Explainable AI

* Explainable AI allows us humans to understand how a AI model comes up with its results and consequently build trust in those results. It takes predicated Output from model with data to generate explanations. The explanations are analysed with clinical knowledge .If predictions are correct it is used to generate insights. If predictions are incorrect then the model is improved. Explainable AI has an user centric approached.
* Explainable AI also help in troubleshooting and improving performance. It allows us to investigate model behaviours through tracking model insights on deployment status, fairness, quality and drift.
* LIME - Local interpretable model agnostic explanations

The LIME method interprets individual model predictions based on locally approximating the model around a given prediction

* + Work on any blackbox model
  + Model internals are hidden
  + Works with many data types
  + Using prior knowledge data types

Paper - “Why Should I Trust You?” Explaining the Predictions of Any Classifier

* SHAP - Shapley additive explanations

Shapley value tell us the average contribution of a feature in prediction.(each feature contribute differently in our to prediction)

Paper - A Unified Approach to Interpreting Model Predictions

Marginal value/contribution - calculating contribution of each feature in each subset and then simply averaging over all of these contributions

* Counterfactual explanations

A counterfactual is the smallest change in the input features, that changes the prediction to another (predefined) output.

Paper - COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR