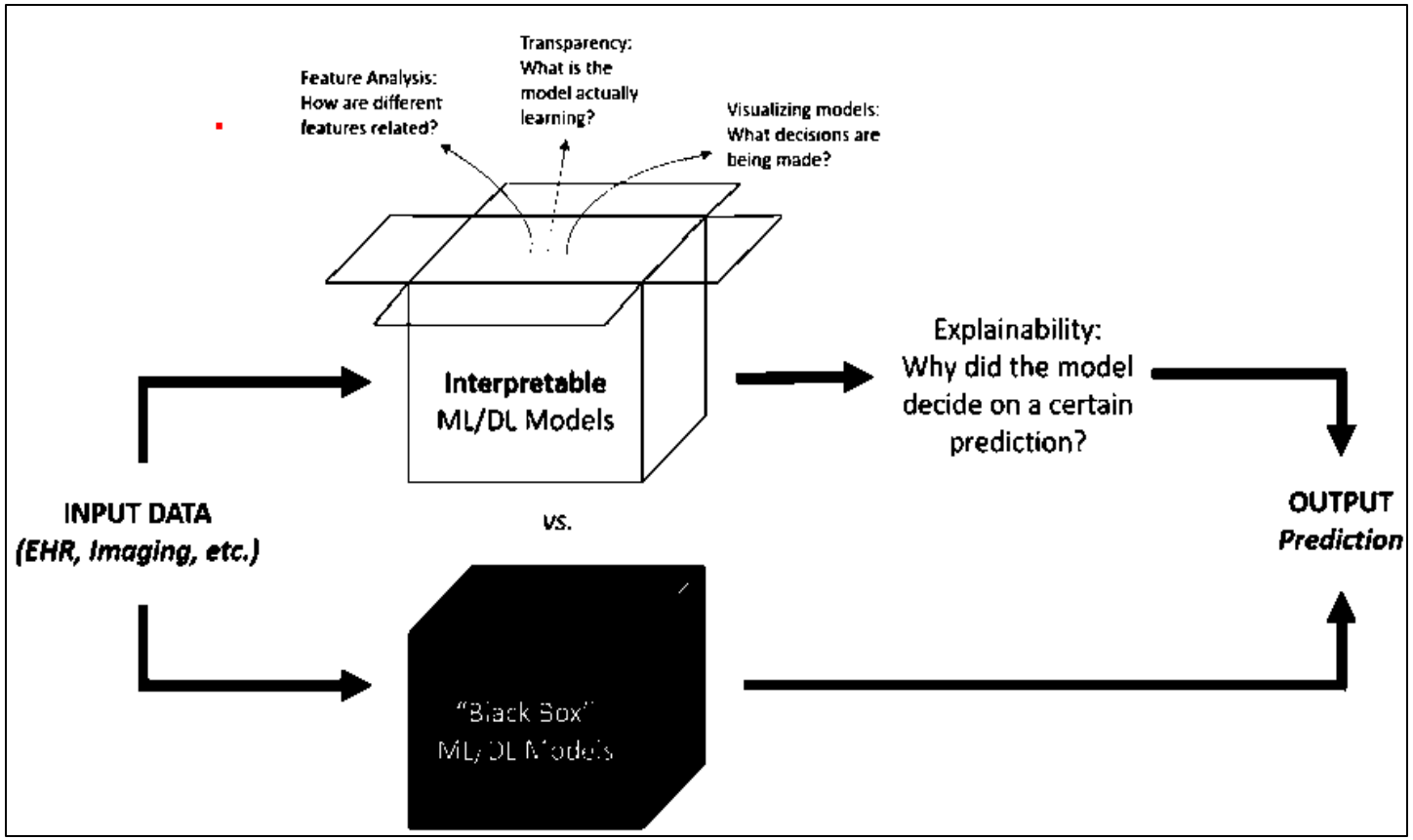
Explainable AI with Shapley Value

Presented By-

Vishnuja Chand- 22727
Madhuri N- 22662
Sriya R G- 22414
Sidhant S- 22768

GTMD 2024

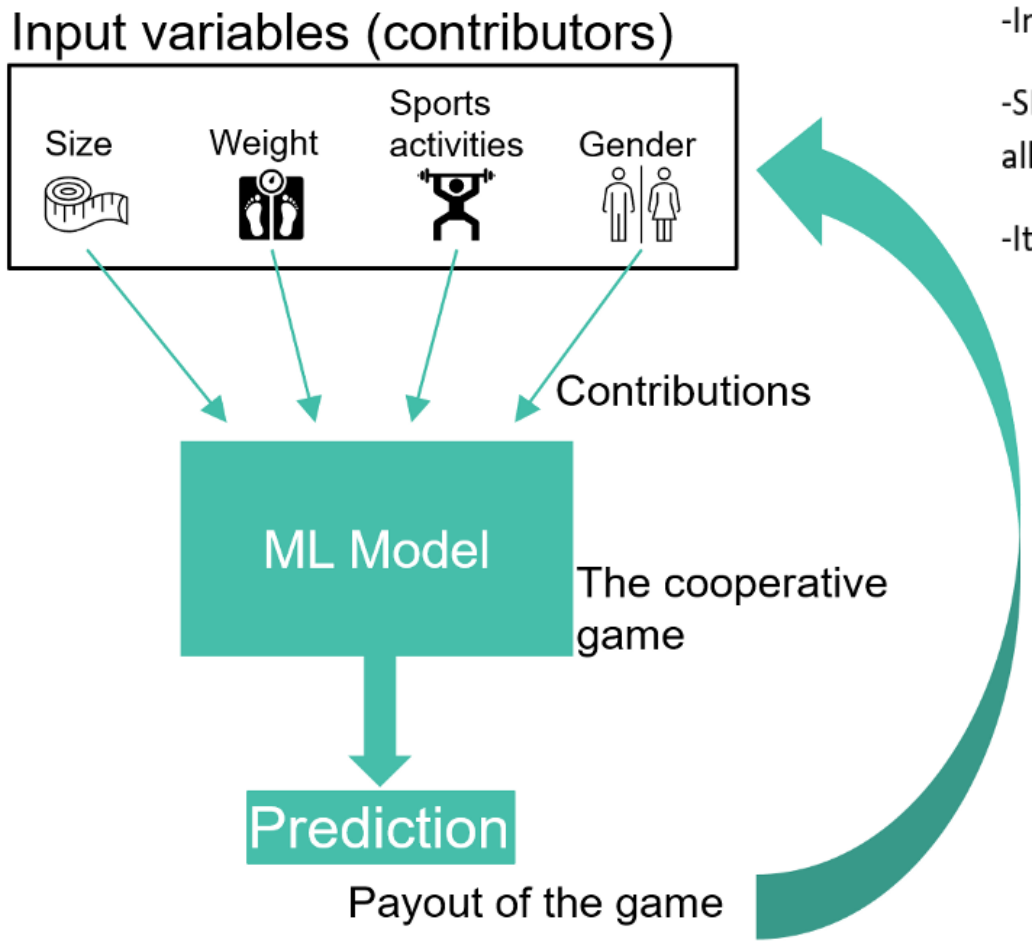
WHY XAI??



HOW

- Local Feature Attribution Methods
  - LIME
  - Focus on Local Interpretability
- Additive Feature Attribution Methods
  - SHAP
  - Focus more on global perspective
  - Useful in terms of model comparison

SHapley Additive exPlanations : From Cooperative game theory to XAI



How to distribute the payout fairly among the contributors?

