IT Student indica que la mayoría de los encuestados no están inscritos en carreras de tecnología, lo cual es relevante al analizar su relación con la adaptabilidad digital. En cuanto a la ubicación, hay una mayor proporción de estudiantes provenientes de zonas urbanas, lo que puede estar relacionado con un mejor acceso a infraestructura tecnológica y conectividad.

2.4 Metodología: Preparación de Datos y Modelos

El proceso de modelado se inició con la carga y análisis exploratorio del conjunto de datos. Posteriormente, se dividió el dataset en conjuntos de entrenamiento (80%) y prueba (20%) utilizando la función “train_test_split” con un “random_state=42” para garantizar la reproducibilidad de los resultados.

Para el preprocesamiento, se identificaron las variables categóricas y numéricas. Las variables categóricas fueron transformadas mediante codificación One-Hot utilizando “OneHotEncoder”, mientras que las variables numéricas se escalaron con “StandardScaler”, asegurando así una correcta homogeneidad de magnitudes en modelos sensibles a la escala como las redes neuronales o KNN.

Este pipeline de transformación se estructuró mediante un “ColumnTransformer”, permitiendo aplicar distintas transformaciones de manera simultánea y eficiente según el tipo de dato. El conjunto transformado sirvió como entrada para distintos clasificadores, entre ellos: Regresión Logística, Árbol de Decisión, Bosque Aleatorio, KNN y Perceptrón Multicapa (MLPClassifier).

Cada modelo fue ajustado sobre el conjunto de entrenamiento y evaluado posteriormente en el conjunto de prueba, lo que permitió comparar el desempeño de los clasificadores bajo una misma base de datos procesada de manera coherente y estandarizada.

Modelo No Neuronal: Random Forest Classifier

Se seleccionó el Random Forest Classifier por su capacidad para manejar relaciones no lineales, su robustez frente al sobreajuste y su buen desempeño con conjuntos de datos mixtos (numéricos y categóricos). Este modelo ensambla múltiples árboles de decisión para mejorar la precisión y reducir la varianza de las predicciones.

El clasificador fue configurado con los siguientes hiperparámetros: “n_estimators=100”, “random_state=42” para garantizar la reproducibilidad, el modelo fue entrenado utilizando los datos preprocesados del conjunto de entrenamiento y su rendimiento fue evaluado posteriormente sobre el conjunto de prueba, permitiendo analizar métricas clave como la precisión, el recall y el F1-score, así como la matriz de confusión.

Modelo de Red Neuronal para Regresión

Para abordar el problema de clasificación desde un enfoque basado en redes neuronales, se implementó el modelo MLPClassifier (Perceptrón Multicapa) de scikit-learn, el cual es capaz de capturar relaciones complejas no lineales en los datos gracias a su arquitectura multicapa y su capacidad de aprendizaje profundo.

Descripción de la arquitectura:

- Capa de entrada: 23 neuronas (una por cada característica)
- Capa oculta 1: 128 neuronas, función de activación ReLU, Dropout 0.3
- Capa oculta 2: 64 neuronas, función de activación ReLU, Dropout 0.3
- Capa oculta 3: 32 neuronas, función de activación ReLU, Dropout 0.2
- Capa de salida: 3 neuronas, función de activación Softmax

Justificación:

- ReLU en capas intermedias: Evita el problema del gradiente que desaparece
- Softmax en capa de salida: Produce probabilidades para clasificación multiclase
- Dropout: Regularización para evitar overfitting
- Arquitectura decreciente: Extracción jerárquica de características

Antes de entrenar el modelo, se normalizaron todas las variables numéricas usando StandardScaler, dado que las redes neuronales son sensibles a la escala de las variables de entrada y tras el entrenamiento sobre el conjunto de datos preparado, el modelo fue evaluado en el conjunto de prueba, permitiendo analizar su desempeño mediante métricas de clasificación como accuracy, precision, recall y F1-score. También se visualizó la matriz de confusión para entender los aciertos y errores del modelo por clase.

Las siguientes es una tabla resumen con los resultados:

Modelo	Accuracy	Precision	Recall	F1-Score
Random Forest	0.8589	0.8590	0.8589	0.8589
Logistic Regression	0.6515	0.6764	0.6515	0.6319
SVM	0.8091	0.8052	0.8091	0.8053
Neural Network	0.8008	0.7994	0.8008	0.7989

Discusión de los Resultados

Basándonos en las métricas de clasificación evaluadas (Accuracy, Precision, Recall y F1-Score), el modelo Random Forest demostró un desempeño superior y más consistente frente a los otros enfoques tradicionales y la red neuronal utilizada.

