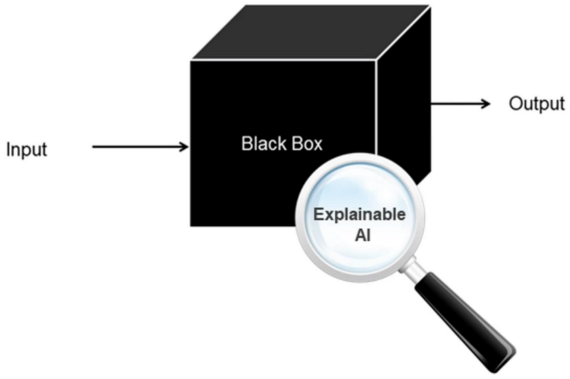
martes, 13 de junio de 2023 11:44

Una vez que tenemos desarrollado un modelo de ML, dependiendo de la aplicación puede resultar de interés conocer por qué el modelo se comporta de la forma en que lo hace. Esto es, conocer qué variables están teniendo un mayor efecto sobre la salida, cuales tienen un efecto positivo y cuales uno negativo, etc.

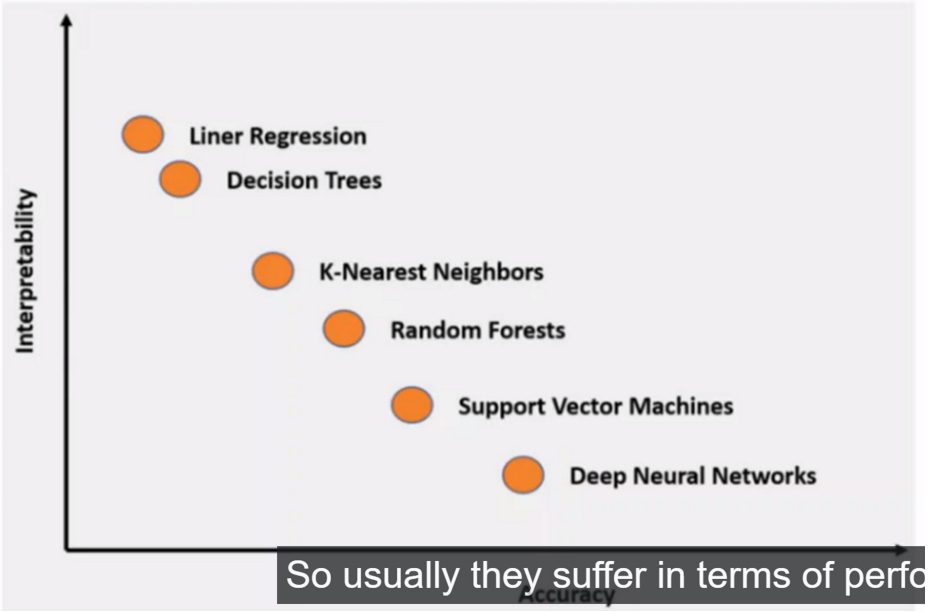
Esta es una tarea necesaria en la última etapa de validación de nuestro modelo, antes del despliegue.

Los modelos matemáticos cuyo comportamiento no es interpretable se conocen como modelos de caja negra. En algunas aplicaciones, como modelos de scoring crediticio, modelos que predicen culpabilidad en juicios, etc. Los modelos son auditados y deben ser explicables, ya que ciertas features relacionadas al genero, edad, etnicidad, etc. No pueden ser tenidas en cuenta ya que no sería ético.



Relación de compromiso: explicabilidad - precisión

En líneas generales podemos identificar que a medida que una mayor complejidad de nuestro modelo nos permite obtener mejores métricas de evaluación, también disminuye la interpretrabilidad del comportamiento.



Entonces tenemos una relación de compromiso entre explicabilidad y precisión.

¿Qué es explicabilidad en AI (XAI)?

Es un conjunto de herramientas y métodos que nos permiten interpretar de una forma comprensible para humanos, las predicciones realizadas por modelos de machine learning.

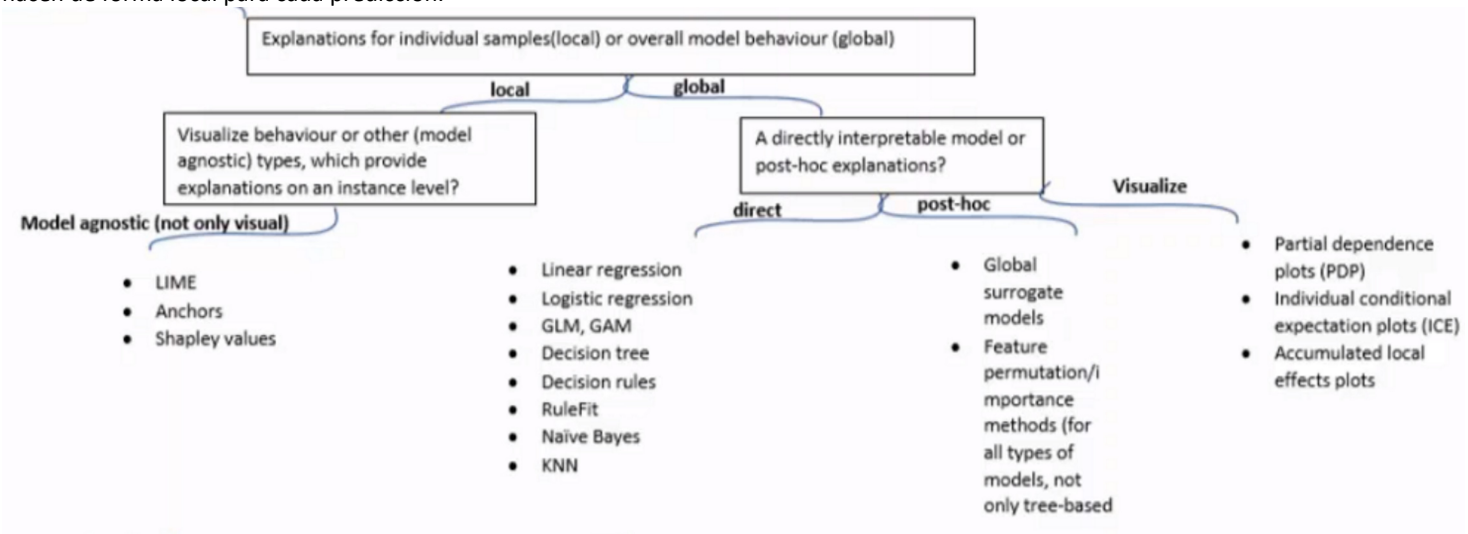
Tipos de métodos de explicabilidad

Existen distintos tipos de enfoques que podemos aplicar para aportar mayor explicabilidad a nuestros modelos, estos se pueden agrupar teniendo en cuenta lo siguiente:

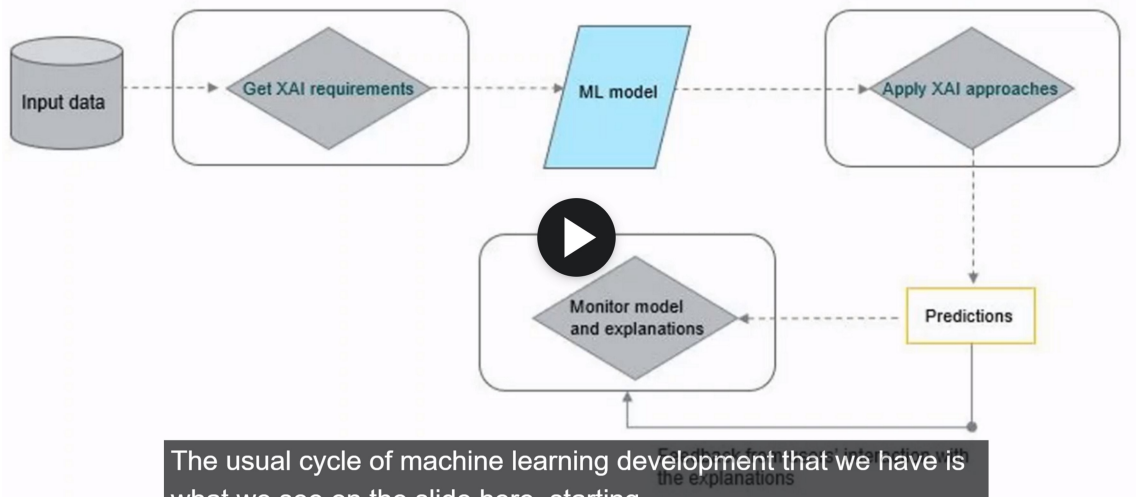
- Métodos previos al entrenamiento del modelo: son criterios que podemos aplicar antes del entrenamiento como restringir la cantidad de features, eliminar ciertas features para disminuir la complejidad, etc.
- Métodos posteriores al entrenamiento del modelo: son métodos que se aplican evaluando las predicciones brindadas por el modelo.

Métodos específicos para ciertos tipos de modelos y otro agnósticos, es decir que pueden ser utilizados independientemente de la arquitectura del modelo.

Métodos que brindan una explicación del comportamiento global del modelo, mientras que otros lo hacen de forma local para cada predicción.



A veces, puede ser conveniente definir los requerimientos de explicabilidad antes de desarrollar el modelo de ML.



Explicaciones visuales

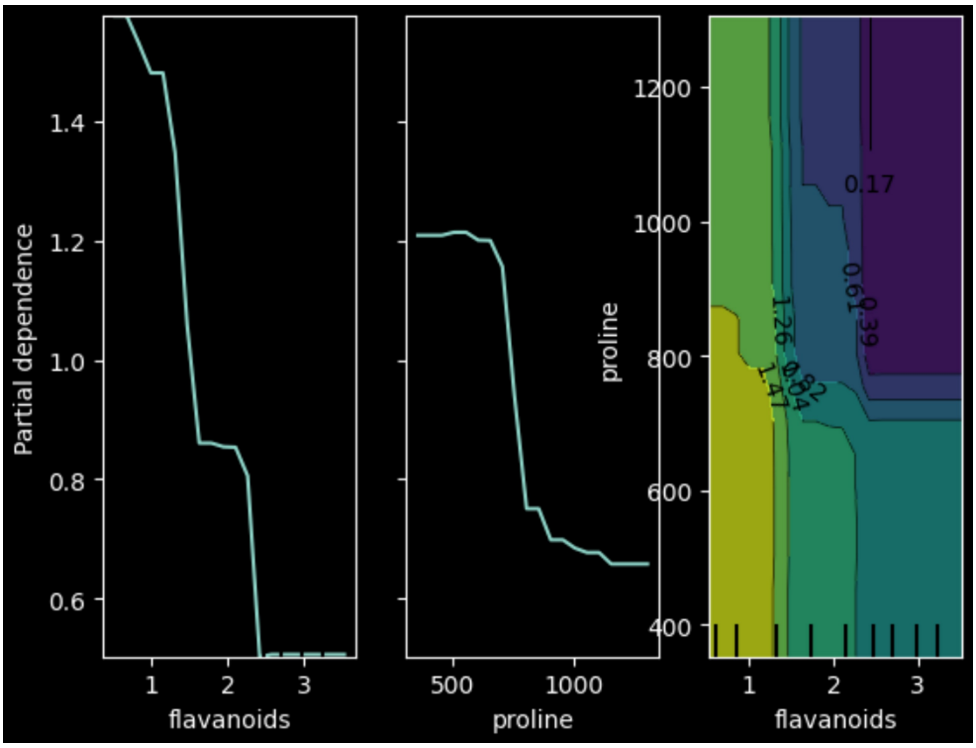
Como comentamos anteriormente, en caso de que debamos utilizar un modelo "transparente", podemos utilizar modelos como árboles de decisión, regresión lineal, regresión logística, RuleFit, etc. Que son modelos que por su naturaleza nos permiten realizar una lectura directa de como las features estan afectando el target predicho.

