Informe de Predicción de Riesgo de Salud Mental mediante Modelos de Machine Learning

Por Sofía Clavijo

1. Introducción

En el contexto del entorno laboral, el cuidado de la salud mental ha adquirido una relevancia cada vez mayor. Las organizaciones están reconociendo la importancia de promover el bienestar mental de sus empleados, no solo como una responsabilidad ética, sino también como una estrategia para mejorar la productividad, el ambiente de trabajo y la retención del talento. A medida que la tecnología avanza, las herramientas de ciencia de datos y aprendizaje automático permiten analizar grandes volúmenes de información para identificar patrones y predecir conductas o riesgos potenciales.

Este proyecto tiene como objetivo aplicar modelos de aprendizaje automático para predecir el riesgo de salud mental en individuos, a partir de una serie de indicadores personales y laborales. El trabajo fue desarrollado como parte de la asignatura de Programación de la Universidad EIA, y busca no solo construir modelos precisos, sino también desarrollar una comprensión profunda del proceso de preparación de datos, modelado y evaluación comparativa de resultados.

2. Objetivo

El objetivo principal de este proyecto es predecir si una persona se encuentra en riesgo de salud mental utilizando un conjunto de indicadores que incluyen edad, género, historial de salud mental, nivel de estrés, entre otros. Para lograrlo, se emplean técnicas de ciencia de datos que abarcan desde el análisis exploratorio y la limpieza de datos, hasta la construcción de modelos de clasificación y su posterior evaluación mediante métricas estándares como la precisión, el recall y el F1-score.

3. Dataset

Se utilizó un dataset simulado proveniente de Kaggle, con 10.000 instancias y 15 atributos relevantes para la predicción del riesgo de salud mental. Los atributos fueron cuidadosamente seleccionados para representar diferentes dimensiones del bienestar de un individuo, como el estado laboral, el entorno de trabajo, el apoyo social, los hábitos de sueño y actividad física, y los niveles de ansiedad y depresión.

Atributos considerados:

- age
- gender
- employment_status
- work_environment
- mental_health_history
- seeks_treatment
- stress_level
- sleep_hours
- physical_activity_days
- depression_score
- anxiety_score
- social_support_score
- productivity_score

El atributo de identificación (ID) fue eliminado para mantener la privacidad y evitar datos irrelevantes que puedan sesgar los modelos.

Datos faltantes

Dado que el dataset original no contenía valores nulos, se simularon valores faltantes bajo el enfoque MCAR (Missing Completely At Random), asegurando que al menos el 5% de los datos estuvieran ausentes de forma aleatoria. Esta simulación permitió poner a prueba técnicas de imputación y evaluación de calidad del dataset, aspecto fundamental en proyectos reales donde los datos faltantes son comunes.

Tratamiento de datos nulos

Se evaluaron dos enfoques:

1. Eliminación de filas incompletas: rápido y directo, pero con pérdida de información.
2. Imputación de valores faltantes: se aplicó mediana para variables numéricas y moda para variables categóricas. Esta estrategia conservó la mayor parte de los datos sin introducir sesgos significativos.

Finalmente, se optó por la imputación, ya que conserva mayor información sin reducir el tamaño de la muestra, asegurando que los modelos puedan generalizar mejor.

4. Análisis Exploratorio de Datos

El análisis exploratorio permitió conocer la distribución de las variables, identificar patrones y correlaciones, y preparar el terreno para un preprocesamiento efectivo. Se emplearon histogramas, diagramas de barras, mapas de calor de correlaciones y gráficos de pares

