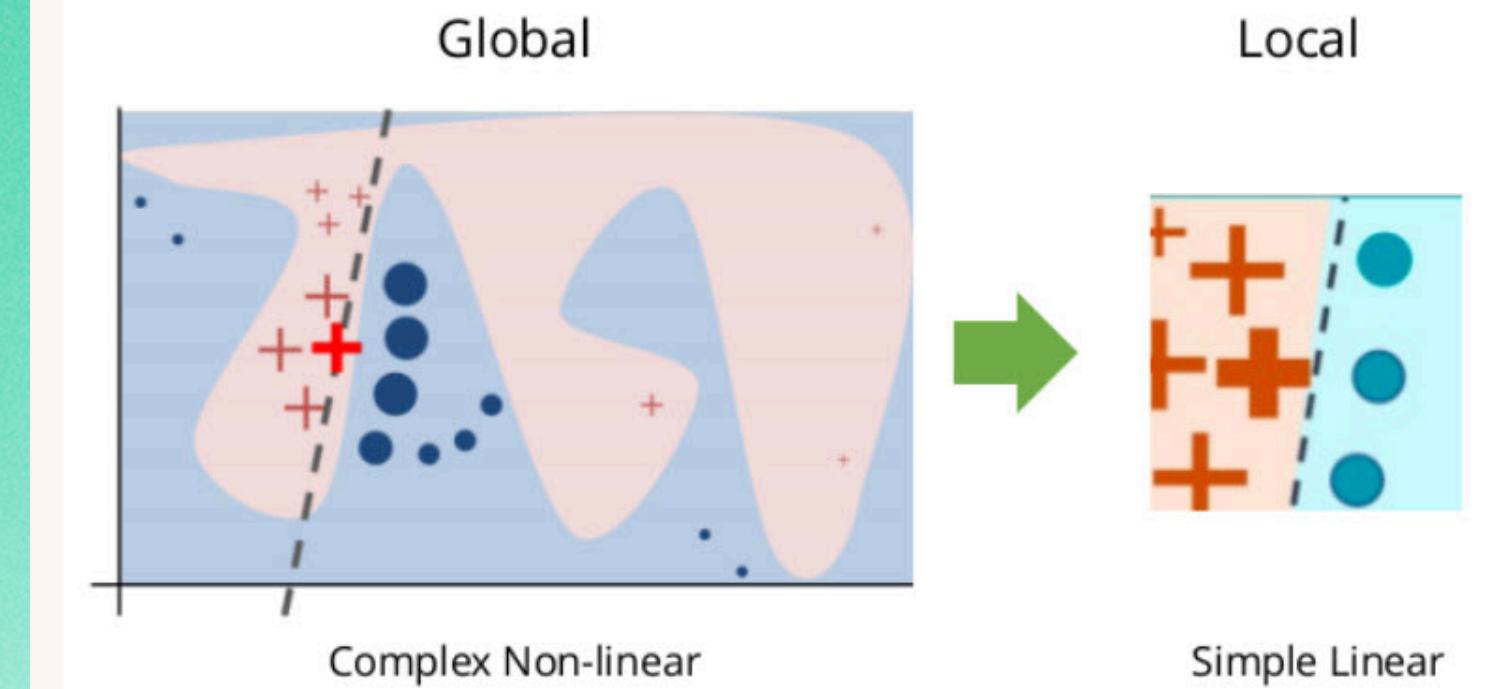


XAI: SHAP Y LIME



índice

- Motivación: problema y contexto
- SHAP y LIME
- Demostraciones
- Aplicaciones: casos reales
- Conclusión



