Proyecto Tema 5

Machine Learning Interpretable

Noel Rodríguez Pérez

El Machine Learning Interpretable (IML), también conocido como Explainable Artificial Intelligence (XAI), se refiere a un conjunto de métodos y técnicas que permiten a los usuarios confiar y entender las decisiones y resultados de los modelos de aprendizaje automático que dichos métodos utilizan. Esta interpretabilidad es especialmente crucial en campos sensibles como la salud y la biomedicina, donde las decisiones pueden tener implicaciones significativas para la vida de las personas; ya sea a la hora de detectar o diagnosticar enfermedades en pacientes, como en la investigación de nuevas patologías y enfermedades.

Impacto en el ámbito de la biomedicina y la salud:

En el sector salud, la adopción de modelos de machine learning puede mejorar diagnósticos, pronósticos y tratamientos personalizados. Sin embargo, la complejidad de muchos de estos modelos puede generar desconfianza entre los profesionales médicos. La capacidad de interpretar el cómo y el por qué un modelo llega a una determinada conclusión o decisión, es esencial para su integración efectiva en la práctica clínica de cara a que el personal médico encargado sea capaz de fiarse de una forma mayoritariamente determinista de dicha elección.

Un estudio realizado por investigadores de diversas universidades en E.E.U.U. exploró las percepciones de médicos sobre el uso de modelos de aprendizaje automático para la detección de sepsis y posibles detecciones de diversas patologías no triviales de ver. A pesar de no tener una comprensión profunda del aprendizaje automático, los médicos desarrollaron confianza en estos sistemas a través de la experiencia práctica; donde se les recomendó además que, para una mejor praxis del ejercicio médico, sería conveniente que se especializasen un poco en dicho campo. Este análisis cualitativo identificó que, aunque los médicos reconocen el valor potencial de la IA, la falta de comprensión detallada sobre su funcionamiento puede ser una barrera para su adopción plena en entornos clínicos. Por lo que, sería necesario ampliar en la formación de los profesionales de la salud, introducir conocimientos necesarios para poder tener la capacidad de corregir cuando sea necesario, las decisiones tomadas por IA.

Además, la falta de transparencia en los sistemas de IA puede limitar su adopción en la práctica clínica, siendo necesario tener en cuenta los algoritmos que cada sistema de predicción implementa en cada caso y decidir cuál es el óptimo en cada ocasión. La confianza en que los sistemas de IA pueden ser confiables se ve afectada por la falta de explicaciones claras sobre cómo llegan a sus decisiones y la manera que tiene de colectar la información a comparar. Un artículo destaca que la falta de transparencia es una de las principales barreras para la implementación de IA en la atención médica, subrayando la necesidad de sistemas de IA explicables para su aceptación y uso efectivo

por parte de los profesionales de la salud. Este es uno de los principales motivos por los que todavía, muchos médicos no implementan dicha praxis en su día a día, pudiendo tener como solución a esto es una pequeña formación sobre la aplicación de dichos métodos a su vida laboral, donde se les explique los distintos algoritmos que son aplicables y para qué deben ser aplicados, acompañados de sistemas IA más visibles para personas no expertas en este ámbito del ML.

La integración de la IA en la atención médica presenta numerosos beneficios, como diagnósticos más rápidos y precisos, planes de tratamiento personalizados y reducción de costos, optimización a la hora de la elección de recursos a emplear y distintos puntos de vista a adoptar en ciertas situaciones. Sin embargo, superar los desafíos relacionados con la interpretabilidad y la confianza en estos sistemas es crucial para su adopción y uso efectivo en la práctica clínica. Quedando por encima de los beneficios que aportarían, la forma en la que los profesionales pueden ver si se ajusta a las necesidades dichas y no comete apenas errores a la hora de la toma de decisiones.

