Explainable AI with Shapley Value

Sriya R G
sriyar@iisc.ac.in

Vishnuja Chand
vishnujac@iisc.ac.in

Sidhant Sharma
ssidhant@iisc.ac.in

Madhuri Nissankararao
madhurin@iisc.ac.in

Abstract

Causal Shapley Values is a model-agnostic approach within the realm of Explainable AI aimed at providing interpretable explanations for individual predictions generated by complex machine learning models. This method utilizes causal knowledge to dissect the contribution of each feature in the model's input to the final prediction. Causal Shapley Values can be decomposed into direct and indirect effects, enhancing the interpretability of feature contributions. This study emphasizes the significance of causality in understanding complex systems and suggests that the proposed approach aligns with human cognitive processes. The research discusses practical implementation, comparisons with other Shapley values, and real-world applications, highlighting potential benefits and risks associated with explanation tools in AI.

1 Introduction

Explainable Artificial Intelligence(XAI) [7], refers to the set of techniques and methods employed to make the outputs of AI models understandable to humans. As AI systems become increasingly integrated into various aspects of our lives, understanding how these systems arrive at their decisions is crucial for trust, accountability, and safety. The core aim of XAI is to make the decisions, predictions, and functioning of AI models more comprehensible to human users. This comprehensibility is essential for several reasons: it fosters trust in AI systems, facilitates their adoption in critical sectors, aids in debugging and improving model performance, and ensures compliance with regulatory requirements that mandate transparency in automated decision-making.

The importance of XAI lies in its potential to enhance transparency and trust in AI systems. By providing explanations for AI decisions, XAI enables users to understand why a particular decision was made, which is essential in domains such as healthcare, finance, and criminal justice where decisions can have significant impacts on individuals' lives.

Within the domain of XAI, feature attribution methods play a pivotal role. These methods aim to quantify the contribution of each input feature to the model's output, providing insights into which features are most influential in the decision-making process. By identifying the features that significantly impact the model's predictions, developers and users can gain a deeper understanding of how the model operates, leading to better interpretability and trust.

Among the myriad of XAI techniques, Shapley values have garnered significant attention for their ability to offer nuanced explanations of model predictions. Originating from cooperative game theory, Shapley values provide a principled framework for attributing the contributions of individual features to model outcomes. This methodology not only enhances transparency but also fosters trust and comprehension among users, regulators, and other stakeholders.

Furthermore, introduction of Asymmetric [1] and Causal Shapley Values [6] has expanded the applicability and utility of this framework in XAI. Asymmetric Shapley values relax the assumption of symmetry in traditional Shapley values, allowing for the consideration of partial causal knowledge about the data-generating process. On the other hand, causal Shapley values incorporate causal relationships among features into the explanation process, offering a more nuanced understanding of feature importance.

2 Evolution of Feature Attribution Methods in XAI

The journey of feature attribution methods in Explainable AI (XAI) traces the path from simple, model-specific techniques to more complex, model-agnostic methods, culminating in the development of causal Shapley values. Here's an outline of this evolution:

2.1 Early Methods and Model-Specific Approaches

Initially, feature attribution in XAI was largely model-specific, designed for particular types of models like linear regression or decision trees. Techniques such as coefficient inspection in linear models or feature importance scores in tree-based models provided straightforward ways to evaluate the influence of each feature. However, these methods lacked generalizability across different types of models and often did not offer insight into complex interactions between features.

2.2 Rise of Model-Agnostic Techniques

The need for methods that could be applied across various model types led to the development of model-agnostic feature attribution techniques. Methods such as LIME[9] (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations)[2] gained popularity. LIME approximates local decision boundaries with simpler models, whereas SHAP leverages the concept of Shapley values to offer consistent and fair attributions across all features for any given prediction.

2.3 Integration of Causal Inference: Towards Causal Shapley Values

The evolution continued with the integration of causal inference principles into feature attribution methods, addressing the limitations of correlational interpretations provided by earlier techniques. This shift recognized the importance of distinguishing between correlation and causation in understanding model decisions. Asymmetric and Causal Shapley values represent a significant advancement in this direction.