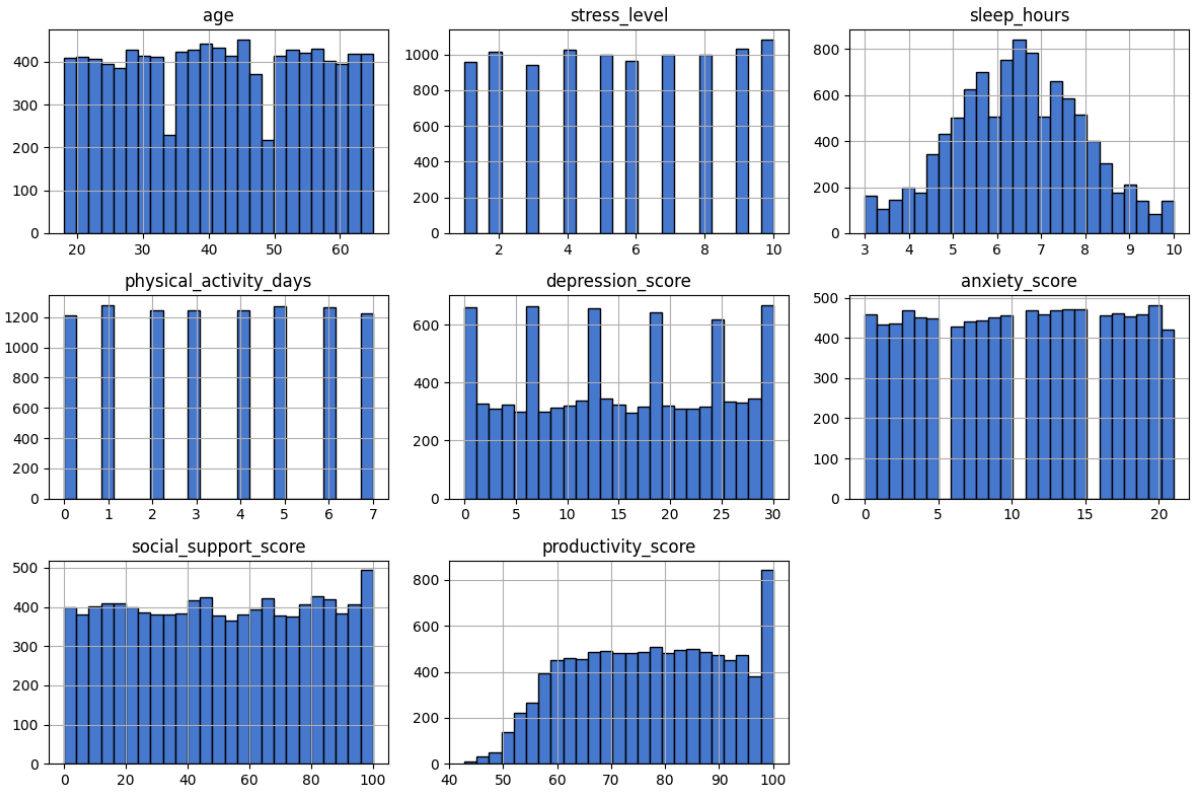


Imagen: distribución de datos numéricos

- La población se encuentra generalmente equitativamente distribuida en el rango de **edad** de los 18 a los 64 años; a excepción de los 32 a 48 años que puede deberse a una conducta por migración laboral desde la empresa o desde propia decisión de los empleados.
- En las **horas** de sueño encontramos un distribución casi normal con q1 5 y q2 8
- La actividad física se encuentra con una distribución uniforme continua al igual que el soporte social.
- El puntaje de productividad se encuentra sesgada a la derecha pero encontramos uno datos anómalos que no siguen un comportamiento que encaje en una distribución, esto puede deberse a como se toma estos datos, donde puede ser subjetivo donde las personas influyen directamente en el puntaje, viendo una relación directa en cómo se auto perciben y el riesgo en la salud mental, como se ve más adelante

- **Distribución de género**: La mayoría de los participantes se identifican como hombres o mujeres, con proporciones similares. Alrededor del 10% se identifican como no binarios o prefieren no decirlo, lo que aporta diversidad al dataset.

- **Tipos de datos**: 8 variables numéricas y 7 categóricas, con entre 2 y 4 categorías por variable categórica.