Tipos de modelos y técnicas de interpretabilidad:

Debido a los problemas de interpretabilidad que los sistemas IA generan en el ámbito la salud, surgió la idea de la creación de modelos de ML donde sea más fácil la manera de ver cómo se adoptan las decisiones tomadas y ver si es necesario aplicar un método u otro.

Los modelos de aprendizaje automático se clasifican en dos categorías principales según su interpretabilidad:

• **Modelos intrínsecamente interpretables (White-box models)**: Estos modelos son transparentes por naturaleza, siendo visible el método que se está utilizando y como se están interpretando dichos datos. Por ejemplo, podemos incluir la regresión logística y los árboles de decisión, donde las relaciones entre las variables independientes y la variable dependiente son claras y directas. Y los métodos de evaluación empleados, se aplican directamente a los datos de las muestras a comparar.

• **Modelos no intrínsecamente interpretables (Black-box models)**: Modelos como Random Forest, XGBoost y redes neuronales profundas, ofrecen alta precisión, pero carecen de transparencia en sus procesos de toma de decisiones debido a la alta complejidad de la metodología en la que se interpretan los datos. Para poder entender cómo se desarrollan las decisiones que se toman con dichos métodos, es necesario tener un conocimiento asentados sobre bases del ML.

Para abordar esta falta de transparencia, se han desarrollado técnicas de interpretabilidad que se dividen en:

• **Específicas del modelo vs. Agnósticas al modelo**: Las técnicas específicas del modelo están diseñadas para ser utilizadas con tipos particulares de modelos, donde se utilizan técnicas características de dicho modelo a interpretar. Mientras que las técnicas agnósticas o post hoc pueden aplicarse a cualquier modelo, independientemente de su arquitectura y metodología de actuación con los datos a tratar, aprovechando rasgos

generales técnicos que se emplean en todos los tipos de ML. Por ejemplo, los valores de Shapley se consideran técnicas agnósticas al modelo; mientras que los caminos de decisiones que se forman desde la raíz a las hojas en el método de Árboles de decisión sería una técnica específica del modelo.

• **Locales vs. Globales**: Las técnicas locales explican decisiones individuales, proporcionando información sobre predicciones únicas específicas, sin tener en cuenta el resto de las decisiones del modelo. Por el contrario, las técnicas globales ofrecen una visión general del comportamiento del modelo en su conjunto, viendo cómo se llega a la decisión de cada predicción y comparando todos los resultados en conjunto.

Ejemplos de métodos de interpretabilidad:

• **Modelos específicos**:

Métodos diseñados para trabajar con tipos particulares de modelos y explotan sus estructuras internas para facilitar la interpretación.

Globales: Análisis de la importancia de variables en árboles de decisión, ya que dividen los datos en función de características específicas para tomar decisiones. Analizar la importancia de las variables en estos árboles permite identificar qué características son más relevantes en el proceso de toma de decisiones del modelo. Por otro lado, la evaluación de los pesos de variables en regresiones logísticas indica la influencia que tienen sobre dicha predicción del resultado. Evaluar estos pesos ayuda a entender cómo cada característica contribuye al resultado final del modelo.

• **Modelos agnósticos**:

Pueden aplicarse a cualquier modelo, independientemente de su arquitectura, y se dividen en técnicas globales y locales.

Globales: Gráficos de dependencia parcial (PDP), los cuales muestran la relación entre una o dos características y la predicción del modelo, manteniendo constantes las demás variables. Ayudan a visualizar cómo varía la predicción a medida que cambia una característica específica. Por otro lado, Individual Conditional Expectation (ICE), son similares a los PDP, pero en lugar de mostrar una tendencia promedio, muestran cómo cambia la predicción para cada instancia individual a medida que varía una característica. Esto permite identificar heterogeneidades en las respuestas del modelo.

