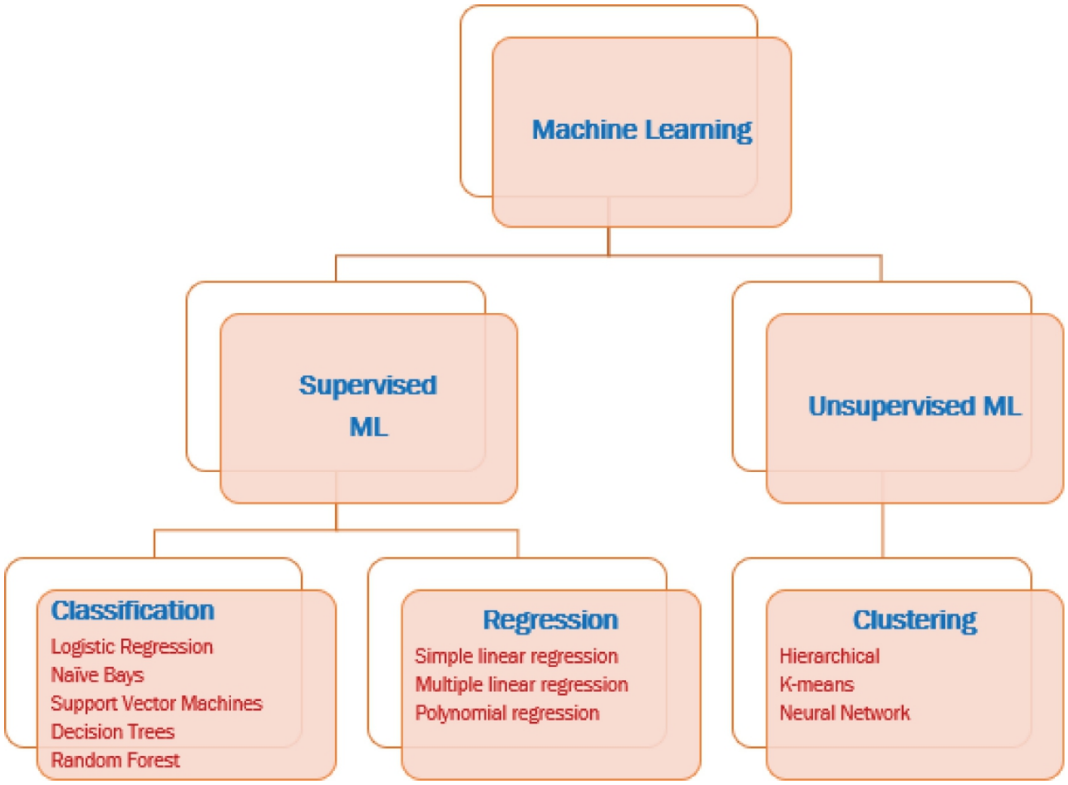


Figure 2.4: ML algorithm taxonomy.

2.3 XAI Evaluation

Mechanisms to evaluate the explanations provided by the XAI model are important. Let's consider the healthcare use case. There exist different users of the system such as the end users, decision makers such as the medical officers, nurses, etc., in a healthcare environment who make critical decisions related to the patient's medical treatment based upon the results generated by the system, data scientists who look to improve the model and regulatory agencies who verify the compliance to rules. Each of these users has different goals for the explanations as presented in Figure 2.5 and evaluation of the explanations should be tuned towards the users and their explanation goals, such as simply understanding the predictions and increasing the trust, model verification and model debugging.