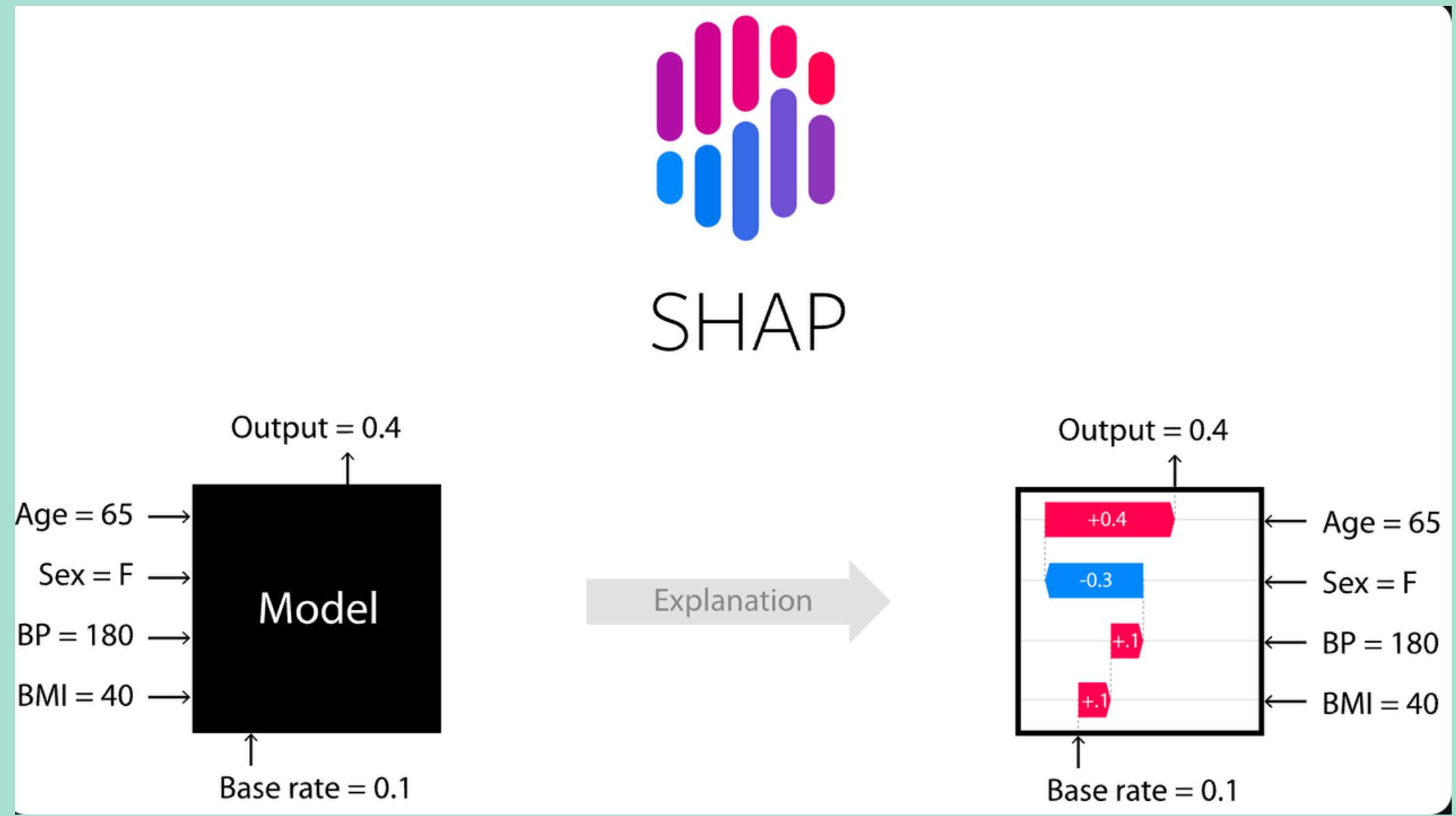
Motivación: problema y contexto

- Los modelos de aprendizaje automático complejos (redes profundas, ensamblados) son muy precisos pero opacos, generando cajas negras. ¿Cómo toman sus decisiones?
- La falta de interpretabilidad puede ser crítica en algunos ámbitos
- Objetivo XAI (Inteligencia Artificial Explicable): hacer que los modelos sean mas transparentes y comprensibles



SHAP

- Entender cómo y por qué toma la decisión
- Basado en valores de Shapley
- Los valores SHAP miden cuánto contribuye cada característica, (ingresos, la edad, etc.) a la predicción del modelo
- Pueden ayudar a ver qué características son las más importantes para el modelo y cómo afectan al resultado.



$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]$$

Ventajas y desafíos



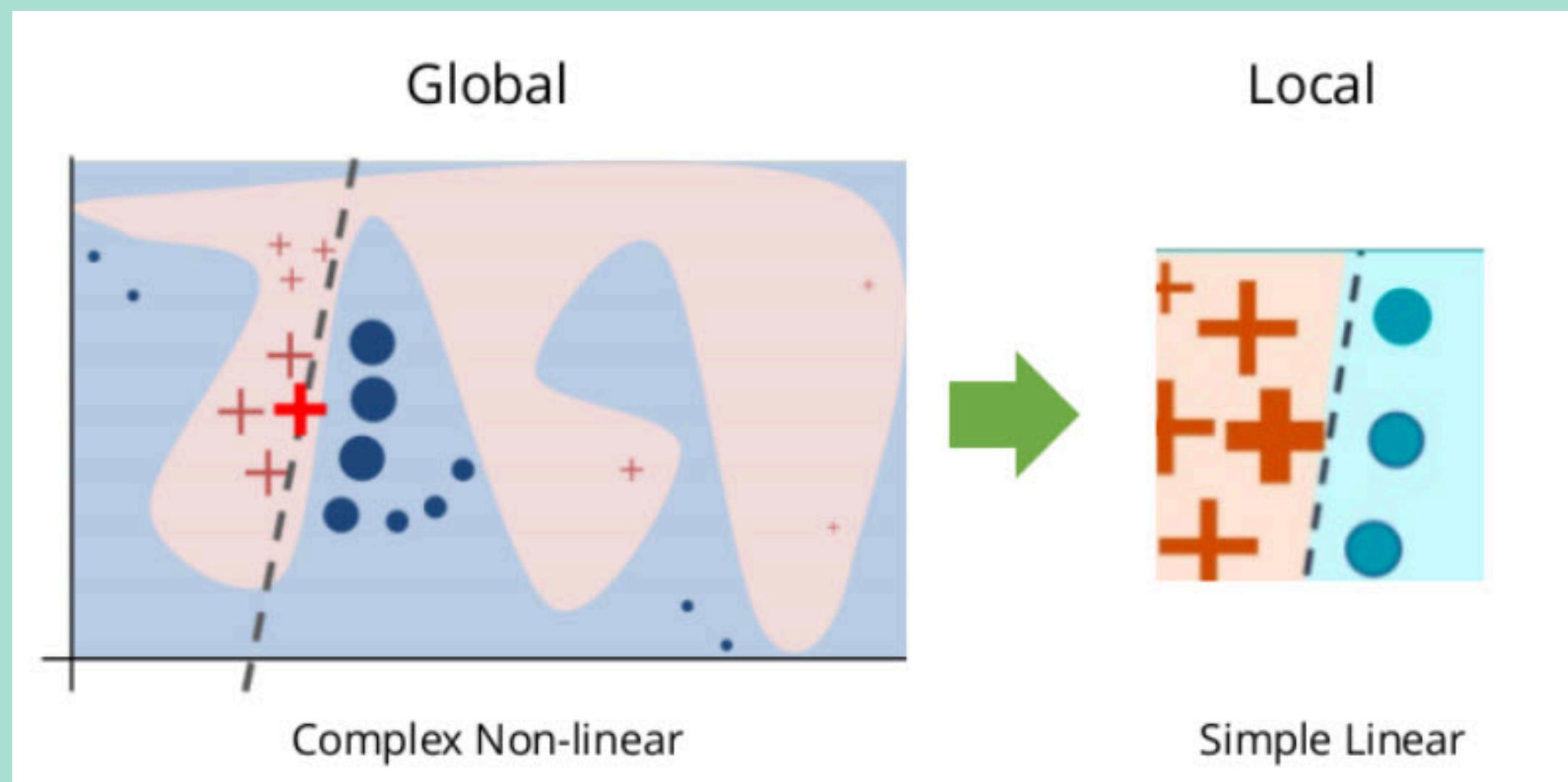
- Si una característica tiene mayor impacto en la predicción, recibe un valor proporcionalmente mayor
- Puede explicar tanto una sola predicción (nivel local) como el comportamiento general del modelo (nivel global).
- Existen versiones optimizadas, como Tree SHAP, que lo hacen eficiente para modelos basados en árboles



- Costo computacional: calcular todas las combinaciones posibles es lento en modelos grandes, aunque existen optimizaciones como Tree SHAP para modelos basados en árboles.

