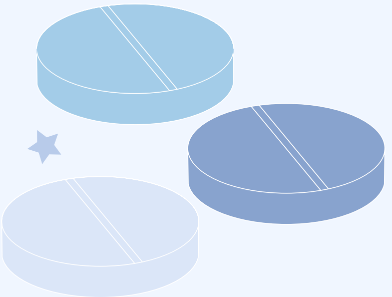
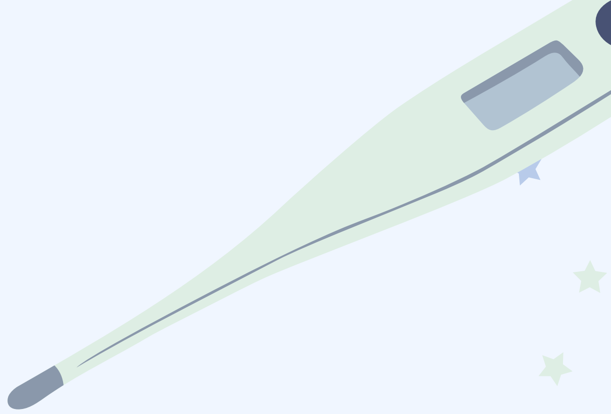




AI Heatlh Workshop Explainability

Muhammad Adil Saleem

25th April 2025

Agenda

Black box models

Understanding why opaque decision-making in models limits trust and usability.

Global vs Local

Differentiating between broad model behavior insights (global) and specific prediction insights (local).