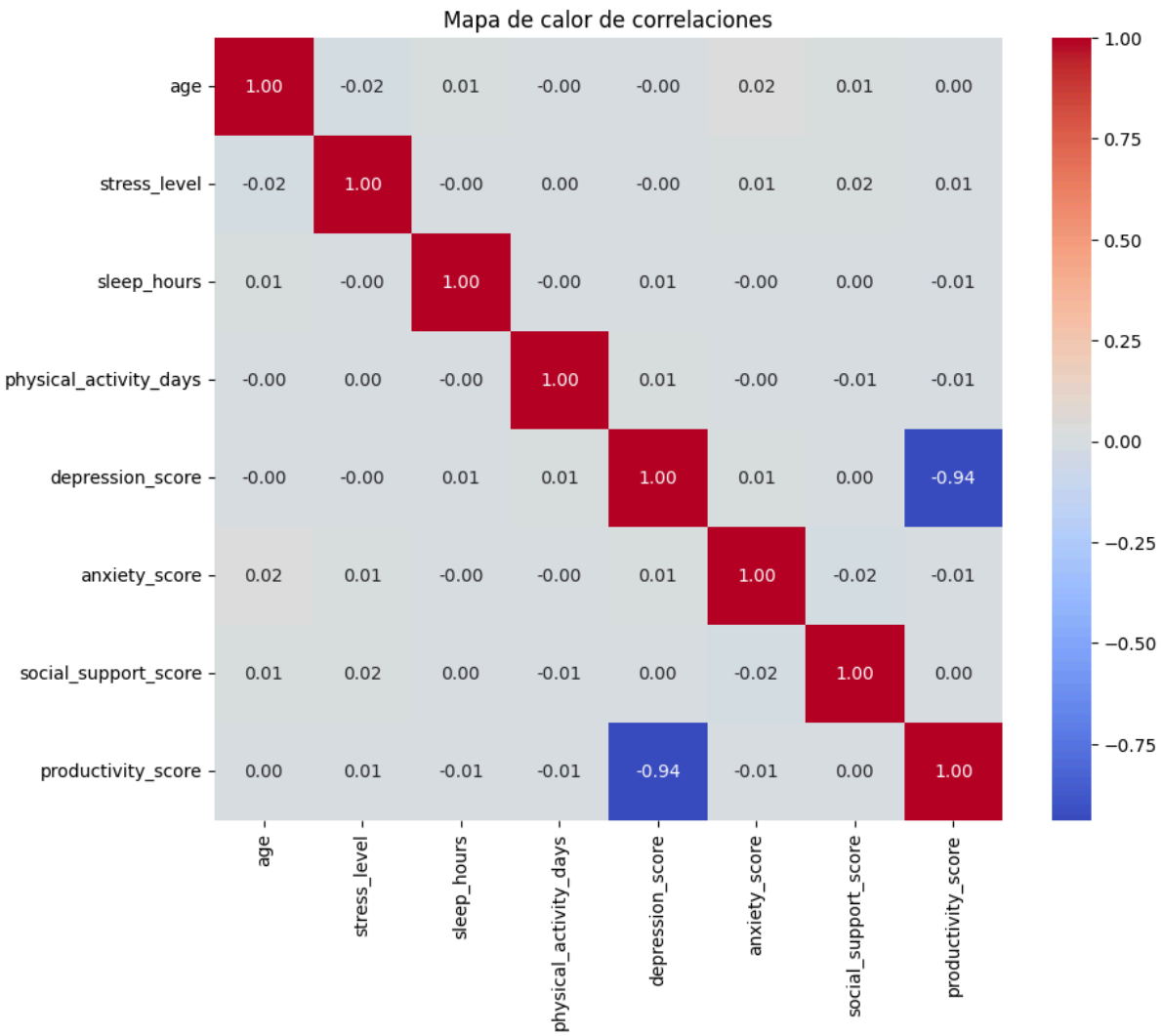


imagen: mapa de calor de las correlaciones

- **Relación productividad - riesgo**: Se observa una correlación inversa clara entre `productivity_score` y el riesgo de salud mental. Individuos con mayor productividad tienden a reportar menor riesgo.