LIME

- Centrado en explicar predicciones individuales
- Crea un modelo simple alrededor de esa predicción, lo que lo hace más fácil de entender, pero no siempre tan preciso



Ventajas y desafíos



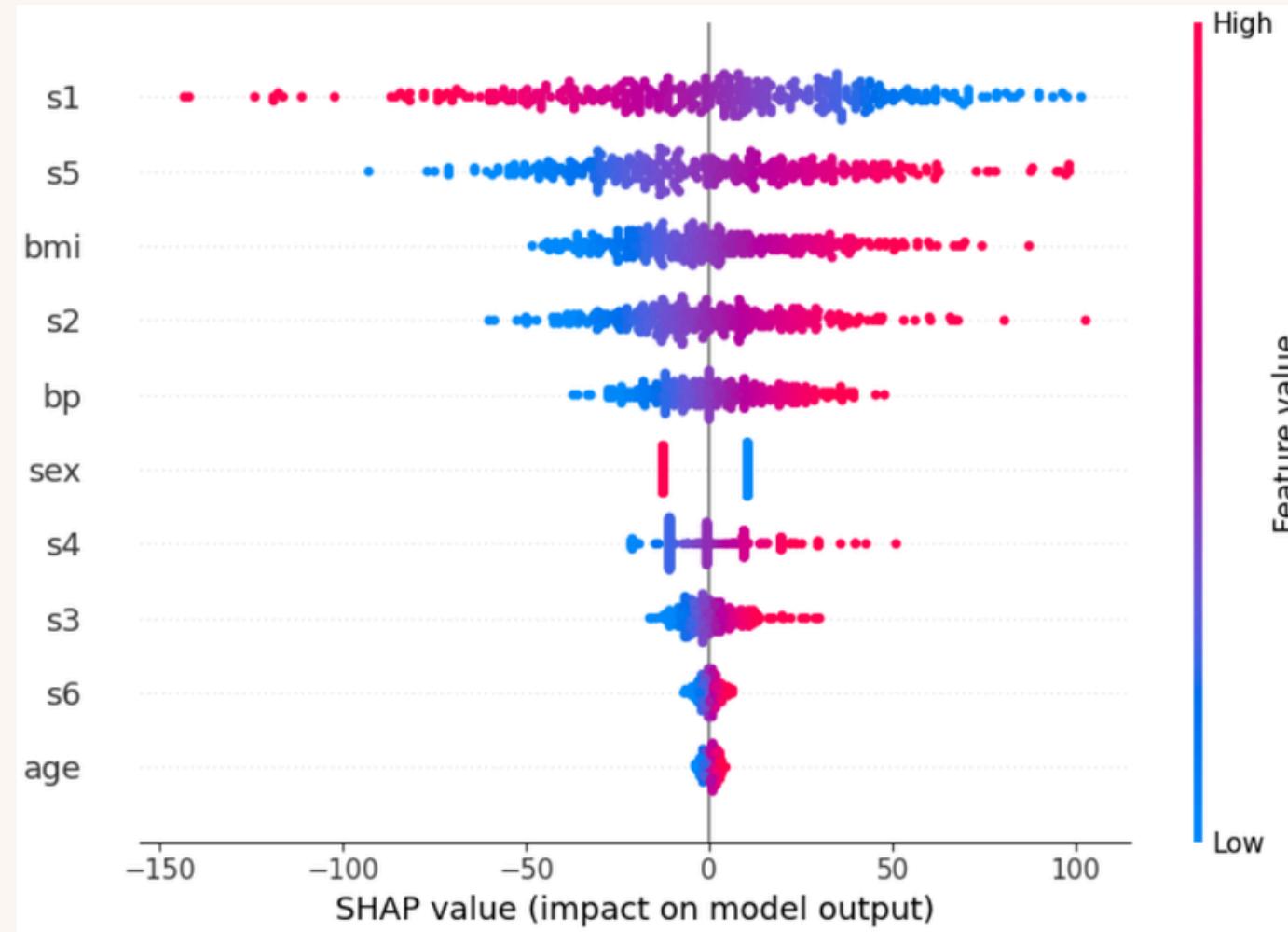
- Puede trabajar con cualquier tipo de modelo de aprendizaje automático
- Al construir modelos locales simples, las explicaciones generadas son fáciles de entender incluso para usuarios sin experiencia técnica.
- Comparado con técnicas más complejas como SHAP, LIME tiende a ser más rápido en la generación de explicaciones locales, ya que no requiere evaluar todas las combinaciones posibles de características.



- Dado que LIME se enfoca en explicaciones locales, no proporciona una visión general del comportamiento del modelo completo