Asymmetric Shapley values(ASV) extends the SHAP framework by addressing its symmetry assumption, where all features are considered to contribute equally to interactions. ASV acknowledges that in many real-world scenarios, the contribution of a feature to a model's prediction can be asymmetric, varying significantly depending on the presence or absence of other features. This method fine-tunes the interpretability provided by SHAP, offering more precise insights into complex feature interactions and dependencies within models, thereby enhancing our understanding of model behavior in cases where traditional SHAP values may offer oversimplified explanations.

Compared to ASV, causal Shapley values provide a more direct and robust way to incorporate causal knowledge. Causal Shapley values build upon the foundation laid by traditional Shapley values but integrate causal relationships between features into the attribution process. This approach not only quantifies the contribution of each feature to the prediction but also clarifies how changes in feature values causally affect the outcome. By doing so, causal Shapley values offer a more nuanced and accurate explanation of model behavior, bridging the gap between mere correlation and true causal influence.

This evolution reflects the growing complexity and sophistication of XAI techniques, mirroring the advancement in AI models themselves. As models become more complex, the methods for explaining them also evolve, with causal Shapley values marking the current frontier in providing deep, actionable insights into model decisions.

3 Shapley Values

3.1 Overview of Shapley Values from Cooperative Game Theory

The Shapley value is a concept originating from cooperative game theory that aims to fairly distribute the total payoff among players who contribute to a cooperative endeavor. It was introduced by Lloyd Shapley in 1953 [3]. The Shapley value considers all possible permutations of players and calculates each player's marginal contribution to the cooperative outcome. This contribution is determined by evaluating the difference in

payoff when a player joins or leaves a coalition, and then averaging this difference across all possible orderings of players.

The Shapley value ensures that each player receives a fair share of the total payoff based on their individual contributions to achieving the cooperative goal. The Shapley value, denoted by $\phi_v(i)$, is calculated using the formula:

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

where:

- $N = \{1, 2, \ldots, n\}$ represents the set of players.
- $v(N)$ denotes the value generated by the coalition of all players.
- $v(S)$ represents the value generated by any subset $S \subseteq N$.
- Π is the set of all permutations of N, i.e., all possible orderings of the players.
- $\pi(j) < \pi(i)$ indicates that player j precedes player i in the ordering π.

The Shapley value $\phi_v(i)$ thus represents the marginal contribution that player i makes upon joining the team, averaged over all orderings in which the team can be built.

Shapley values stand as the unique attribution method satisfying the following four axioms:

1. Axiom 1 (Efficiency): $\sum_{i \in N} \phi_v(i) = v(N) - v(\{\})$.
2. Axiom 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. Axiom 3 (Nullity): $\phi_v(i) = 0$ whenever $v(S \cup \{i\}) = v(S)$ for all $S \subseteq N \setminus \{i\}$.
4. Axiom 4 (Symmetry): $\phi_v(i) = \phi_v(j)$ if $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq N \setminus \{i, j\}$.

3.2 Global Shapley Values

The Shapley values $\phi_{f_y(x)}(i)$ presented above provide a local explanation of the individual prediction $f_y(x)$. Global Shapley values for model f are defined by averaging local explanations:

$$\Phi_f(i) = \mathbb{E}_{p(x,y)}[\phi_{f_y(x)}(i)]$$

over the distribution $p(x, y)$ from which the data is sampled. Global Shapley values explain the model's general behavior across the data, remaining consistent with the Shapley axioms. In particular, the global Shapley value $\Phi_f(i)$ can be interpreted as the portion of model f's accuracy attributable to feature i. This follows from the sum rule:

$$\sum_{i \in N} \Phi_f(i) = \mathbb{E}_{p(x,y)}[f_y(x)] - \mathbb{E}_{p(x_0)}\mathbb{E}_{p(y)}[f_y(x_0)]$$

The first term on the right is the accuracy one achieves by sampling labels from f's predicted probability distribution over classes. (Note that this is distinct from the accuracy of predicting the max-probability class.) The offset term is the accuracy one is left with using none of the features: predicting the label of x by sampling from the model's output $f_y(x_0)$ on randomly drawn x_0.