-Instead of Player in a coalition game we have features in predictive models

-SHAP calculate contribution of each feature to model prediction for specific instance by considering all possible combination of features

-It evaluates how including each features changes prediction compared to excluding it

Traditional Shapley values from Cooperative game theory

$$\phi_v(i) = \sum_{\pi \in \Pi} \frac{1}{n!} \left(v\left(\{j : \pi(j) \leq \pi(i)\}\right) - v\left(\{j : \pi(j) < \pi(i)\}\right) \right)$$

Lloyd S Shapley, 1953

SHAP

$$\phi_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

AXIOMS

1) Efficiency $\sum_{i \in N} \phi_v(i) = v(N)$
2) Linearity $\phi_{\alpha u + \beta v} = \alpha \phi_u + \beta \phi_v$ for any value functions u, v, and any $\alpha, \beta \in \mathbb{R}$.
3) Nullity $\phi_v(i) = 0$ whenever $v(S \cup \{i\}) = v(S)$ for all $S \subseteq N \setminus \{i\}$.
4) Symmetry $\phi_v(i) = \phi_v(j)$ if $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq N \setminus \{i, j\}$.

Asymmetric Shapley Value (ASV)

$$\phi_v^{(w)}(i) = \sum_{\pi \in \Pi} w(\pi) \left[v(\{j : \pi(j) \leq \pi(i)\}) - v(\{j : \pi(j) < \pi(i)\}) \right]$$

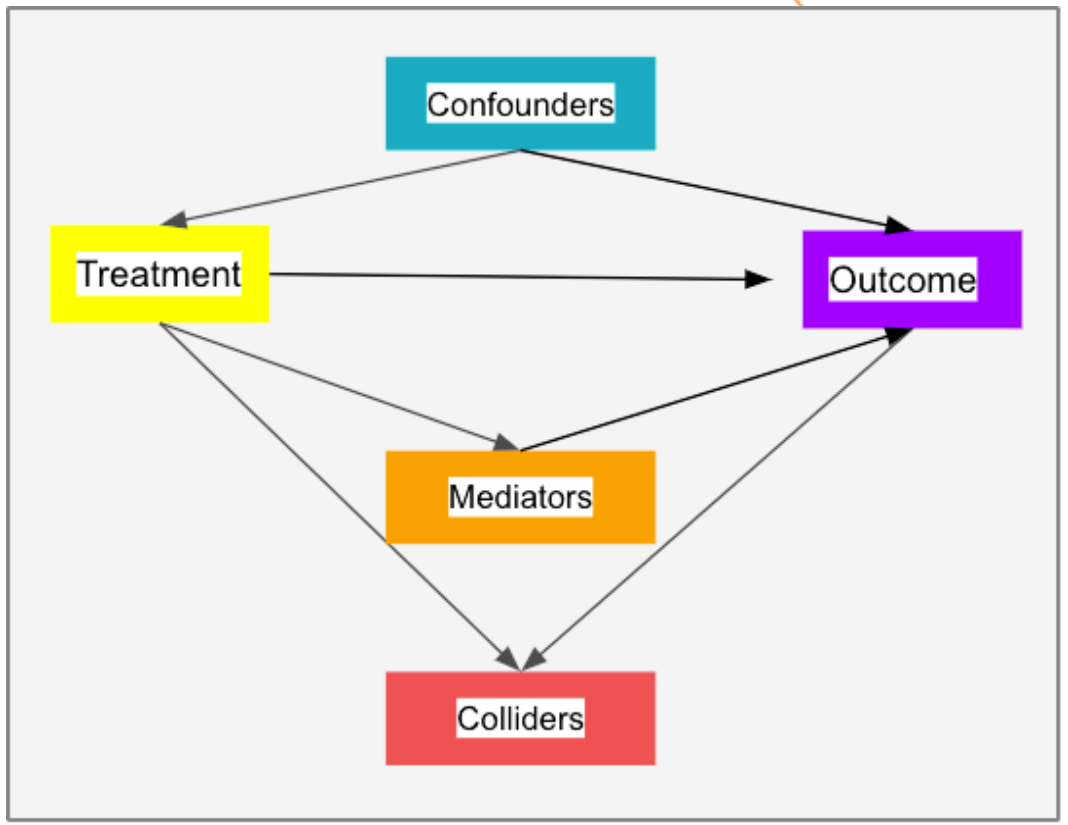
ASVs incorporate the ability to assign varying probabilities to the sequence in which features are presented to the model, thereby determining the influence of each feature on the model's prediction.