Demostraciones

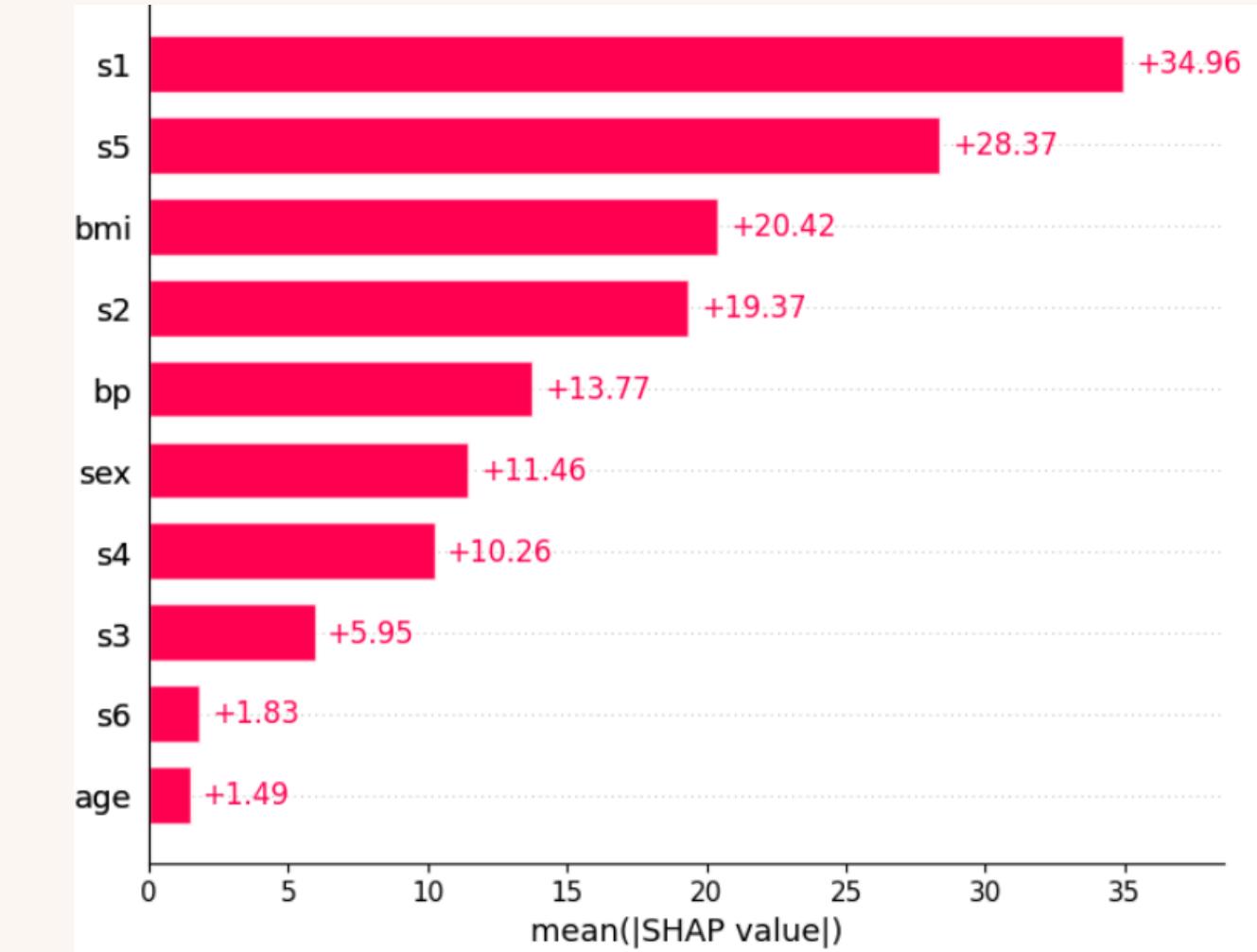
1. Demostración 1 SHAP



Este gráfico muestra cómo cada característica afecta la predicción del modelo.

Cada punto representa un paciente.

Si el punto está a la derecha, esa característica aumenta la predicción. Si está a la izquierda, la disminuye.

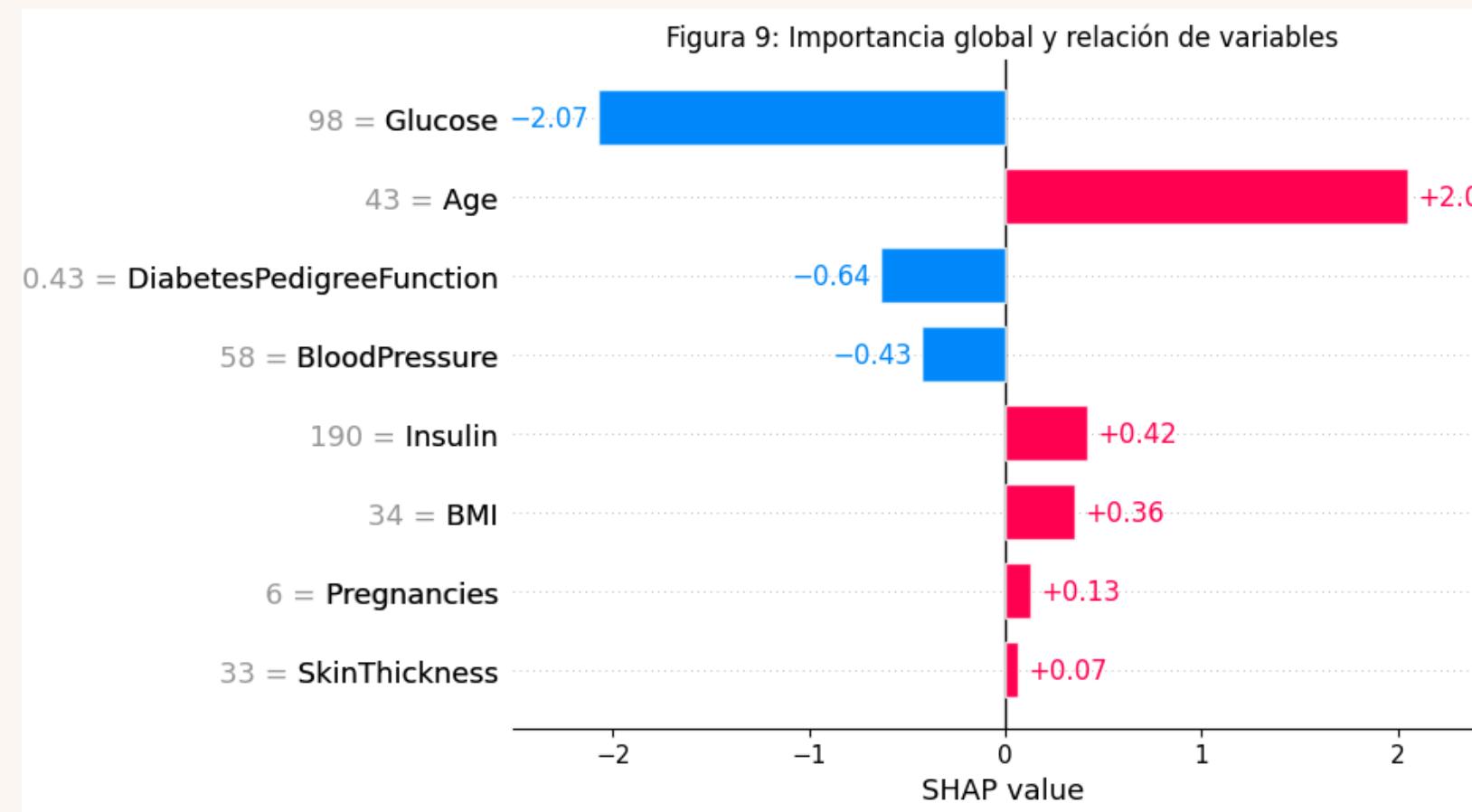


Este gráfico muestra qué características son más importantes para el modelo. Cuanto más alta la barra, más influencia tiene esa característica en las predicciones.