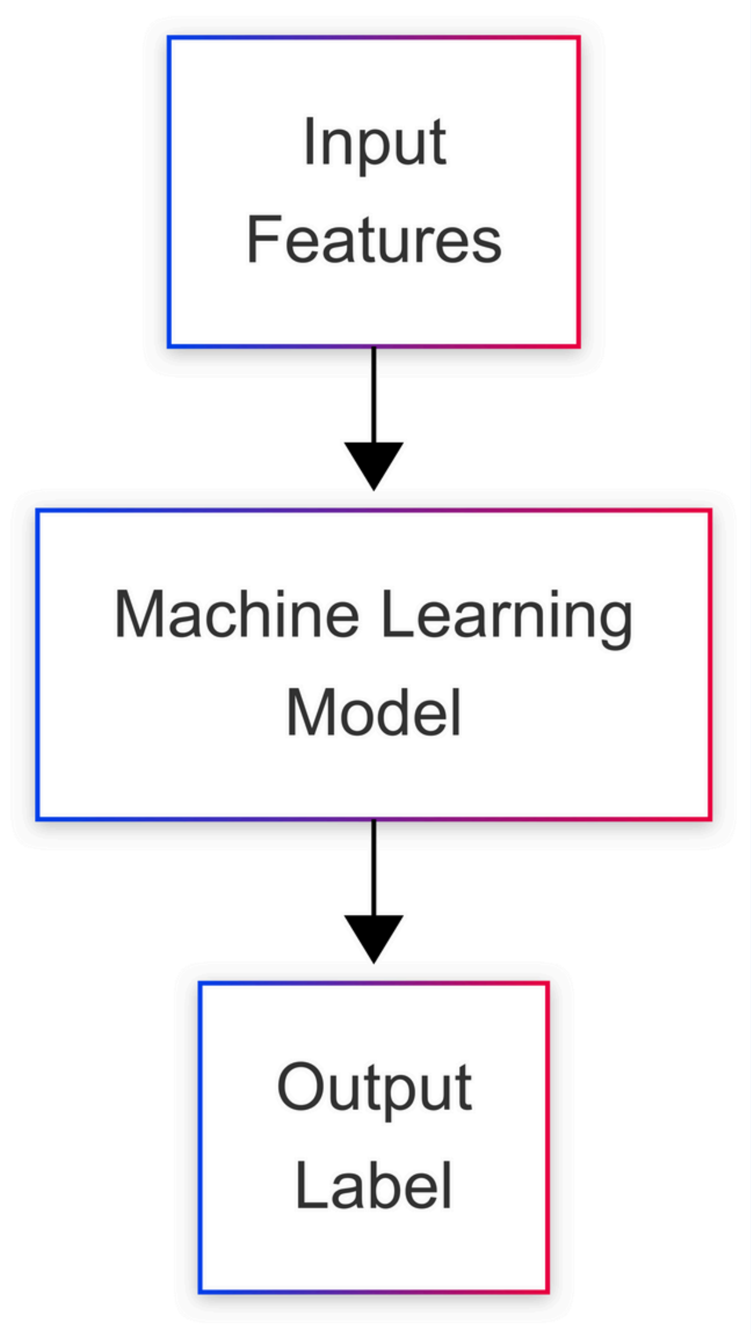
Interpretable models

Exploring models that are inherently transparent and easier to explain.

Post-hoc explanations

Discussing techniques to interpret and explain blackbox models after they've been trained.

How ML models work?



Input Features

Machine Learning Model

Output Label

Machine Learning (ML) models usually work by taking input (features) and output (label), learning patterns from the data during training, and then using those patterns to make predictions or decisions on new, unseen inputs.

Often, these model don't show how they reached their decision.

Explainability to the rescue

Explainability in machine learning refers to how a human can understand why a machine model made a certain prediction or decision.

Example

A machine learning model predicts that a patient is likely to develop diabetes. Using explainability, we find that the key reasons for this prediction are the patient's high glucose level, BMI, and number of pregnancies.

This explanation helps a doctor understand why the model gave that prediction, allowing them to verify it with medical knowledge.

Before we start

Internet

Ensure the internet is enabled in the notebook

Accelerator

Ensure the accelerator is turned on - this will speed up code execution where possible

Dataset

Ensure the dataset is load - we'll use samples from a world dataset

Follow the instructions provided in the notebook and let us know if you need any assistance

Global vs Local Explanations

Global Explanations

Global explanations provide an overview of the model’s behavior and which features generally influence predictions across the whole dataset.

Local Explanations

Local explanations dive into individual predictions, showing which features contributed most to a specific decision made by the model.

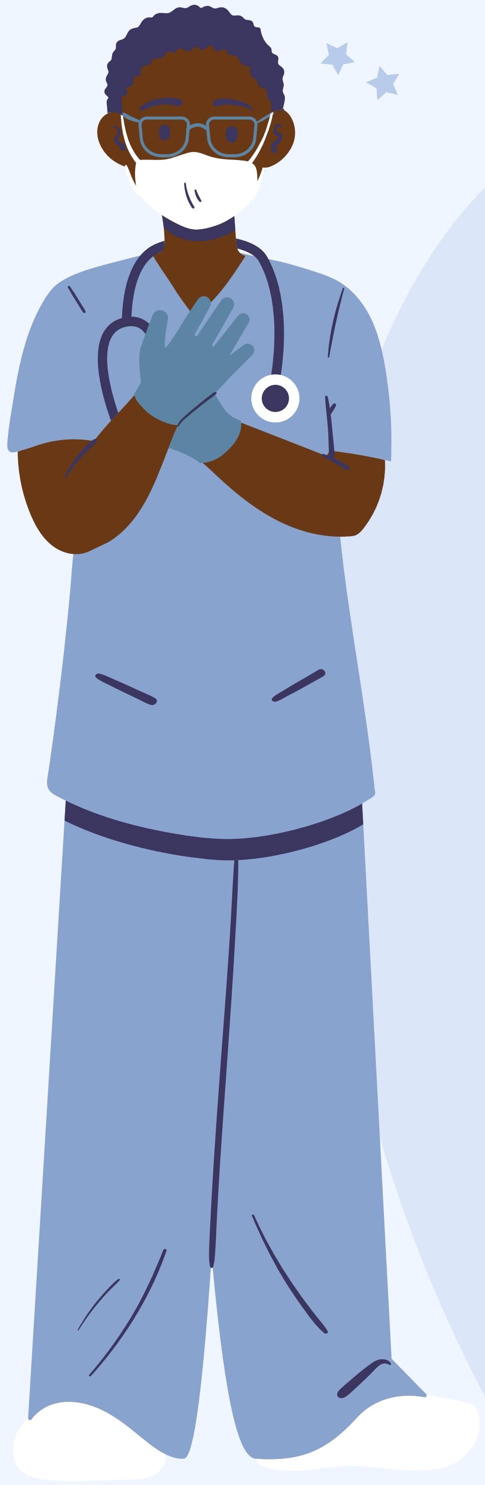
Global Explanation using Logistic Regression