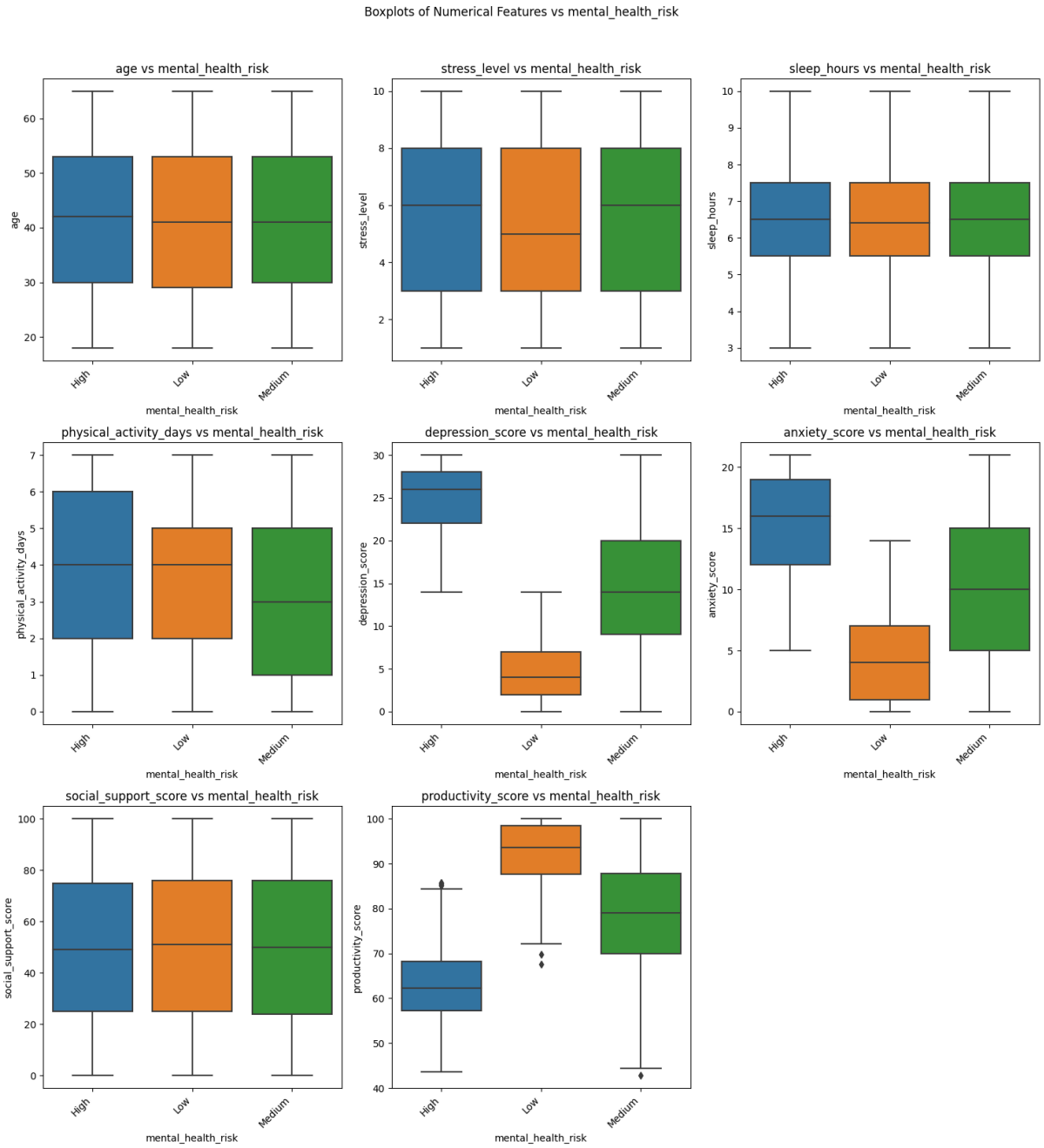


Imagen: Diagrama de cajas y bigotes de la distribuciones categóricas vs mental_healt_risk

- **Outliers**: Se detectaron valores extremos en variables como `productivity_score`, los cuales fueron inspeccionados y tratados para evitar que distorsionaran el entrenamiento del modelo.