Partial dependency plots (PDP)

Es una manera gráfica de ver la contribución marginal promedio de cada feature sobre el target.

Es útil porque nos muestra si la relación que hay entre una determinada característica del modelo es lineal, monótona o más compleja.

Una de las suposiciones en las cuales se basa este método es que las características que no están siendo evaluadas en los PDP no están correlacionadas con las que sí se están evaluando.



Ventajas:
PDP son gráficos muy fácil de entender para stakeholders

Desventajas:
Su uso está limitado a dos features como máximo
La suposición de independencia es muy considerable
Solo muestra contribuciones marginales, por lo que efectos compuestos no son visibles.

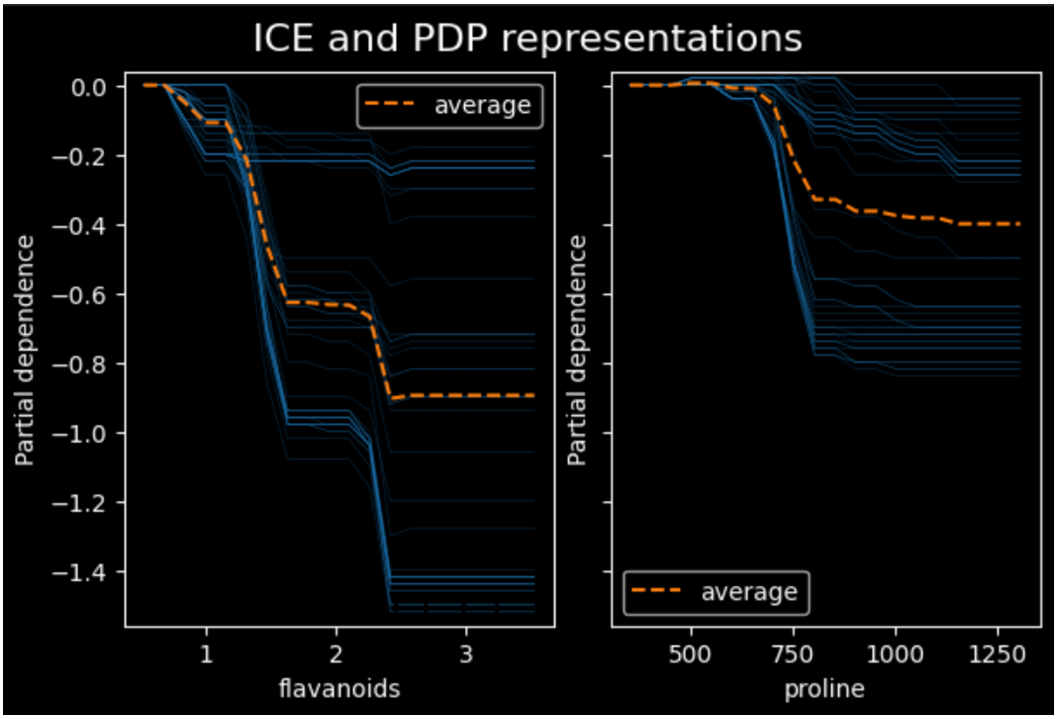
Individual conditional expectation (ICE)

Son una forma de visualizar y comprender cómo el valor esperado de la variable de respuesta cambia con respecto a una variable específica, manteniendo las demás variables constantes. Es decir, para cada fila de nuestro dataset, se dejan el resto de las variables fijas y se varía el valor de la característica bajo estudio.