Intuition

The magnitude of the coefficients (w) is proportional to how important a feature is. The sign also indicates whether it influences positively or negatively.

Logistic regression is like linear regression, but instead of predicting a number, it predicts a label. It still uses
z = w1*x1 + w2*x2 + ... + wn*xn +b
but a little differently than linear regression

Local explanation using Logistic Regression



Intuition

The magnitude of the coefficients (w) multiplied with the current input is proportional to how important a feature is in making a prediction for the patient under observation.

The features are usually standardized (e.g. between 1 and -1) so that the features are comparable.

Local explanation using Decision Trees

Intuition

Decision trees works by making branches based upon the features of the input data. When we test an unseen record, it takes a decision path which gives us explainability.

Sometimes, we are concerned about what features are important rather than how much each feature is important.

Interpretable models vs post-hoc explainability

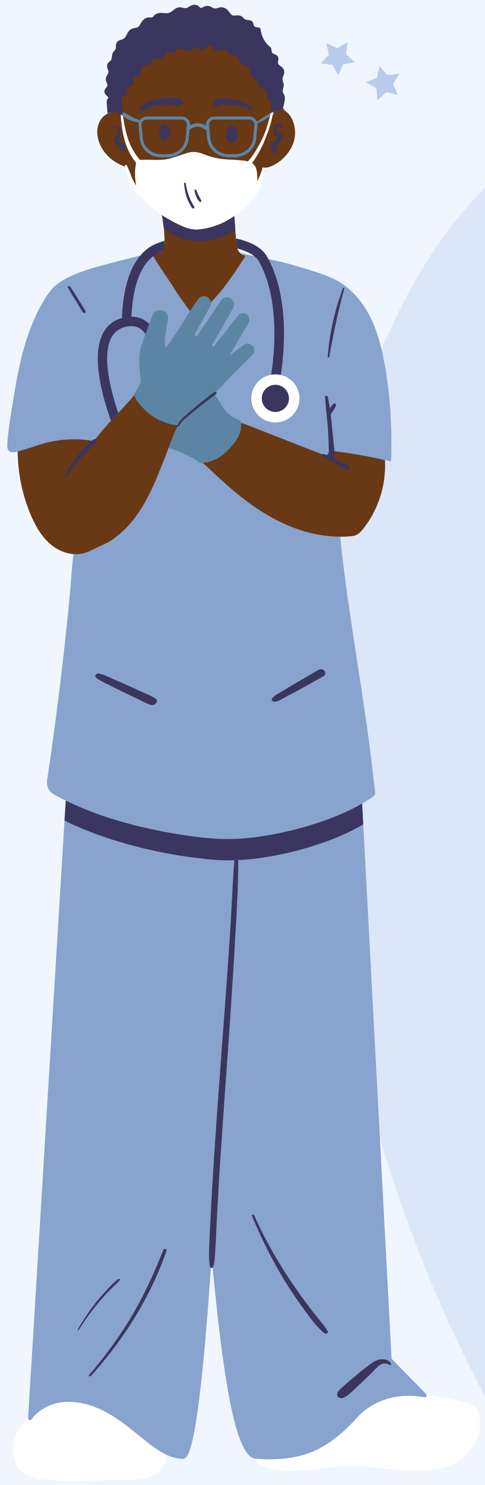
Interpretable models

Interpretable ML models are inherently explainable but they are generally weak.