Causal Shapley Value

$$\phi_i = \sum_{S \subseteq N \setminus i} \frac{|S|!(n - |S| - 1)!}{n!} \left[v(S \cup i) - v(S) \right]$$

Getting causal inferences

- DAG
- ML based methods
- Causal graphs
- Do calculus

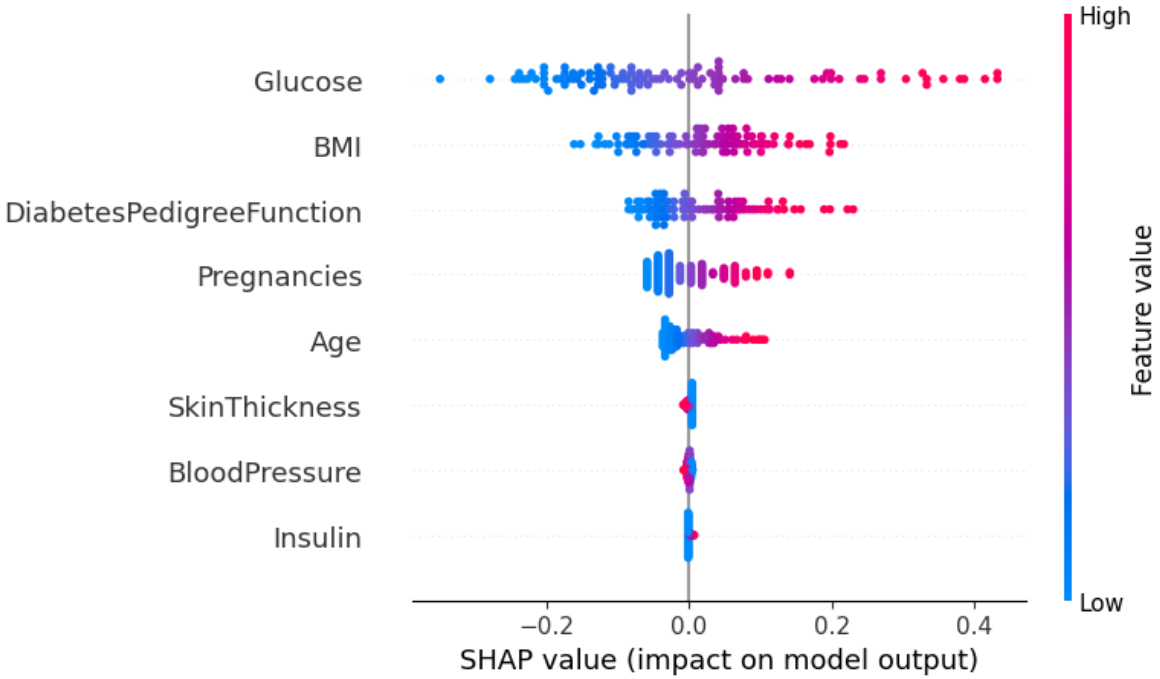


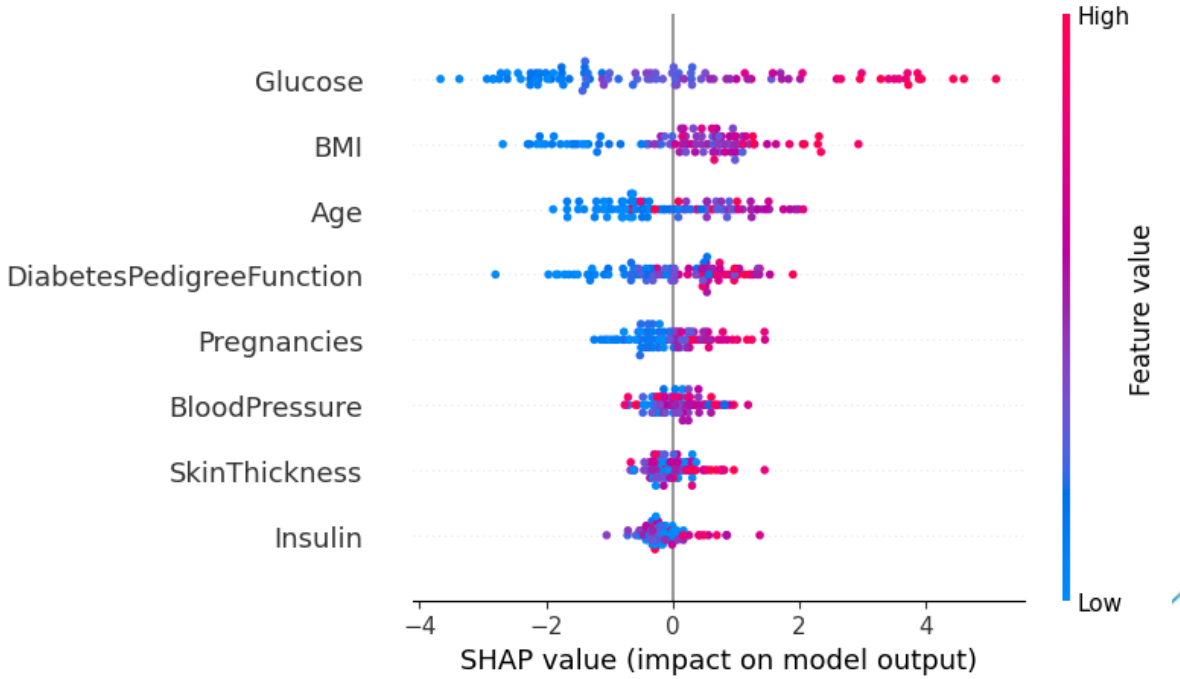
Basic interpretation of SHAP PLOTS

- SHAP magnitude quantifies feature impact on model output.
- Positive (Pink) values increase prediction; negative (Blue) decrease.
- Additive explanation: Sum equals model prediction.
- Captures interaction effects between features.
- Enhances model understanding and trust.
- Offers both global and local interpretability.

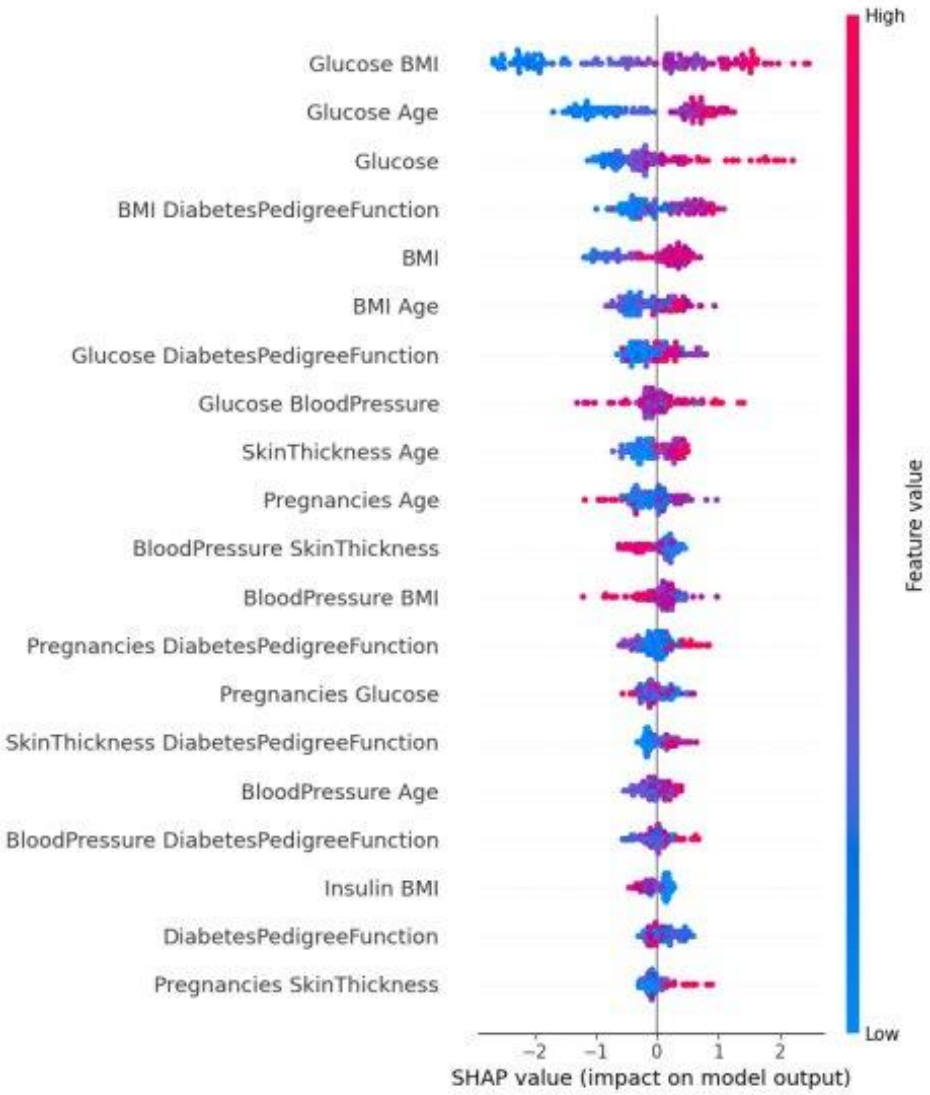
Experimenting with model specific SHAP packages- Python