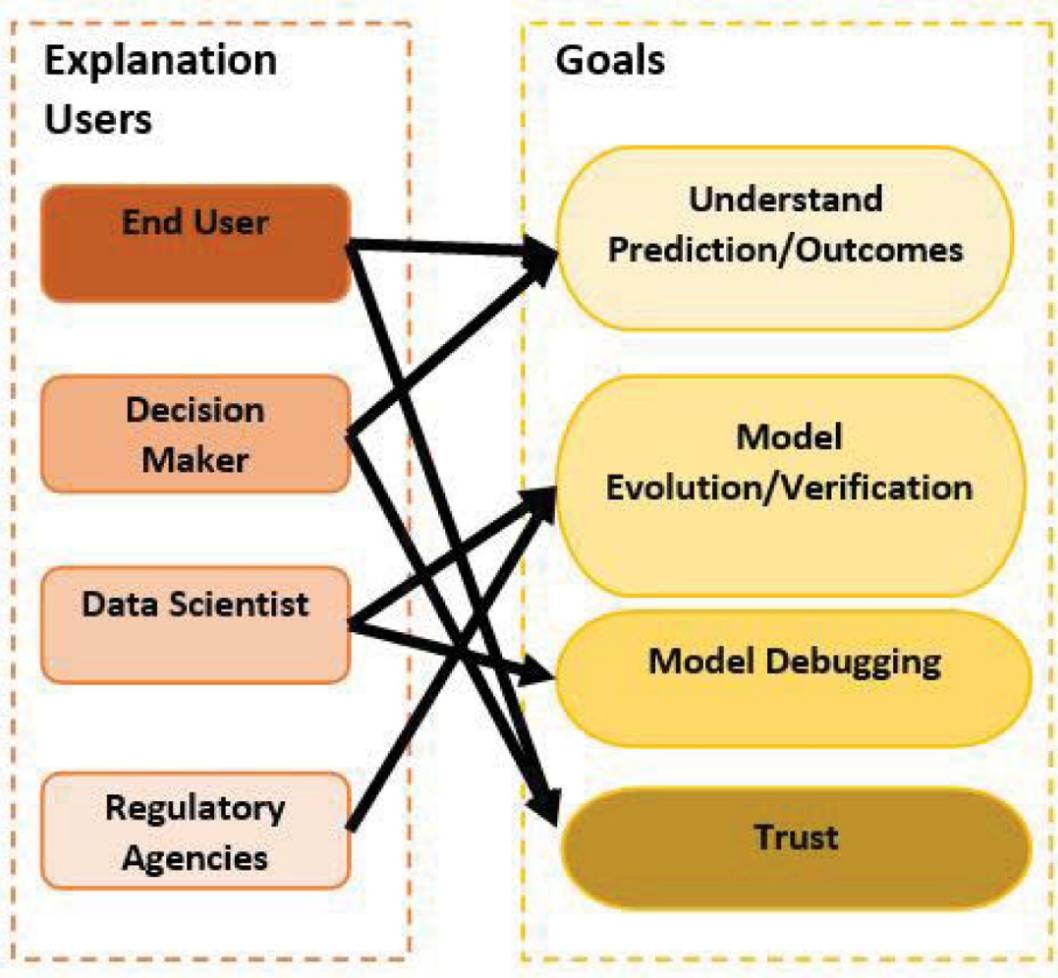
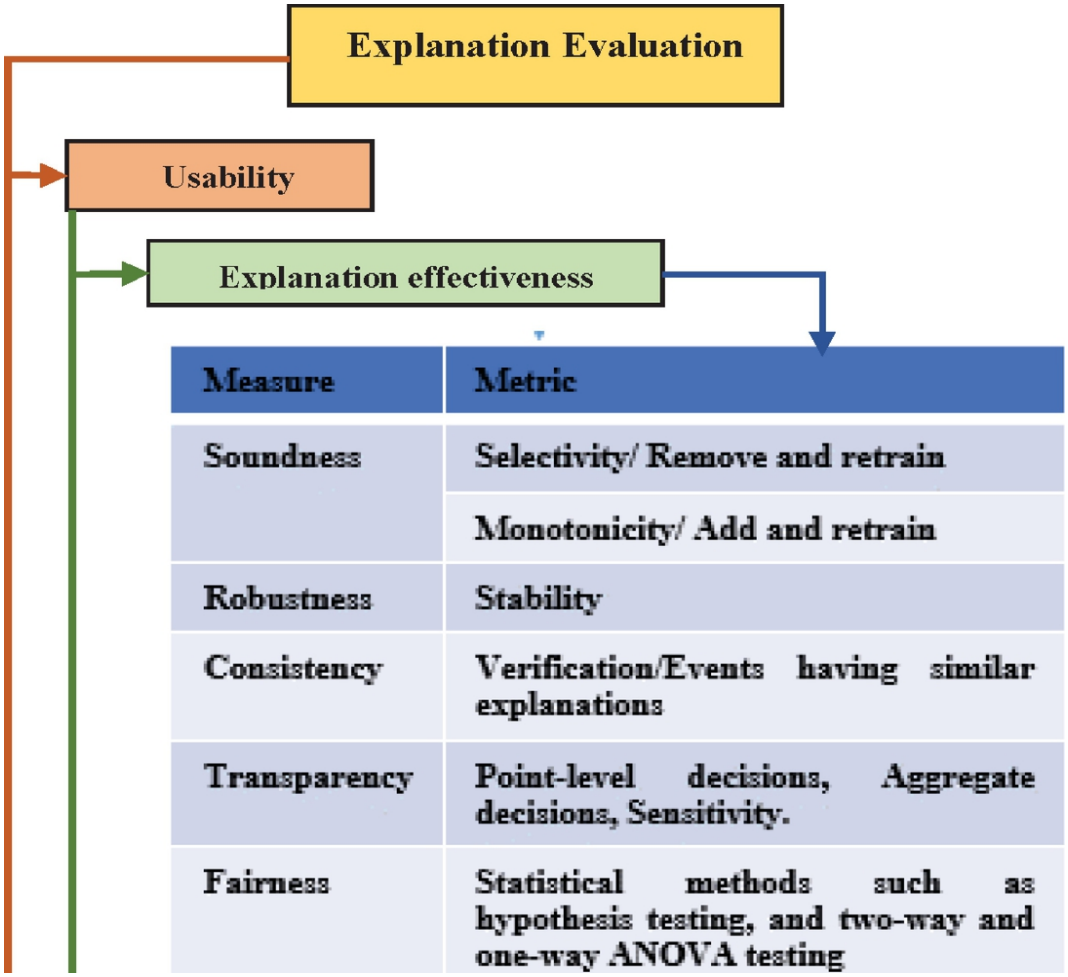