Post-hoc explainability

More powerful ML models are not inherently explainable. They require, what we call, post-hoc explainability, to make them explainable.

Explainability with LIME



Intuition

LIME explains a prediction by creating many small variations of the input and observing how the model's output changes.

LIME or Local Interpretable Model Agnostic Explanations uses a surrogate model (an interpretable model) to make a black-box model explainable.

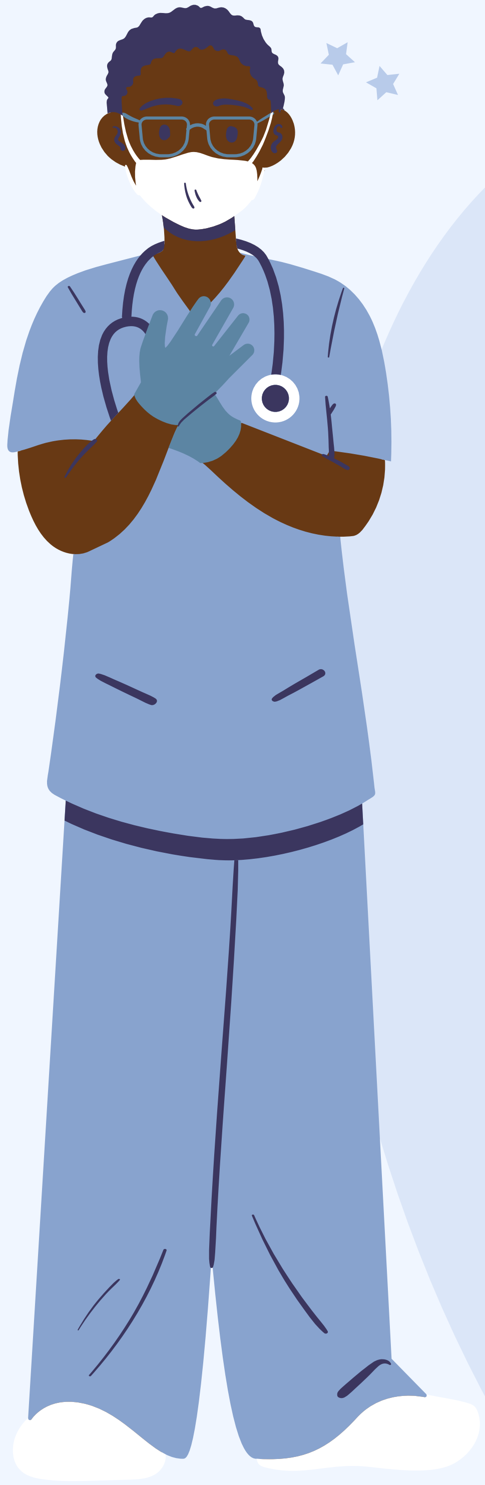
Explainability with SHAP

Intuition

SHAP explains a prediction by calculating the contribution of each feature using game theory, treating features like players in a game.

SHAP or SHapley Additive exPlanations can be computationally expensive and is often approximated using different techniques.

Explainability with Counterfactual Explanations



Intuition

Minimally changing the original input such that the output (label) changes also provides insights on what features are important.

Counterfactual Explanations are unique as they are actionable explanations i.e. it suggests change in original value to flip the output
e.g. diabetes -> no_diabetes

Pitfalls of explainability

Over-simplification

Surrogate models may not capture nuances of very complex models

Computational Cost

Some techniques are computationally prohibitive

Accuracy vs explainability

Interpretable models are not always best models for a task.

Conclusion

Use explainability

Explainability can foster trust of healthcare practitioners in AI based solutions

But choose the right one

There is no silver bullet for explainability, so choose the best one the fits your use case.

Thank you for your attention