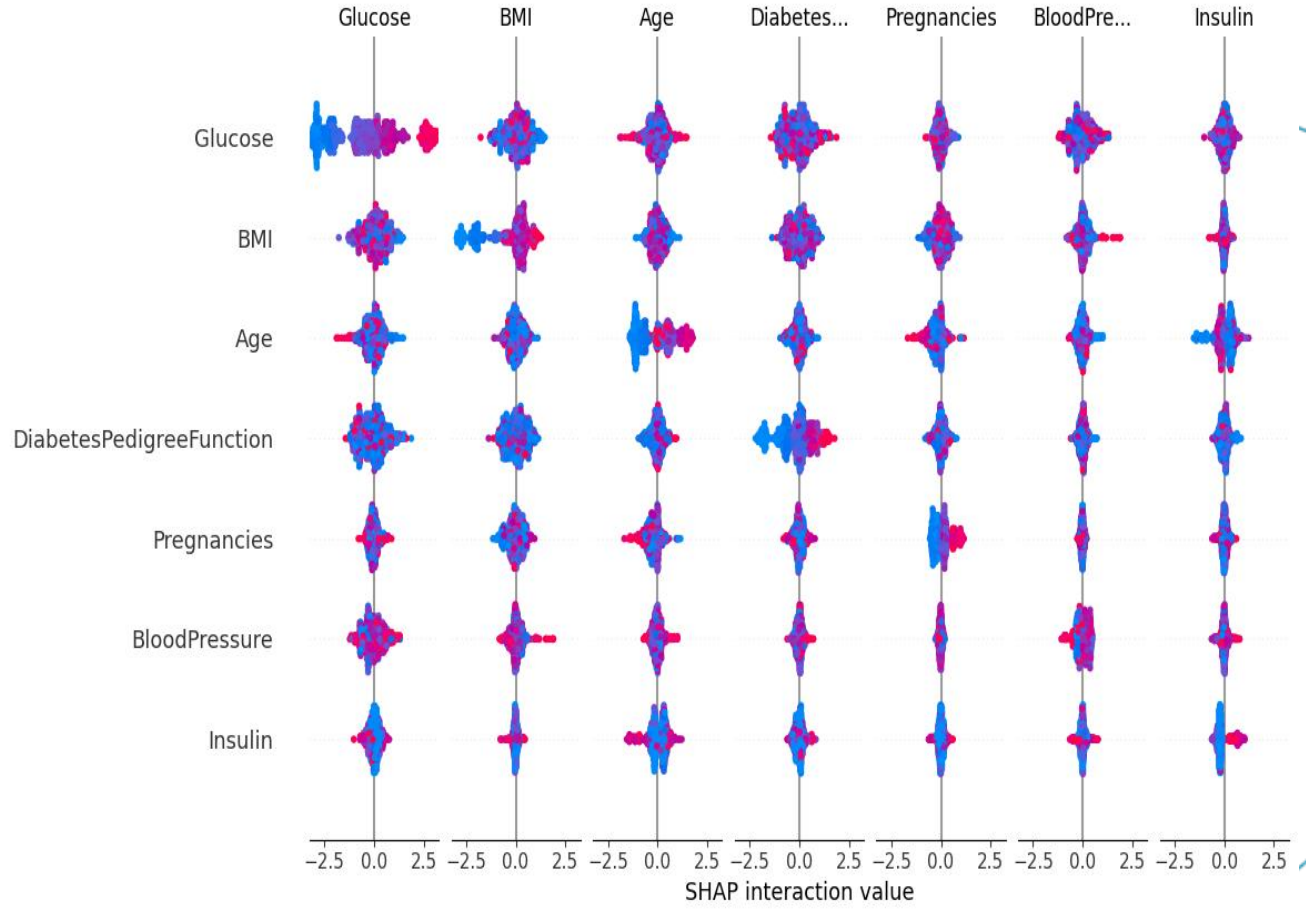
Linear data (Diabetes)






```
Accuracy for XGBoost model: 0.717391304347826
Accuracy for Linear Regression model: 0.7753623188405797
```





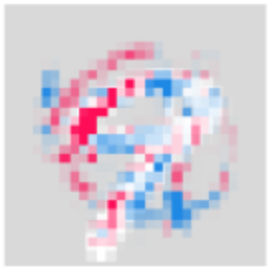
Feature interactions

Deep Explainer

For more complex model (For MNist dataset)

Explanation is OK

Explanation is Bad



No Explanation


```
Test accuracy for CNN model: 0.9896000027656555
```

Advancements

- Several Advancements addressing Computational Cost have been made in recent years:
  - FastSHAP (2020): Estimates Shapley values in a **single forward pass** using a learned explainer model.
  - HarsanyiNet(2023), which simultaneously conducts model inference on the input sample and computes the **exact** Shapley value for each input variable in a **single forward propagation** .
  - Fast Weighted Shapley (FW-Shapley)(2024): Framework for efficiently computing weighted Shapley values using a learned estimator.
- Rational Shapley values
  - Method that satisfies efficiency, linearity, sensitivity, symmetry, and rationality

Challenges And Future Research Directions

- Challenges
  - Computational efficiency and dimensionality of data
  - Interpretability (Especially if we don’t have domain knowledge)
    - Ambiguity in feature interactions
  - Model Sensitivity: Shapley value can be sensitive to changes in the model or training
- Future Research Directions
  - Model Diagnosis : Shapley Values can offer insights into model anomalies for operators or designers.
  - Model Optimization: Encoding prior knowledge through Shapley Value during Model Training.