5. Preprocesamiento

El preprocesamiento es una etapa crucial, ya que transforma los datos en un formato que los modelos pueden interpretar correctamente. Las acciones realizadas fueron:

- Codificación de la variable objetivo `seeks_treatment` en LOW, MEDIUM, HIGH, transformadas a [0,1,2] mediante LabelEncoder.

- Variables categóricas se codificaron con OneHotEncoder, generando variables dummy para representar cada categoría.
- Variables numéricas fueron escaladas mediante normalización para garantizar que todas tuvieran el mismo rango de influencia.
- Los datos finales fueron almacenados en archivos `.joblib` para facilitar su reutilización y permitir una ejecución eficiente de los modelos sin necesidad de preprocesar cada vez.

6. Modelado

Se implementaron dos modelos de clasificación:

6.1 Regresión Logística

Un modelo lineal ampliamente utilizado por su interpretabilidad y bajo costo computacional.

- **Mejores hiper parámetros**: C=100, penalty='l1'
- **F1 Macro**: 0.9317
- **Reporte de clasificación**:
    - High: 92%
    - Low: 92%
    - Medium: 95%
    - Accuracy total: 94%

Este modelo mostró buen rendimiento general, especialmente en la clase mayoritaria (Medium), aunque tuvo algunas dificultades para distinguir entre las clases Low y High.

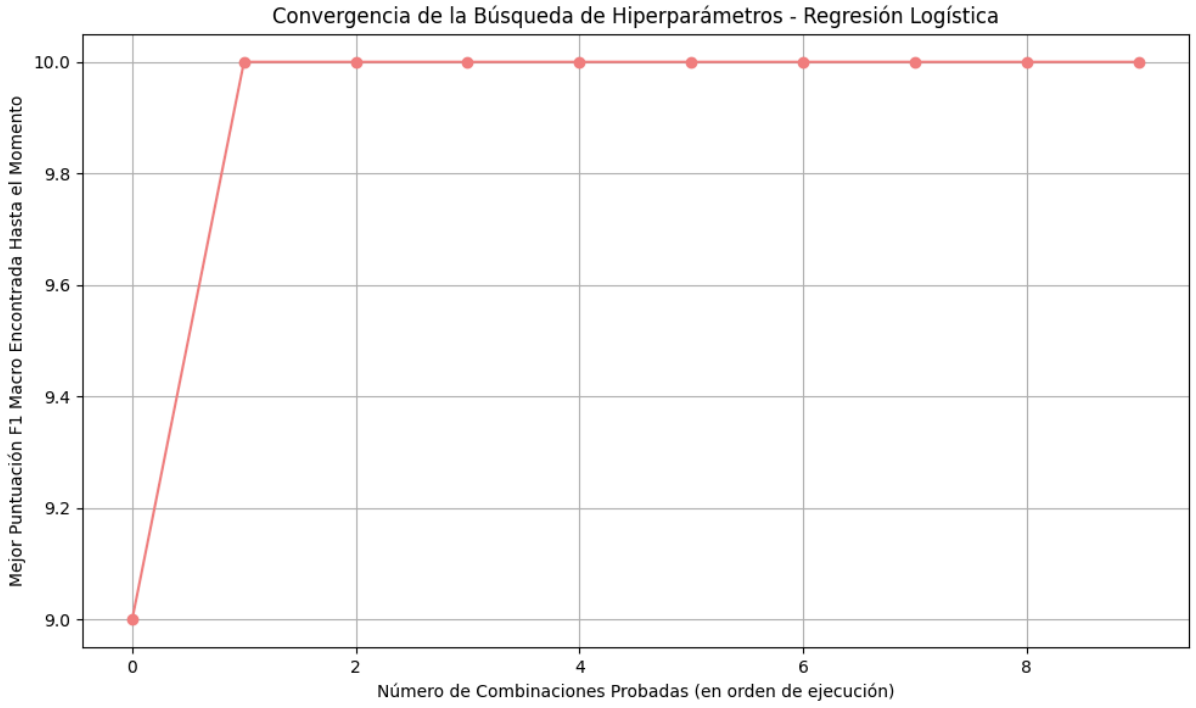


Imagen: Convergencia en la búsqueda de hyper parámetros.