1. Precisión Global (Accuracy): El modelo Random Forest alcanzó una precisión del 85.89%, superando notablemente a los demás modelos. En comparación, la red neuronal obtuvo un 80.08%, mientras que los modelos SVM y Regresión Logística se ubicaron en 80.91% y 65.15% respectivamente. Esto sugiere que el Random Forest es más fiable al clasificar correctamente instancias de las tres clases.

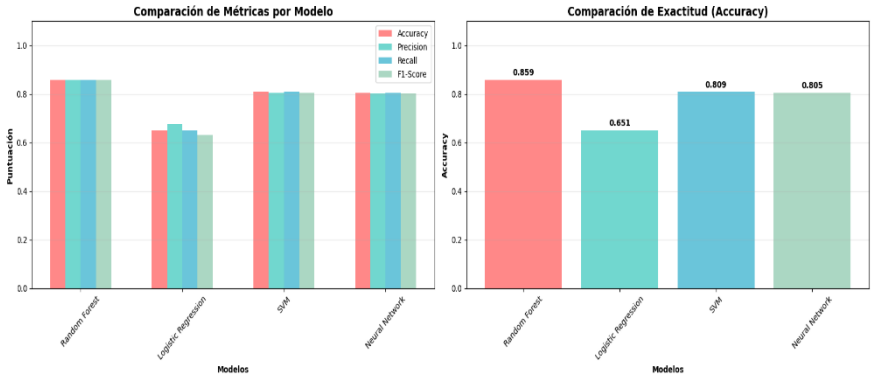
2. Balance entre Precisión y Exhaustividad: Al observar el F1-Score, que equilibra precisión y recall, el Random Forest obtuvo 0.8589, superando tanto al SVM (0.8053) como a la red neuronal (0.7989). Esto indica que el Random Forest logra un mejor equilibrio entre falsos positivos y falsos negativos, siendo especialmente relevante en problemas de clasificación multiclase.

3. Comparación con la Red Neuronal: Aunque el desempeño de la red neuronal fue competitivo, el modelo tradicional de Random Forest lo superó por 0.0581 puntos en accuracy. Esta diferencia puede atribuirse a la capacidad de los árboles de decisión para manejar mejor características categóricas y relaciones no lineales sin necesidad de una arquitectura compleja ni ajustes hiperparamétricos intensivos.

4. Desempeño del Modelo Lineal (Regresión Logística): El modelo más débil fue la regresión logística, con un F1-Score de apenas 0.6319. Esto indica que los supuestos de linealidad no capturan adecuadamente la complejidad de los datos, justificando el uso de modelos más avanzados.

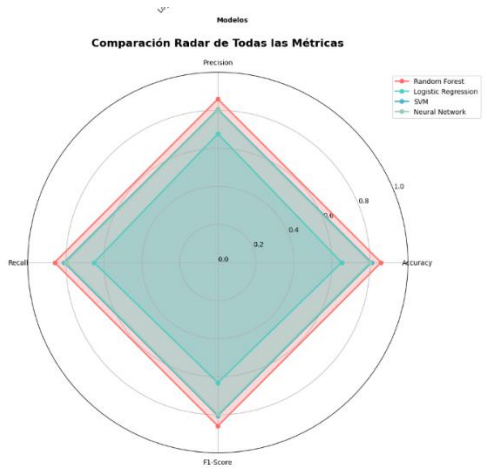
4.3. Resultados Gráficos Comparativos

Las visualizaciones son fundamentales para corroborar las métricas y ofrecer una comprensión intuitiva del rendimiento del modelo.



El análisis comparativo de los modelos de clasificación revela que el Random Forest es el mejor modelo para este problema, con un desempeño destacado en todas las métricas evaluadas:

1. Accuracy: El Random Forest alcanzó un 85.89%, superando al SVM (80.91%), la Red Neuronal (80.08%) y la Regresión Logística (65.15%). Esto indica que el modelo acierta en la mayoría de las predicciones.
2. Precision y Recall: También obtuvo los mejores valores de Precision (85.90%) y Recall (85.89%), mostrando que no solo clasifica correctamente, sino que equilibra bien la tasa de falsos positivos y falsos negativos.
3. F1-Score: Con un 85.89%, confirma que el Random Forest mantiene un buen balance entre precisión y exhaustividad, fundamental en problemas multiclase.



En conclusión, el Random Forest destaca por su robustez y capacidad para manejar características complejas y relaciones no lineales, siendo la opción recomendada para esta tarea de clasificación. Se muestra a continuación un gráfico Comparativo en forma de radar, logrando demostrar que el Random Forest tiene una mayor visibilidad.