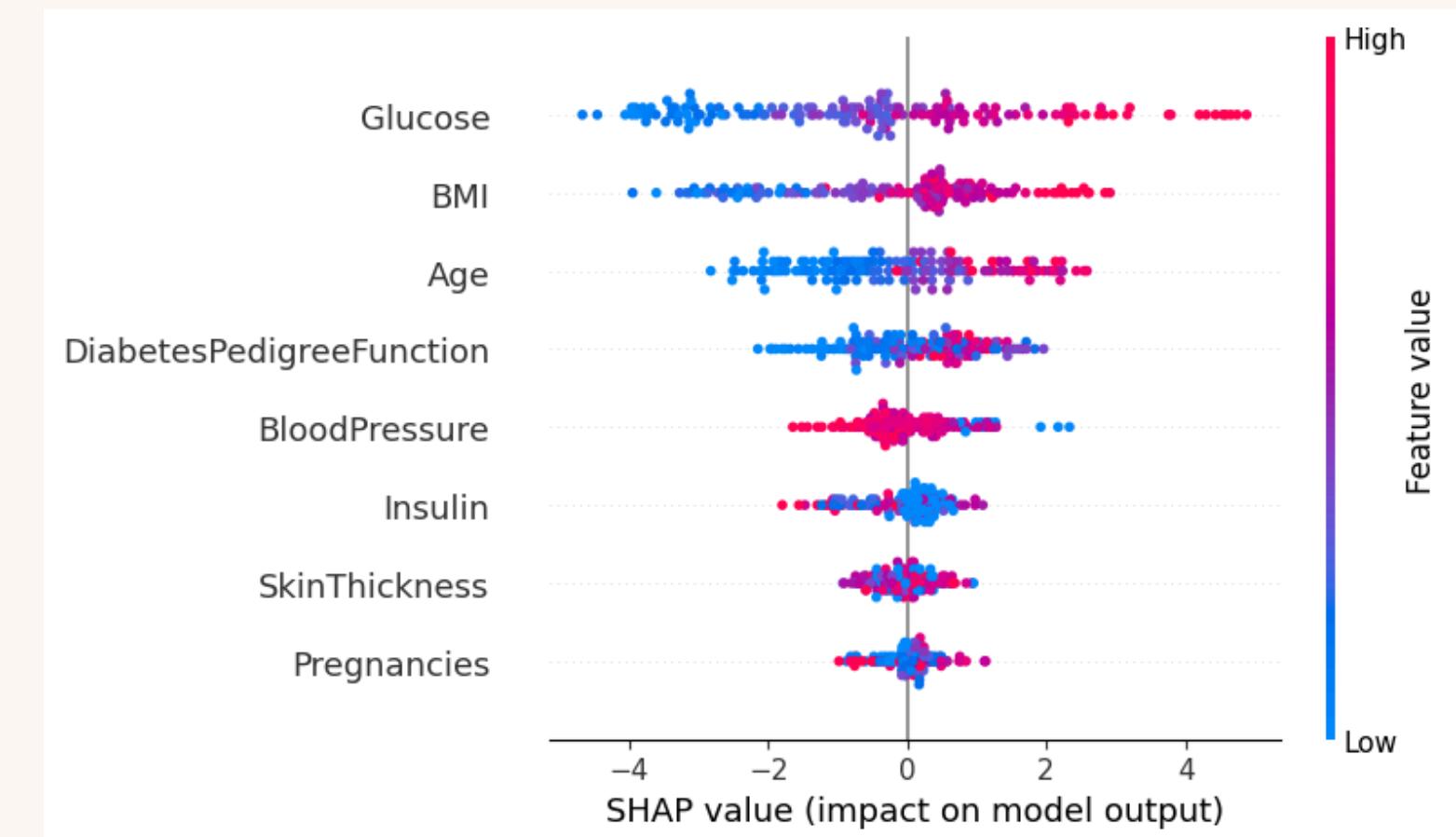
Por ejemplo, 's1' es la más influyente .

Demostraciones

2. Demostración 2: Interpretabilidad local y global con SHAP



Este gráfico muestra cómo cada característica afecta la predicción para un paciente específico.
Por ejemplo, la glucosa baja reduce la predicción, mientras que la edad alta la aumenta.
Cada barra indica si la característica empuja la predicción hacia arriba o hacia abajo.



Este gráfico es parecido al anterior, pero ahora usamos un modelo más complejo. Lo importante aquí es que SHAP sigue mostrando cómo cada característica afecta la predicción, incluso en modelos no lineales

Demostraciones

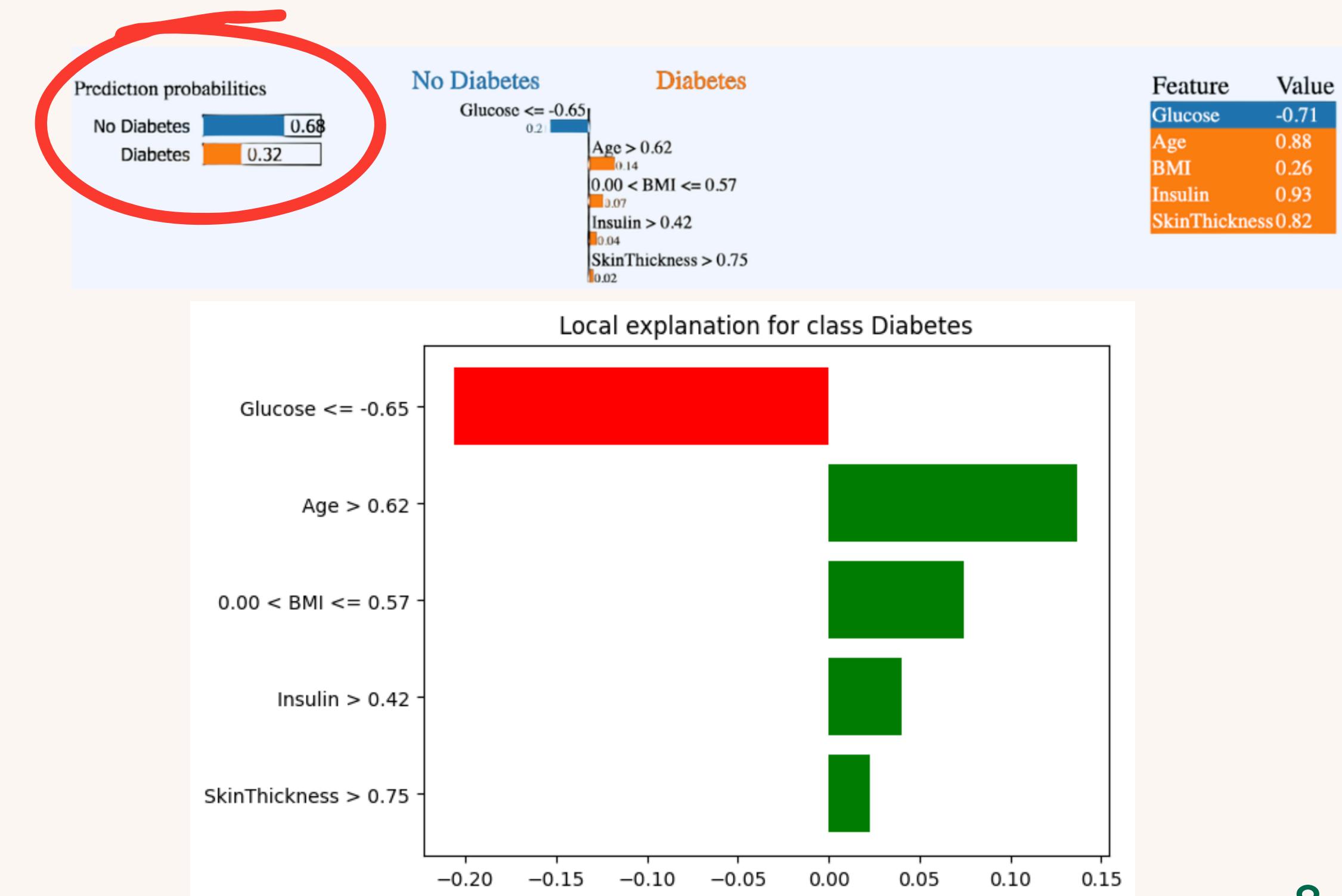
3. Demostración 3: Explicación local con LIME