3.3 SHAP (SHapley Additive exPlanations)

Rooted in the concept of Shapley values, SHAP [2] offers local interpretability by attributing the model's output to each feature, providing insights into feature importance at a granular level. The calculation of SHAP values involves quantifying the contribution of each feature to the prediction by considering the difference between the model's prediction for a specific instance and the average prediction across all instances. This additive feature attribution method allows SHAP to offer valuable insights into how individual features influence model predictions, enabling users to understand the decision-making process of complex black-box models such as neural networks and gradient boosting machines. With its ability to provide transparent and understandable explanations for individual predictions, SHAP enhances trust and comprehension in AI systems, making it a valuable tool in the XAI toolkit. In SHAP (SHapley Additive exPlanations), the formula for calculating SHAP values is as follows:

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

where:

- N is the set of all features.
- S is a subset of features excluding feature i.
- $f(S)$ represents the model's output when only the features in set S are considered.
- $f(S \cup \{i\})$ represents the model's output when feature i is added to set S.
- $|S|$ denotes the number of features in subset S.
- $|N|$ is the total number of features.

This formula calculates the SHAP value $\phi_i(x)$ for feature i in the context of a specific instance x. It attributes the difference between the model's prediction when feature i is included and excluded from a subset of features to feature i, scaled by the number of possible feature combinations. This additive feature attribution method based on Shapley values allows SHAP to provide interpretable explanations for individual predictions of machine learning models.

3.4 Asymmetric Shapley Values

Asymmetric Shapley values (ASVs) were introduced [1] to enhance Shapley value explanations by incorporating partial causal knowledge about the data-generating process. This involves relaxing the Symmetry axiom to address scenarios where certain features exhibit deterministic causal relationships. The Symmetry axiom in traditional Shapley values implies that features with the same effect on the value function will have identical Shapley values. However, if one feature is known to be a deterministic causal ancestor of another, it may be preferable to attribute all importance to the ancestor feature. ASVs achieve this by introducing a probability distribution over permutations, denoted as $w : \Pi(N) \to [0, 1]$. This distribution is used to weigh the contributions of features based on their causal relationships.

The formula for ASVs is given by:

$$\phi_w^v(i) = \sum_{\pi \in \Pi(N)} w(\pi)[v(\{j : j \prec_\pi i\}) - v(\{j : j \prec i\})]$$

Where $i \prec_\pi j$ denotes that i is a predecessor to j in permutation π. If w is the uniform distribution, the original Shapley values are recovered. The choice of probability distribution w allows for the incorporation of causal prior knowledge into the explanation.

Given partial causal knowledge represented by a partial order $\prec$, where $i \prec j$ if i is an ancestor of j in the causal directed acyclic graph (DAG), two approaches are specified to incorporate this knowledge: one attributes more importance to distal (root) causes, and the other to proximate (direct) causes. The distal

approach assigns weights to permutations that preserve the causal ancestry relationships, while the proximate approach assigns weights to permutations that place causal ancestors after their descendants.

Furthermore, it's noted that ASVs can be combined with either the conditional value function $v_{fy}(x|c)$ or the marginal value function. The choice between these options depends on factors such as computational complexity and the availability of conditional distributions of the dataset, but both approaches offer advantages and disadvantages similar to those of traditional Shapley values.

ASVs uniquely satisfy Axioms $1-3$. They do not satisfy Axiom 4 (Symmetry) unless the distribution $w \in \Delta(\Pi)$ is uniform, in which case they reduce to the Shapley values of Eq. (1).

3.5 Causal Shapley Values

To address interdependence among model features, SHAP calculation can utilize the conditional distribution of excluded features. Additionally, considering the causal structure of features involve conditioning absent features on included ones through intervention. This is achieved by using the interventional distribution in the sampling procedure, typically represented by the do-operator in do-calculus. By applying do-calculus rules, conditional SHAP can be modified to incorporate interventional distribution, allowing the calculation of causal SHAP values. Causal SHAP values account for both direct and indirect effects of model features, capturing changes in the model output due to alterations in specific features while considering the influence on absent features. This inclusion of indirect effects distinguishes causal SHAP values from marginal SHAP values, which only account for direct effects.