Cada línea de un gráfico ICE representa el cambio en el valor esperado de la variable de respuesta para una observación individual a medida que una variable predictora específica varía.

El PDP se obtiene a partir de promediar los resultados del ICE.

Al igual que en el PDP, se asume que la feature bajo estudio no se encuentra correlacionada con el resto de las variables del dataset.



Global explanations

Este enfoque hace referencia a encontrar una explicación del comportamiento GLOBAL del modelo. Existen distintas técnicas, una de ellas puede ser la conocida como Surrogate Models.

Surrogate models consisten en entrenar un modelo de caja negra que satisfaga las necesidades predictivas del problema. Una vez que el modelo predictor se encuentra entrenado, un modelo más "transparente" o más facilmente explicable, es entrenado sobre las predicciones del modelo de caja negra.
A partir de la interpretabilidad otorgada por este último modelo de menor complejidad, podemos obtener una visión global de por qué el modelo de caja negra se comporta de cierta manera.

Para validar que los resultados obtenidos sean correctos, debemos chequear que las métricas obtenidas por el segundo modelo sean adecuadas.

Ventajas
Es un método fácil de implementar, flexible y es muy sencillo de interpretar y explicar

Desventajas
La explicabilidad que obtenemos de este método, está sujeta a que tanto se ajustan las predicciones del segundo modelo a las originales. Es un método que nos da una idea global aproximada.
No siempre es claro que tan buenas tienen que ser las métricas que evalúan el modelo surrogado para poder adoptar los resultados obtenidos.

Feature importance

Feature importance es un método que comunmente es propio de los modelos basados en árboles, pero que también puede ser calculado para otros tipos de arquitecturas con alguna librería externa.

El concepto de feature importance es conocer, cuanto varia una determinada métrica del modelo al cambiar el valor de una de las features de entrada. Este enfoque es conocido como Permutación y consiste en variar aleatoriamente el valor de una feature y ver el impacto sobre la métrica de evaluación. A medida que el impacto es mas grande, mayor es la importancia de esa feature.

Local explanations

Este enfoque se basa en explicar las predicciones de manera individual en vez del comportamiento

general del modelo como se vio en el caso de Global explanations.

Uno de los métodos que utilizan este enfoque se llama: Local interpretable model-agnostic explanations (LIME)

Es un algoritmo que utiliza permutación para genera un nuevo dataset

1. Selección de la instancia: Se elige la instancia de datos para la cual se desea explicar la predicción del modelo.
2. Generación de muestras vecinas: Se generan muestras ligeramente modificadas alrededor de la instancia seleccionada. Estas modificaciones se realizan mediante perturbaciones controladas, como eliminar o cambiar valores en las características de la instancia original.
3. Predicciones de las muestras vecinas: Las muestras vecinas se pasan al modelo de aprendizaje automático y se obtienen las predicciones correspondientes.
4. Construcción del modelo interpretable: Se ajusta un modelo interpretable (generalmente más simple) en las muestras vecinas generadas. Esto se hace para comprender cómo el modelo interpretable está tomando decisiones basadas en características seleccionadas.
5. Cálculo de la importancia de características: Se calcula la importancia de las características en el modelo interpretable para la instancia seleccionada. Esto proporciona una indicación de qué características están influyendo más en la predicción del modelo original.
6. Generación de explicaciones: Se generan explicaciones basadas en la importancia de características calculada. Estas explicaciones pueden ser en forma de puntuaciones, resúmenes textuales, visualizaciones, etc.

SHAP

Los SHAPley values indican que tanto cada feature "empuja" el valor de la predicción por encima o por debajo del valor promedio del target en el dataset.

El procedimiento es el siguiente:
1. Se define un valor de base para el target, comunmente el promedio del dataset.
2. Se crean múltiples "coaliciones", que son escenarios donde se van variando los valores asociados a cada feature para ver cómo impactan en el target.
3. Se promedian las "importancias" calculadas en el ítem anterior para obtener un valor promedio para cada feature en esa realización en particular