Entrenamos un modelo Random Forest para predecir diabetes. Seleccionamos un paciente del conjunto de prueba y usamos LIME para explicar esa predicción.

Para este paciente, el modelo predice 32% de diabetes y 68% de no diabetes.

Una edad elevada, niveles altos de insulina, un BMI alto y un mayor grosor de piel incrementan la probabilidad de diabetes; en cambio, una glucosa baja la reduce.

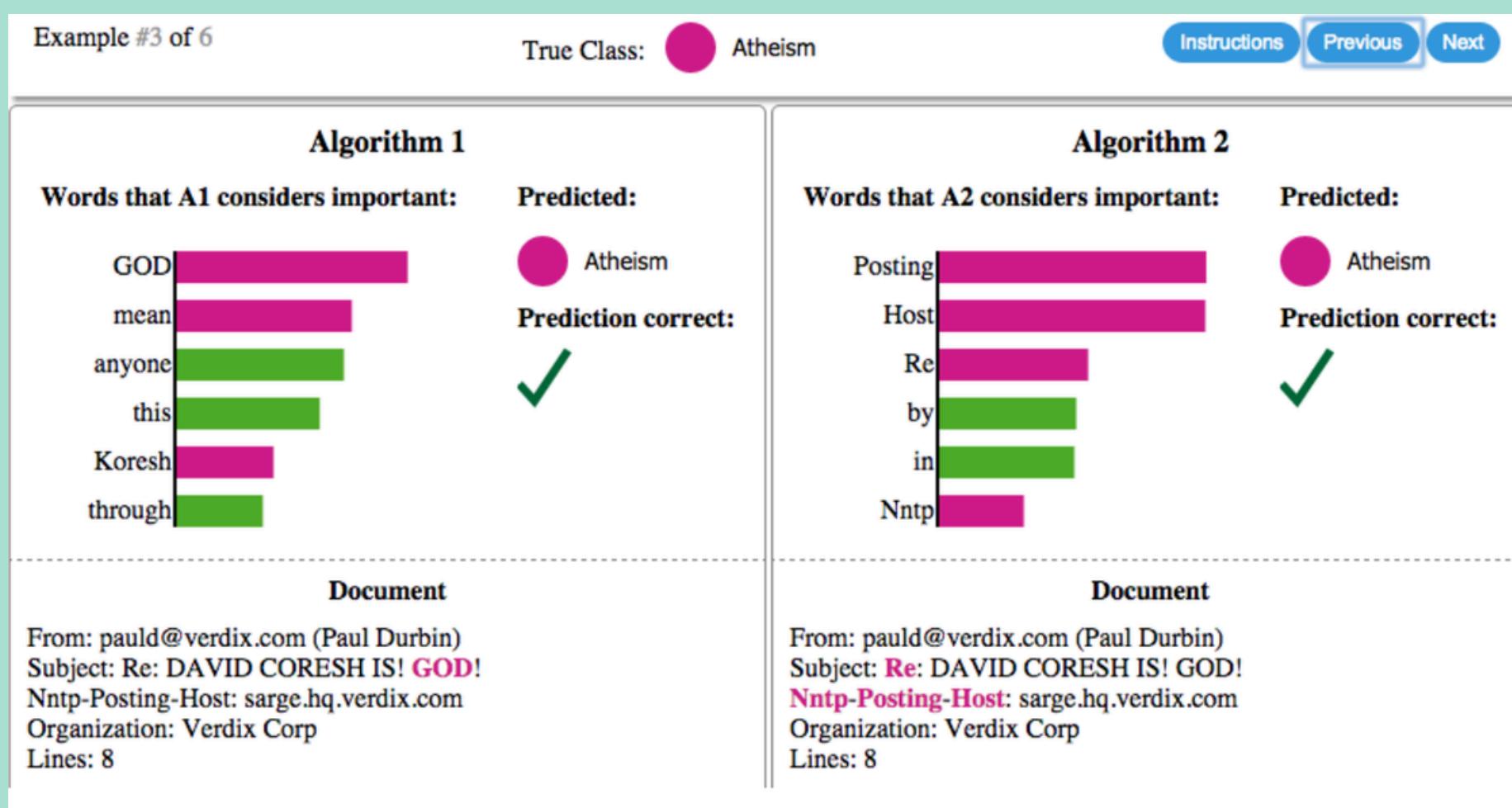
El gráfico muestra las reglas que LIME encontró para este paciente.



