| | |
|---|---|
| Autores:<br>Fabiola Contreras, 22787<br>María José Villafuerte, 22129 | Docente: Luís Roberto Furlán Collver<br>Laboratorio 7 |
| Sección: 21 | Fecha: 28/09/2025 |

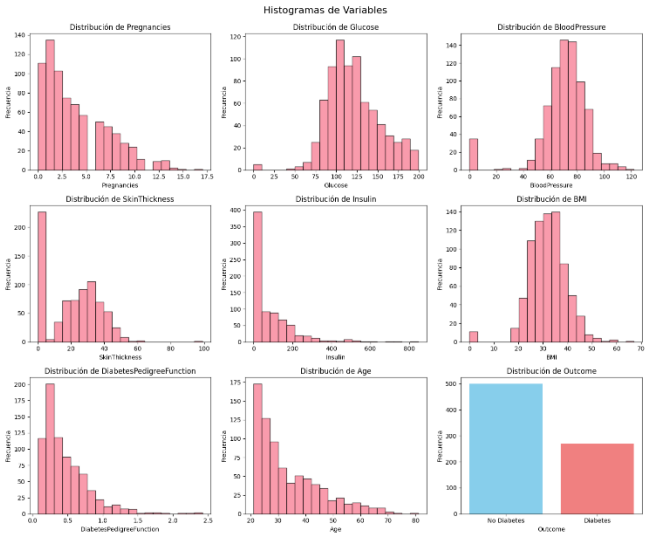

# Laboratorio 7. Predicción de datos con AutoGluon

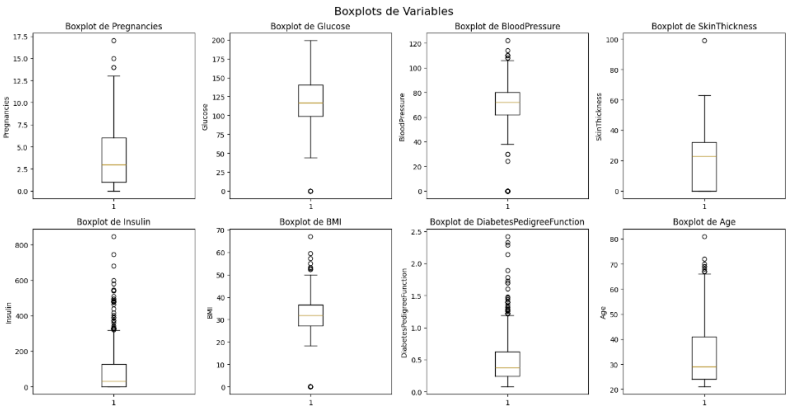
## I. Análisis exploratorio

### 1. Obtenga estadísticas descriptivas básicas del conjunto de datos.

Este dataset contiene 768 registros de pacientes con 9 variables predictivas relacionadas con el diagnóstico de diabetes, incluyendo factores demográficos, clínicos y genéticos. Las variables están principalmente en formato entero (int64) excepto el BMI y la función de pedigree de diabetes que son decimales (float64), lo que permite un análisis cuantitativo preciso. La muestra de las primeras 5 filas muestra una diversidad en los perfiles de pacientes, desde una mujer de 50 años con múltiples embarazos y diabetes confirmada, hasta una joven de 21 años sin diabetes, lo que sugiere un dataset balanceado para entrenar modelos predictivos.

### 2. Visualice la distribución de las variables con histogramas y boxplots.




Los histogramas revelan que la mayoría de variables presentan distribuciones asimétricas hacia la derecha, siendo especialmente notable en Pregnancies, SkinThickness, Insulin y DiabetesPedigreeFunction, mientras que Glucose, BloodPressure y BMI muestran distribuciones más cercanas a la normalidad. Los boxplots confirman estas asimetrías y evidencian la presencia significativa de valores atípicos en todas las variables, particularmente extremos en Insulin (hasta 846) y algunos outliers en Age, Glucose y BloodPressure que podrían requerir atención especial durante el análisis. La variable Outcome muestra un desbalance de clases con aproximadamente 65%

de pacientes sin diabetes (500 casos) versus 35% con diabetes (268 casos), lo que es típico en datasets médicos y deberá considerarse en el modelado predictivo.