6.2 Random Forest

Un ensamble de árboles de decisión que mejora la capacidad de generalización.

- **Mejores hiperparámetros**: n_estimators=100, max_depth=None, min_samples_split=2
- **F1 Macro**: 0.9402
- **Reporte de clasificación**:
    - High: 94%
    - Low: 93%
    - Medium: 96%
    - Accuracy total: 95%

Aunque Random Forest requirió más tiempo de entrenamiento, obtuvo mejores resultados en todas las clases, mostrando mayor robustez frente a la complejidad de los datos.

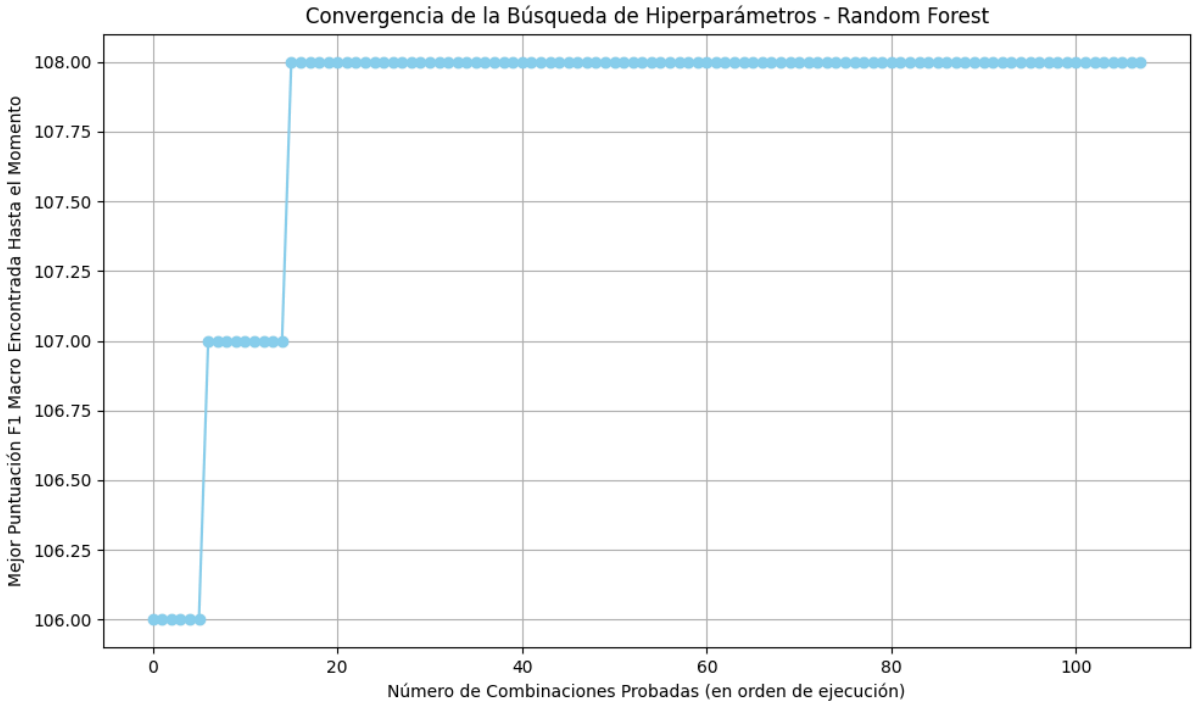


Imagen: Convergencia en la búsqueda de hyper parámetros.

