Submitted by: Shashank Dammalapati, Benoit Atari

1. What is the meaning of explainable machine learning?

Explainability in the context of explainable machine learning is to combine interpretable properties with domain knowledge, in addition to an analysis goal.
Explainable ML models offer transparent and understandable explanations for their decisions while integrating domain-specific knowledge into the model to align it with human expertise in a given field for enhanced reliability.

2. Why do we need to understand the decision-making mechanism of a machine learning algorithm?

1. To ensure transparency and fairness
2. To debug biases in the ML algorithm (Deep Dream - Dog Face Bias)
3. To help improve the performance of the algorithm - Explainable ML can help identify areas where the model can improve.
4. In Self Driving Car scene, it's important to understand the reasoning behind the perception and planning algorithms which are designed as an ML problem

3. What are the differences between explainable models and interpretable models? (give two examples for each type.)

Explainable models blend interpretable features with domain expertise and analysis. They provide clear justifications for their decisions with domain-specific knowledge to improve their reliability and trust.

Examples:

1. Shallow Convolutional Neural Networks (CNNs),Random Forest
2. Activation Maximization Networks: Activation maximization networks focus on amplifying specific neurons' activations within a CNN to visualize the input patterns that trigger those activations. By visualizing the patterns that lead to high activations, users gain insights into the features the model has learned, enhancing its explainability.

Interpretable models, on the other hand, prioritize making their inner workings understandable without necessarily delving deeply into domain-specific knowledge, placing a higher emphasis on human-friendly insights.

Examples:

1. Linear Regression
2. Decision Trees

4. What are the differences between global explainability methods and local explainability methods? (give two examples for each type.)

Global Explainability methods focus on understanding the behavior of the model with respect to the **entire dataset** or a significant subset of it

Examples:

Feature Importance using permutation importance can be used to determine the importance of each feature in the model's predictions across the entire dataset.

Partial Dependence plots visualize the relationship between a selected feature and the predicted outcome while keeping all other features constant on average. They provide insights into how the model's predictions change as the chosen feature varies.

Local Explainability methods on the other hand focus on understanding the behavior of the model with respect to **a specific instance** of the dataset or a small instance of the dataset

Examples:

LIME (Local Interpretable Model-agnostic Explanations): LIME LIME helps us understand why a specific prediction was made by creating a simple, easy-to-understand model (like linear regression) using slightly tweaked versions of the original data point.

SHAP (Shapley Values) for Individual Predictions: While SHAP can be used for global explanations can also be zoomed in to explain why a particular prediction was made for a specific instance.

5.What are the differences between model-agnostic methods and model-specific methods? (give two examples for each type.)

Model-agnostic methods are techniques designed to provide explanations and insights into the predictions made by machine learning models, without relying on the specific details or inner workings of the model itself. These methods aim to offer transparency and interpretability for any type of machine learning model, regardless of its complexity or algorithm used for training.

Examples:

LIME (Local Interpretable Model-agnostic Explanations): LIME creates interpretable models such as linear regression models around specific instances to explain their predictions, regardless of the type of underlying model.

SHAP (Shapley Additive exPlanations):SHAP values provide a way to measure the impact of each feature on the model's predictions, and they can be applied to any machine learning mode

Model-specific methods are techniques designed to provide explanations and insights into the predictions made by a specific type of machine learning model. These methods are customized to leverage the unique characteristics, architecture, and internal workings of the chosen model.

Examples:

Decision Tree Visualization: Decision trees themselves are interpretable, and their structure can be directly visualized to understand how the model makes decisions.

Feature Importance in Random Forests:Random forests have a built-in method for calculating feature importance, which provides insights into which features are influential in making predictions.

6. What are the differences between Ad-hoc methods and post hoc methods?

Ad-hoc methods refer to techniques that are improvised or created on an as-needed basis, often without a formal or predefined structure.They are designed to address specific, immediate problems or challenges without necessarily following established methodologies.while **Post-hoc Methods** are techniques or analyses that are applied after an experiment, study, or process has been completed.They are used to gain insights, draw conclusions, or make inferences from data or observations that have already been collected.

.

7.What are the differences between data modality agnostic methods and data modality specific methods?

Data modality agnostic methods are techniques that are designed to work across different types of data modalities, such as images, text, audio, or numerical data.They are designed to handle a wide range of data types without being tailored to any specific modality. They do not rely on detailed knowledge about the specific characteristics of a particular data type. while **Data Modality Specific Methods** are techniques or algorithms that are designed specifically for handling a particular type of data modality. They are tailored to the unique characteristics and properties of that modality.