### 3. Verifique la presencia de valores nulos o atípicos y decida cómo manejarlos.

El análisis revela un problema significativo de calidad de datos donde los valores cero probablemente representan datos faltantes codificados incorrectamente, ya que es biológicamente imposible tener glucosa, presión arterial o BMI en cero. La situación es particularmente crítica en Insulin (48.7% de ceros) y SkinThickness (29.6% de ceros), lo que sugiere que estos valores faltantes fueron sistemáticamente registrados como ceros durante la recolección de datos. Este hallazgo contradice la aparente "completitud" del dataset y requiere estrategias de tratamiento de datos faltantes, como imputación por mediana/media o eliminación de registros, antes de proceder con cualquier análisis predictivo para evitar sesgos en los resultados.

### 4. Analice el balance de clases de la variable 'Outcome'.

El dataset presenta un desbalance moderado de clases con 500 pacientes sin diabetes (65.1%) versus 268 pacientes con diabetes (34.9%), reflejando una distribución realista donde la diabetes afecta a una minoría significativa pero no mayoritaria de la población estudiada. Aunque este desbalance no es extremo (no supera el 70-30%), puede influir en el rendimiento de los algoritmos de machine learning, potencialmente sesgándolos hacia la predicción de la clase mayoritaria (no diabetes). Para obtener modelos predictivos más robustos será recomendable considerar técnicas de balanceeo como SMOTE, ajuste de pesos de clase o métricas de evaluación que no dependan únicamente de la precisión global, como el F1-score o el área bajo la curva ROC.

### 5. Genere una matriz de correlación y un mapa de calor para identificar relaciones entre variables.

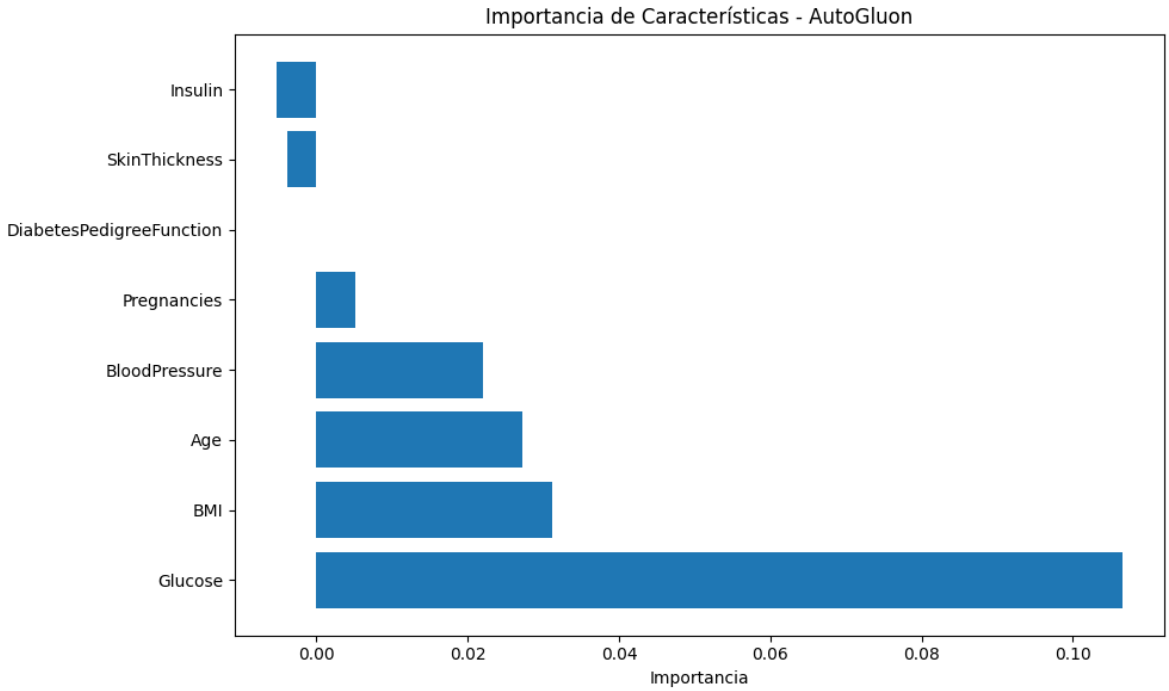
La matriz de correlación revela que Glucose presenta la correlación más fuerte con el Outcome diabético (0.47), confirmando su importancia como predictor primario, seguido por BMI (0.29), Age (0.24) y Pregnancies (0.22) como factores de riesgo secundarios. La correlación más alta del dataset se observa entre Pregnancies y Age (0.54), lo que es biológicamente coherente, junto con correlaciones moderadas entre variables físicas como Insulin-SkinThickness (0.44) y BMI-SkinThickness (0.39). En general, las correlaciones son de débiles a moderadas (rango 0.1-0.5), lo que sugiere que las variables aportan información complementaria sin problemas severos de multicolinealidad, facilitando la construcción de modelos predictivos robustos que puedan capturar patrones complejos en los datos.