7. Comparación de Modelos

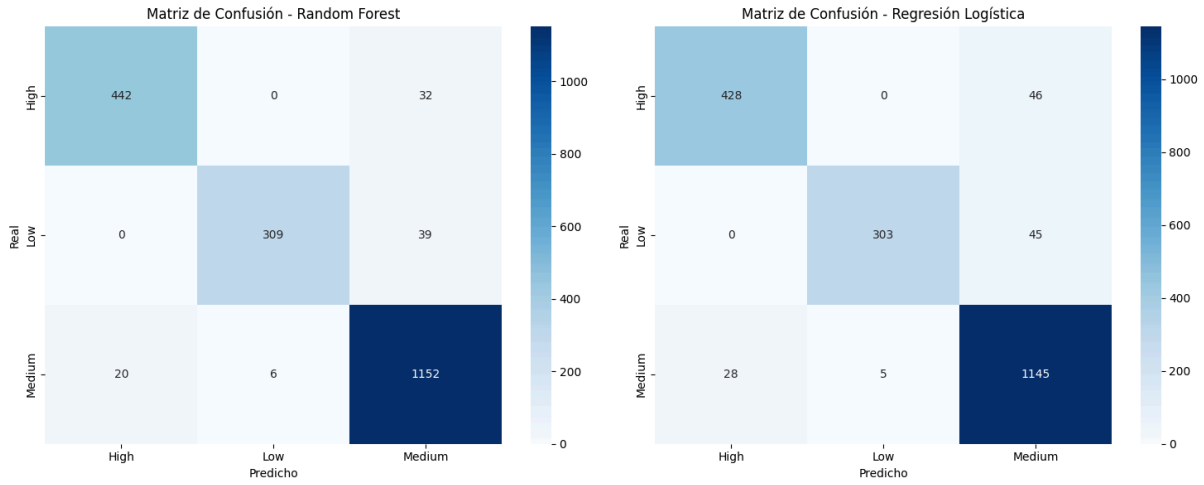
Para comparar el desempeño de los modelos se utilizaron diversas métricas:

- Matrices de confusión
- F1 Macro
- Kappa de Cohen (0.9220): alto nivel de acuerdo entre ambos modelos
- Índice de Jaccard (0.9072): indica similitud de predicciones

Análisis de errores:

- **Errores comunes** (ambos fallan): 68 instancias
- **Errores únicos de RF**: 29 (1.45%)
- **Errores únicos de LR**: 56 (2.8%)

Se concluye que Random Forest no solo tuvo un mejor rendimiento global, sino que también cometió menos errores críticos.



Análisis de las Matrices de Confusión

Las matrices de confusión comparan las predicciones del modelo frente a los valores reales. En este caso, se presentan para dos modelos:

- **Izquierda**: Random Forest
- **Derecha**: Regresión Logística

Cada matriz tiene tres clases: `High`, `Low` y `Medium` riesgo de salud mental.

🔵 Random Forest

- **Clase High**:
    - 442 predicciones correctas
    - 32 personas con riesgo High fueron clasificadas como Medium
    - Precisión sobresaliente en esta clase, con solo ~7% de error.
- **Clase Low**:
    - 309 correctamente clasificadas
    - 39 confundidas con Medium
    - 100% de precisión en evitar falsos positivos para la clase High.
- **Clase Medium**:
    - 1152 predicciones correctas
    - 26 errores distribuidos entre High (20) y Low (6)

🔹 **Conclusión**: Random Forest tiene excelente precisión general, especialmente al diferenciar entre `High` y `Low`. La mayor confusión ocurre en la frontera `Low–Medium`.

8. Conclusiones

- Ambos modelos tienen un excelente rendimiento, pero **Random Forest** demostró mejor adaptabilidad y menor tasa de error.
- Las variables con mayor influencia fueron: `mental_health_history`, `stress_level`, `productivity_score` y `social_support_score`.
- El análisis evidenció que las personas con apoyo social bajo y altos niveles de estrés presentan un riesgo significativamente mayor.
- Este sistema tiene potencial para ser usado en procesos de detección temprana en entornos empresariales y clínicos, facilitando decisiones proactivas en salud ocupacional.

9. Recomendaciones

- Ampliar el dataset con datos reales de diversas industrias para mejorar la generalización.
- Explorar modelos avanzados como XGBoost, LightGBM o redes neuronales profundas.
- Incorporar una etapa de selección de atributos automatizada para reducir dimensionalidad y ruido.
- Desarrollar una interfaz visual que permita aplicar el modelo de forma intuitiva en entornos reales, como empresas o instituciones educativas.