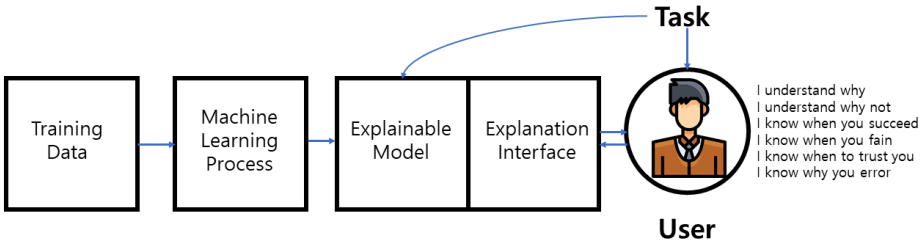
Explainable AI

What is Explainable AI?

Explainable AI (XAI) refers to a set of processes and methods that enable human users to understand and interpret the outcomes of machine learning,deep learning models. As AI systems are increasingly used in various domains, the need for transparency and interpretability has become critical. XAI seeks to make the decisions made by AI systems more comprehensible, ensuring trust and accountability.



Types of Explainability

1. **Global Explainability:** Focuses on providing insights into the overall behavior and decision-making processes of a model. It answers the question, "How does the model work?"
2. **Local Explainability:** Targets specific predictions or decisions made by the model, providing context on why a particular output was generated for a given input. It answers the question, "Why did the model make this specific decision?"

Techniques for Explainable AI

Several techniques have been developed to enhance the interpretability of machine/deep learning models. Here are some popular ones:

1. **LIME (Local Interpretable Model-agnostic Explanations)**

    -> Library: The `lime` library in Python.

2. SHAP (SHapley Additive exPlanations)

    -> Library: The `shap` library in Python.

3. VIT (Visual Interpretable Technique)

    -> Libraries: `matplotlib`, `seaborn` for visualizations.

Example: Abusive Comment Classifier

Imagine we have a classifier that determines whether a comment on a discussion forum is abusive or non-abusive. The classifier uses features such as the presence of specific words or phrases in the comment.

Scenario:

A user submits the comment: "You are a total loser and will never succeed."

How LIME Works:

1. **Original Prediction:**
   - The classifier predicts this comment as abusive.
2. **Using LIME:**
   - LIME takes this specific comment and generates similar comments by slightly altering it (e.g., changing some words).
   - It then predicts the sentiment for these altered comments.
3. **Building a Local Linear Model:**
   - LIME examines the altered comments and determines the contribution of specific words to the prediction:
     - "loser" (weight: -0.7): This word significantly contributes to the abusive classification.
     - "succeed" (weight: +0.3): This word has a positive connotation but does not outweigh the negative impact of "loser."
4. **Final Interpretation:**
   - The LIME output could be represented as:

     Prediction=−0.7×"loser"+0.3×"succeed"

   - This means that if the word "loser" were removed, the prediction would likely shift toward non-abusive due to the influence of the positive word "succeed."