## II. Evaluación y entrenamiento del modelo

En un límite de cinco minutos, AutoGluon fue capaz de generar y probar 37 modelos distintos. Para usarlo aplicamos los parámetros:

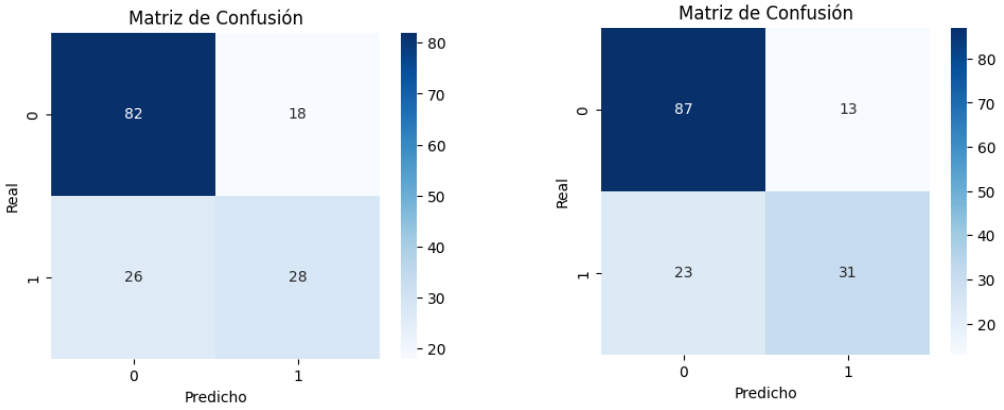
- **preset='best_quality'**: Configuración que prioriza la calidad del modelo sobre la velocidad
- **eval_metric='accuracy'**: Métrica de evaluación principal
- **time_limit=300**: Límite de tiempo de 5 minutos para el entrenamiento
- **verbosity=2**: Nivel de detalle en la salida

Al enfocarnos en los resultados prácticos que nos da AutoGluon para crear un buen modelo, una herramienta muy útil es el análisis de feature importance. En este caso reveló que la glucosa es por mucho la característica más determinante para la predicción, con una importancia superior a 0.10, seguida a distancia considerable por el BMI (~0.03), la edad (~0.025) y la presión arterial (~0.022). Esta jerarquía es clínicamente coherente, ya que los niveles de glucosa son el indicador directo más confiable para el diagnóstico de diabetes, mientras que factores como el índice de masa corporal, la edad y la presión arterial actúan como factores de riesgo secundarios. Es notable que variables como el número de embarazos, la función del pedigree diabético, el grosor de la piel y los niveles de insulina muestran una importancia considerablemente menor, lo que sugiere que, aunque pueden aportar información adicional, su contribución individual a la predicción es limitada comparada con los biomarcadores principales.



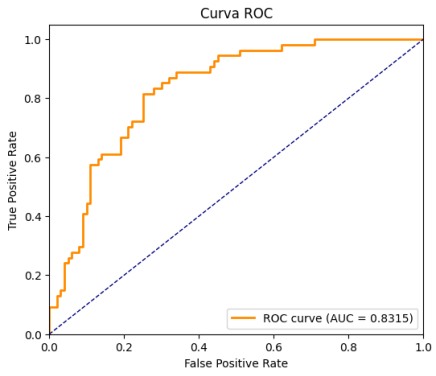
Ahora, enfocándonos en los modelos resultantes podemos comparar AutoGluon con un modelo base de regresión logística que resultó en un accuracy de 0.7143. Al comparar los resultados de los modelos de AutoGluon el mejor modelo obtenido tuvo un accuracy de 0.7866.

Estas son las matrices de confusión del modelo base (izquierda) vs. el mejor modelo de AutoGluon (derecha).




En esta comparación es posible observar que AutoGluon redujo los falsos positivos de 18 a 13 (27.8% de reducción), lo que significa menos casos sanos clasificados incorrectamente como enfermos. También disminuyó los falsos negativos de 26 a 23 (11.5% de reducción), reduciendo el riesgo crítico de no detectar casos positivos reales.

Este modelo seleccionado por AutoGluon demostró un rendimiento sólido y equilibrado, alcanzando una accuracy global del 76.62% con métricas balanceadas entre ambas clases. Para la clase negativa (no diabetes), el modelo logró una precisión del 79.09% y un recall del 87%, resultando en un F1-score de 0.8286, lo que indica una excelente capacidad para identificar correctamente los casos sanos. Para la clase positiva (diabetes), aunque las métricas son menores debido al desbalance natural del dataset, el modelo mantuvo un desempeño respetable con una precisión del 70.45% y un recall del 57.41%, generando un F1-score de 0.6327.