Locales: SHAP (SHapley Additive exPlanations), es una técnica que utiliza la teoría de juegos para asignar a cada característica una contribución justa en la predicción de un modelo. Calcula los valores de Shapley, que distribuyen equitativamente la "recompensa" (en este caso, la predicción) entre las características. Esto permite entender el impacto de cada variable en una predicción específica y en el comportamiento general del modelo. Otro ejemplo sería LIME (Local Interpretable Model-agnostic Explanations), donde se aproxima localmente el comportamiento de un modelo complejo mediante un modelo interpretable sencillo (como una regresión lineal o un árbol de decisión) alrededor de una predicción específica. Modifica ligeramente los

datos de entrada y observa cómo cambian las predicciones, lo que ayuda a entender las decisiones del modelo en casos individuales.

Método SHAP:

SHAP es una técnica que utiliza la teoría de juegos para asignar a cada característica de entrada una contribución justa en la predicción de un modelo. Calcula los valores de Shapley, que distribuyen de forma justa la "recompensa" (en este caso, la predicción) entre las características. Esto permite entender el impacto de cada variable en una predicción específica y en el comportamiento general del modelo.

- **Fundamentos Teóricos:**

Los valores de Shapley provienen de la teoría de juegos cooperativos, donde se busca distribuir una recompensa total entre los participantes en función de su contribución individual. En el contexto de aprendizaje automático, cada característica del modelo se considera un "jugador" que contribuye al "juego" (la predicción). El valor de Shapley mide la contribución promedio de cada característica, considerando todas las posibles combinaciones de características y su impacto en la predicción final.

- **Cálculo de los Valores de Shapley:**

Calcular los valores de Shapley implica evaluar todas las posibles combinaciones de características y determinar cómo la inclusión de cada característica afecta la predicción. Este proceso puede ser computacionalmente intensivo debido al número exponencial de combinaciones posibles. Para abordar este desafío, se han desarrollado métodos de aproximación eficientes, como KernelSHAP y TreeSHAP, que reducen significativamente el tiempo de cálculo manteniendo la precisión de las explicaciones.

- **Aplicaciones de SHAP:**

  - **Interpretación de Modelos:** SHAP proporciona explicaciones claras sobre cómo cada característica influye en las predicciones, facilitando la comprensión de modelos complejos.
  - **Detección de Sesgos:** Al analizar las contribuciones de las características, se pueden identificar y mitigar posibles sesgos en el modelo, promoviendo decisiones más justas y éticas.
  - **Validación y Confianza:** Ofrece transparencia en las decisiones del modelo, aumentando la confianza de los usuarios y stakeholders en su funcionamiento y resultados.

- **Implementación Práctica:**

La biblioteca SHAP en Python facilita la integración de esta técnica en proyectos de aprendizaje automático ML. Proporciona herramientas para calcular y visualizar los valores de Shapley, permitiendo a los usuarios interpretar y validar modelos de manera efectiva. Por ejemplo, se pueden generar gráficos que muestren la influencia de cada característica en las predicciones, identificando variables clave y posibles áreas de mejora en el modelo.

Bibliografía:

• Molnar, C. (s.f.). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.

• Hall, P., & Gill, N. (2019). *An Introduction to Machine Learning Interpretability: An Applied Perspective on Fairness, Accountability, Transparency, and Explainable AI*.

• Biecek, P., & Burzykowski, T. (s.f.). *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*.

• O'Sullivan, C. (2022). *What is Explainable AI (XAI)? An introduction to XAI — the field aimed at making machine learning models understandable to humans*.

• Lu, S.-C., et al. (2021). *On the importance of interpretable machine learning predictions to inform clinical decision making in oncology*. *Frontiers in Artificial Intelligence*, 4, 10013157.

• Byrd, R. (2019). *Machines Treating Patients? It's Already Happening*. *Time*.

• Salih, A., et al. (2023). *A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME*.

• Zilker, S., et al. (2024). *A machine learning framework for interpretable predictions in patient pathways: The case of predicting ICU admission for patients with symptoms