Figure 2.5: Explanation of users and associated goals.

Explanations can be presented in different forms such as Excel sheets, highlighting parts of images that contributed to the decision, heat maps, etc. In many cases, the explanations are not dependent on the choice of metrics or correctness, but it simply depends upon how well people really understand the explanations with ease.

In general, we can split the explanations into two categories based on usability and correctness. Usability does not consider the content of the explanations for evaluation while correctness considers the content of explanations and requires the existence of a ground truth for evaluations like the ground truth in a supervised learning algorithm. The usability aspects are

further classified depending upon the effectiveness of explanation and effectiveness of functionality to end users. Since the dataset against which the model is trained does not have explanations of truth that directly specifies the important features that guide the predictions/outcomes, a quantitative study/metric for evaluation of correctness does not exist. However, significant attempts have been done where the authors [12] have attempted to represent real-world classification tasks by coding the explanations into the dataset using formal grammar and have introduced an evaluation metric called k-accuracy for the same which is believed to act as an agreeable benchmark to compare feature importance AI methods. We propose a taxonomy of explanation evaluation measures and metrics as shown in Figure 2.6.



Measure	Metric
Soundness	Selectivity/ Remove and retrain
	Monotonicity/ Add and retrain
Robustness	Stability
Consistency	Verification/Events having similar explanations
Transparency	Point-level decisions, Aggregate decisions, Sensitivity.
Fairness	Statistical methods such as hypothesis testing, and two-way and one-way ANOVA testing