Aplicaciones de LIME

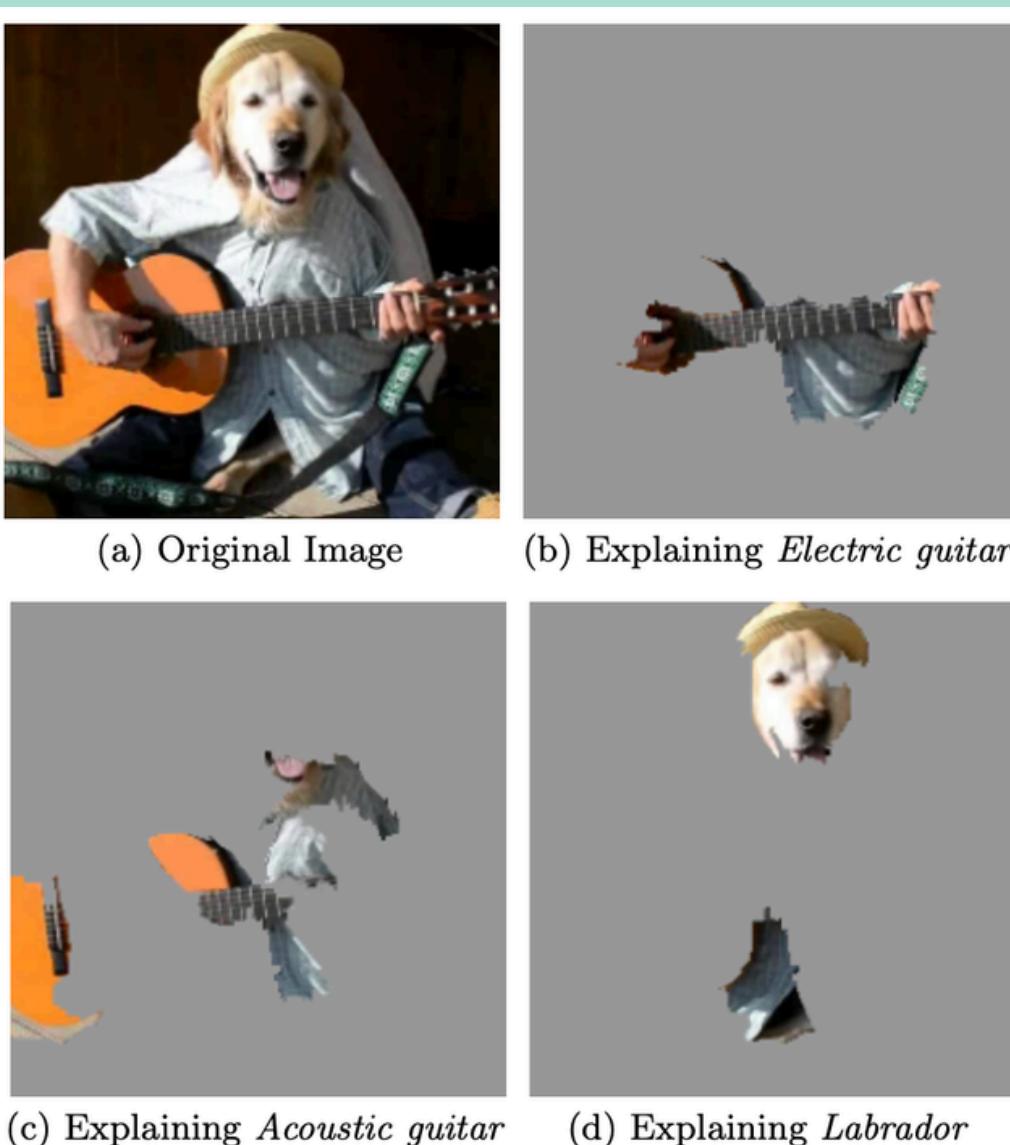
- Aplicaciones en clasificación de texto

LIME puede exponer sesgos invisibles en métricas



- Aplicaciones en clasificación de imágenes

LIME permite ver que el modelo identifica objetos correctamente y no confunde perros con guitarras