La curva ROC confirma la calidad predictiva del modelo con un AUC de 0.8315, clasificándose en el rango de "buena" capacidad discriminativa. La forma de la curva muestra una rápida elevación en las primeras etapas (baja tasa de falsos positivos), alcanzando aproximadamente un 60% de sensibilidad con solo un 20% de falsos positivos, eso demuestra la eficacia del modelo para distinguir entre clases. El AUC de 0.8315 supera considerablemente el valor de 0.5 (clasificador aleatorio), indicando que el modelo posee una capacidad predictiva superior al azar y es adecuado para aplicaciones prácticas de screening y diagnóstico asistido.

# III. Conclusiones

- AutoGluon logró un accuracy de 78.66% superando en 14.23% al modelo base de regresión logística. En solo 5 minutos evaluó 37 modelos diferentes, demostrando alta eficiencia en el desarrollo de soluciones predictivas.
- El modelo redujo falsos positivos en 27.8% y falsos negativos en 11.5%, aspectos cruciales en diagnóstico médico. El AUC-ROC de 0.8315 confirma una capacidad discriminativa "buena" para aplicaciones clínicas.
- AutoGluon confirmó que Glucose es el predictor dominante (importancia >0.10), seguido por BMI, Age y BloodPressure. Esta jerarquía coincide perfectamente con el conocimiento médico establecido sobre factores de riesgo diabético.
- Se detectó que 48.7% de valores cero en Insulin y 29.6% en SkinThickness representan datos faltantes mal codificados. Este hallazgo es crítico para futuras mejoras en la recolección y preprocesamiento de datos médicos.
- El modelo presenta 87% de especificidad y 57% de sensibilidad, apropiado para screening inicial. Sin embargo, requiere validación externa con múltiples poblaciones antes de implementación clínica real.
- AutoGluon automatizó exitosamente la selección de algoritmos y optimización de hiperparámetros sin intervención manual. No obstante, los modelos ensemble resultantes sacrifican interpretabilidad, aspecto crucial en aplicaciones médicas.

# IV. Reflexión

¿Qué ventajas y desventajas encontró al usar AutoGluon y AutoML en general?

Como ventaja principal el uso de estas herramientas nos permite automatizar muchísimo el proceso, eliminando la necesidad de seleccionar manualmente algoritmos, optimizar hiperparámetros y realizar la elección de características. Además, es un proceso bastante sencillo, que requiere mínima configuración y da más tiempo para interpretar los resultados. Permite reducir el tiempo de desarrollo al combinar múltiples herramientas y automatizar las tareas repetitivas.

Por otro lado, al usar esta herramienta tenemos menor flexibilidad de implementar conocimiento específico y hay menos control en las decisiones del modelo final realizado. A parte, requiere mayor poder computacional y memoria comparado con modelos simples.

¿Qué métricas considera más relevantes en este problema?

Según lo evaluado, el recall es muy importante para minimizar falsos negativos, el no detectar diabetes puede traer consecuencias muy graves. El F1-Score nos permite tener un balance entre precisión y recall, y AUC-ROC, evalúa la capacidad del modelo de discrimar.

¿Qué precauciones deberían tomarse al aplicar este tipo de herramientas en salud?

Primordialmente, es importante considerar que los resultados deben ser siempre interpretados por profesionales de salud capacitados y los modelos deben ser validados con datos de múltiples poblaciones y centros médicos antes de implementación.

Al usar específicamente AutoGluon es necesaria la verificación y no limitarse a presentar únicamente el modelo resultante por la herramienta, siempre se puede seguir mejorando en base a lo que resultó. Este debe implementar métodos de interpretabilidad para que los médicos entiendan las bases de las predicciones y establecer sistemas de seguimiento del rendimiento del modelo en producción, para asegurar el buen funcionamiento en todo momento.

¿Cómo compara esta experiencia con construir un modelo manualmente?

AutoGluon es mucho mejor para obtener resultados rápidos y de alta calidad, mientras que hacerlo manualmente sería preferible cuando se requiere máxima interpretabilidad o cuando existen restricciones específicas del dominio médico. Como mencionamos antes, la combinación ideal sería usar AutoGluon para establecer una línea base sólida y luego refinar manualmente aspectos críticos para la aplicación clínica.

# V. GitHub

Link a repositorio: https://github.com/Maria-Villafuerte/Labs_Data_Science/tree/main/Lab_7