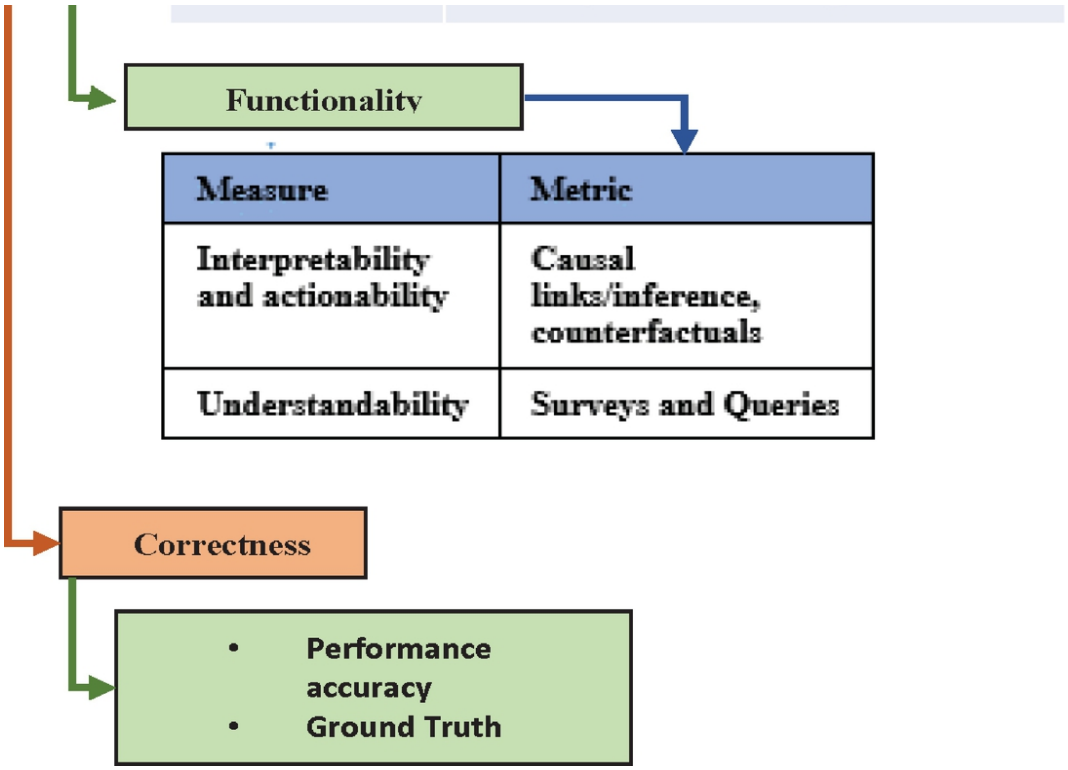


Figure 2.6: Explanation evaluation taxonomy.

The measure that evaluates the dependence of the model on a particular feature refers to soundness. This can be done by choosing a particular feature (for example a feature that has the highest importance value) to be removed and retraining the model to check the outcome (selectivity). The opposite can also be done by starting the training from just one feature and then adding features and retraining the model to see the outcome (monotonicity).

Robustness checks the stability of the model in the presence of undesirable noise features. The addition of noise features to the dataset from a different distribution should not affect the explanations. Transparency refers to the measure that helps verify that the model's decisions are made by considering the right reasons that contribute to the decision. For example, a wolf and a husky dog should be identified by their features such as the shape of ears, eyes, etc. and not by the presence of snow in the background. Metrics such as point-level decisions, aggregate decisions and sensitivity can be used to

evaluate the transparency of the model. Point-level decisions are local explanations that explain individual decisions and present the users with an alternate course of action that can be taken to change/improve the decision. The aggregate decision is a metric that enables the identification of the features that act as the overall drivers of the model. It helps to exactly locate the information and features that the model considers important and explain the reasons for the model's performance on a larger scale. Sensitivity is a metric that enables visualization of the feature's contribution to the decision-making process of the model. Influence sensitivity plots, partial dependence, and accumulated local effects plots are some tools that can be used for visualizing the sensitivity of the model. Interpretability refers to the accuracy with which the model associates the cause with its effect. In simple terms, if a model accepts inputs and consistently provides the same results then the model is interpretable.

Fairness refers to the ability of the model to provide predictions and classifications without any bias. Evaluations are done in terms of impartiality and discrimination. Statistical methods, such as hypothesis testing, ANOVA testing, impact testing, etc., can be used as metrics to evaluate the fairness of explanations.

2.4 XAI Mechanism Used for Prediction

In this section, we discuss the various explainable AI mechanisms used to explain the predictions of the detection algorithms, and a summary is presented in Table 2.1. In [13] the authors have evaluated two different feature sets, namely NetFlow and CICFlowMeter against a dataset and have found that the NetFlow feature enhances the detection accuracy of the tested models namely deep feed forward and random forest. SHAP has been adopted to explain the classification predictions of the models. However,