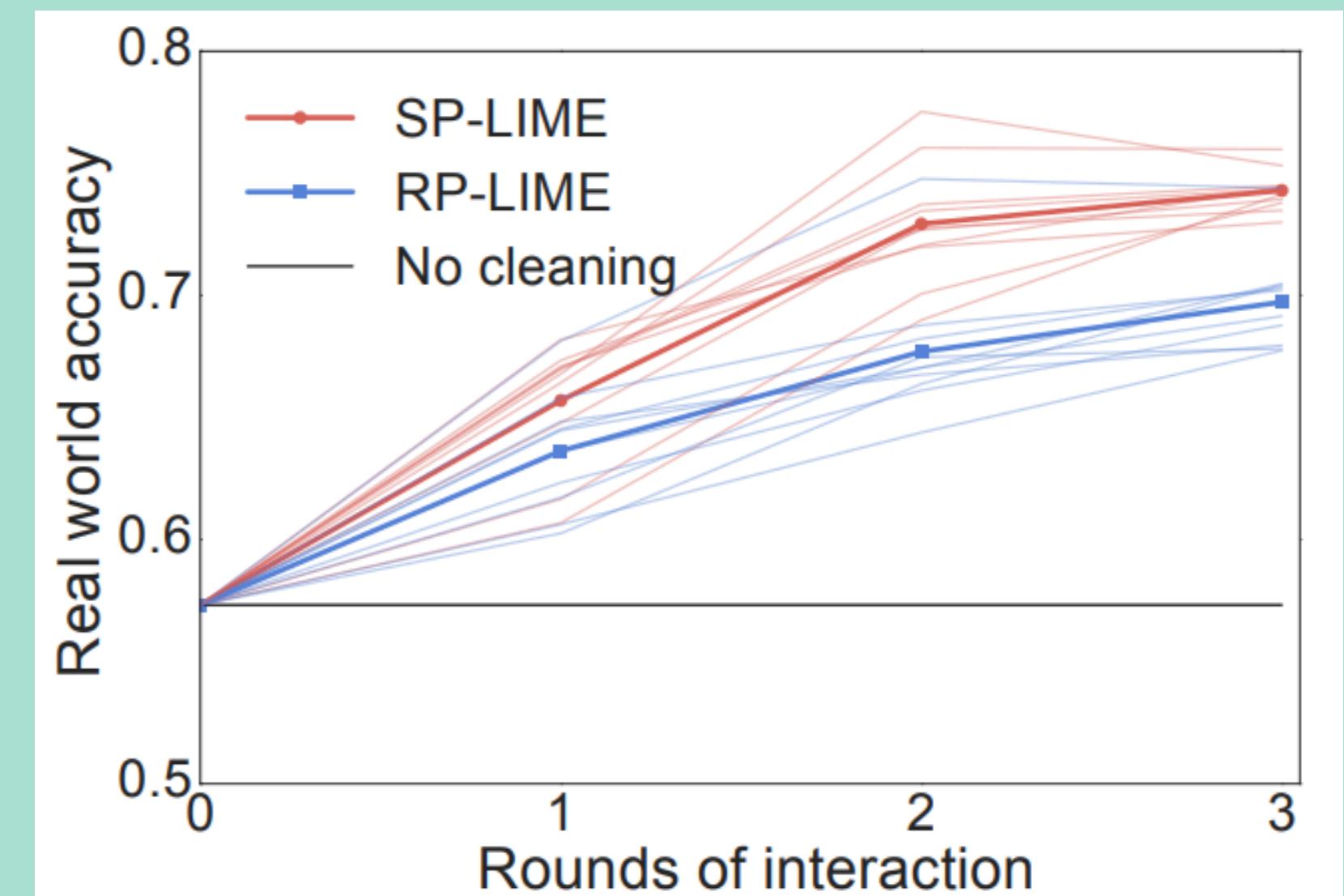
Aplicaciones de LIME

- Evaluación de la confianza en predicciones individuales

Las explicaciones de LIME permiten juzgar confianza mejor que adivinar

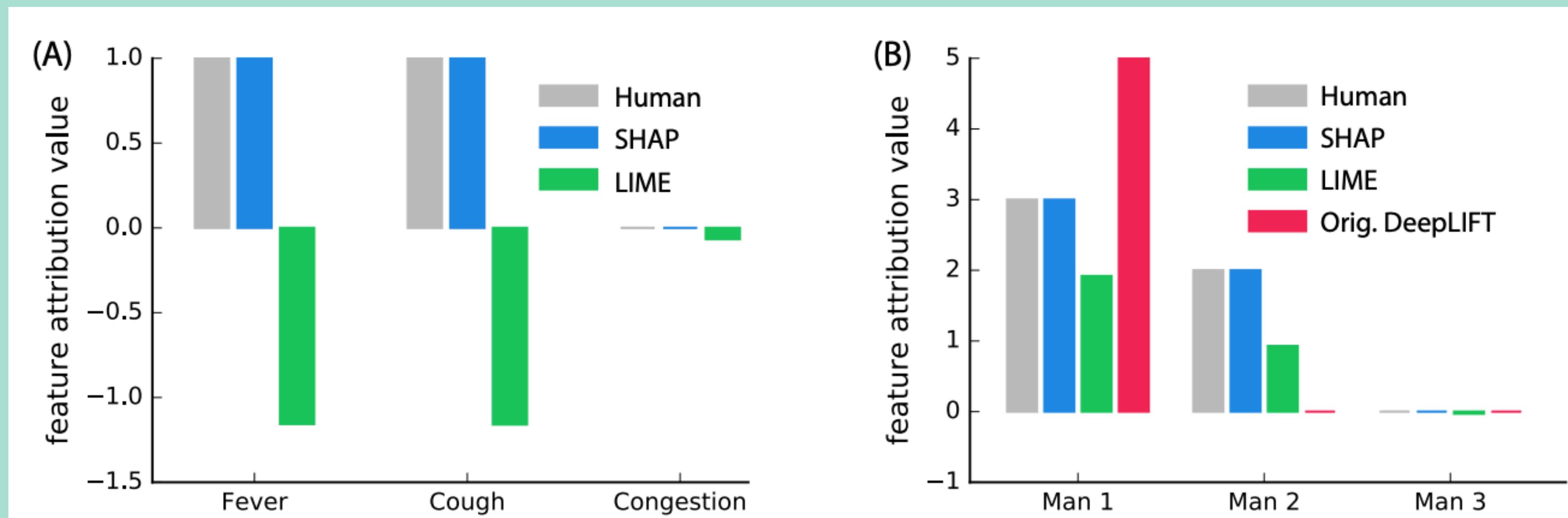
	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	96.6	94.5	96.2	96.7	96.6	91.8	96.1	95.6

- Selección y mejora de modelos mediante explicaciones



Aplicaciones de SHAP

- Aplicaciones en datos tabulares
- Interpretación de modelos complejos y redes neuronales profundas
- Consistencia con la intuición humana y confianza en el modelo



Conclusión

SHAP ofrece una explicación más consistente y matemática, permitiendo una comprensión tanto a nivel local como global de las predicciones, mientras que LIME se enfoca en explicar decisiones específicas de una forma un poco más rápida y sencilla.

Ambas técnicas ayudan a que aumente la confianza en los modelos y a tomar decisiones más informadas.

	LIME	SHAP
Nivel de explicación	Local únicamente	Local y global
Modelo-agnóstico	Sí	Sí
Costo computacional	Bajo a moderado	Moderado alto
Simplicidad	Alta	Moderada



Referencias

- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. <https://arxiv.org/abs/1702.08608>
- Gunning, D. (2017). Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency (DARPA). <https://www.darpa.mil/program/explainable-artificial-intelligence>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. <https://doi.org/10.48550/arXiv.1705.07874>
- MarkovML. (s. f.). LIME vs SHAP: Model explainability techniques compared. <https://www.markovml.com/blog/lime-vs-shap>
- Qu4nt. (s. f.). Nombres de las variables del dataset de diabetes (scikit-learn). https://qu4nt.github.io/sklearn-doc-es/auto_examples/feature_selection/plot_select_from_model_diabetes.html
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. En Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. <https://doi.org/10.1145/2939672.2939778>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. En Proceedings of the 32nd AAAI Conference on Artificial Intelligence (pp. 1527–1535). AAAI Press. <https://doi.org/10.1609/aaai.v32i1.11491>
- SHAP Contributors. (s. f.). Issue #750. GitHub. <https://github.com/shap/shap/issues/750>
- DataCamp. (s. f.). Introduction to SHAP values: Machine learning interpretability. <https://www.datacamp.com/es/tutorial/introduction-to-shap-values-machine-learning-interpretability>
- Demostraciones en Google Collab: https://colab.research.google.com/drive/1y_f1qcWOqPpS5S7GWiomFQuNus4G9TgQ?usp=sharing