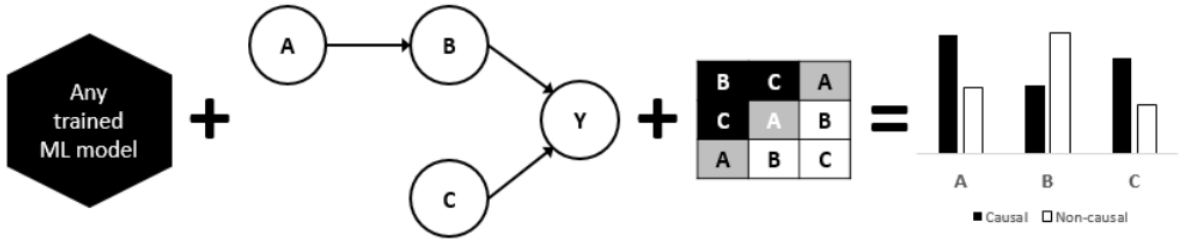


Figure 1: Modelling Causal Shapley values

Causal Shapley Values (CSVs) [6] extend the traditional Shapley values framework by incorporating causal knowledge about the data-generating process. This integration allows for a more nuanced attribution of feature importance in machine learning models, considering causal relationships among features. CSVs address scenarios where certain features may have direct causal effects on others, providing a means to differentiate between direct and indirect contributions.

Mathematically, the formula for CSVs is derived by modifying the traditional Shapley value formula to incorporate causal relationships. Let's denote:

- N as the set of features.
- v as the characteristic function representing the model's performance (e.g., predictive accuracy) given different feature subsets.
- S as a subset of features, where $S \subseteq N$.
- $n(S)$ as the number of features in subset S.
- π as a permutation of features.

The formula for CSVs is as follows:

$$\phi_v(i) = \sum_{\pi \in \Pi(N)} \frac{(v(S \cup \{i\}) - v(S))}{(n(S)+1)!}$$

Where S ranges over all subsets of features that do not contain i, and $S \cup \{i\}$ denotes the feature subset obtained by adding feature i to subset S. This formulation remains similar to traditional Shapley values, but it takes into account the causal relationships among features when evaluating their contributions.

CSVs allow for a more nuanced understanding of feature importance by accounting for causal dependencies among features. By incorporating causal knowledge into the Shapley value framework, CSVs provide insights into the direct and indirect impacts of features on the model's performance, enhancing interpretability and robustness in machine learning models.

4 Key Methodologies for Calculating Causal Shapley Values

Calculating causal Shapley values involves methodologies that integrate the principles of causal inference with the Shapley value framework from cooperative game theory. These methodologies aim to attribute the contribution of each feature to a prediction by considering not only the feature's marginal contribution but also the causal relationships among features. Here's a summary of key methodologies and advancements:

4.1 Integration with Causal Graphs

One approach involves leveraging causal graphs (or directed acyclic graphs - DAGs)[4, 8] to model the causal relationships between features. By using these causal graphs, researchers can more accurately estimate the marginal contributions of features by considering the pathways through which features exert causal influences on the prediction outcome.

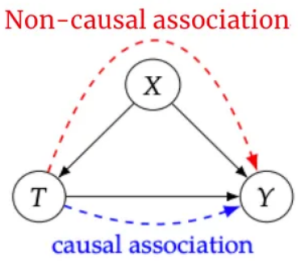


Figure 2: causal graphs

4.2 Do-Calculus and Interventional Distributions

To compute causal Shapley values, researchers utilize do-calculus [5] to derive interventional distributions, which represent the effect of manipulating individual features while holding others constant. By accounting for these interventional distributions, causal Shapley values capture the causal impact of each feature on model predictions.

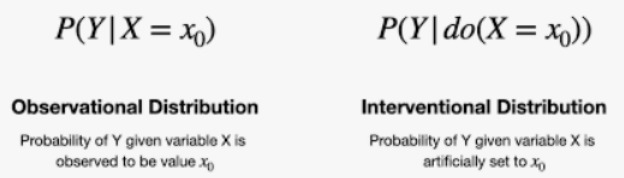


Figure 3: Do-calculus Interventional Distribution

4.3 Counterfactual Explanations

Causal Shapley values can be calculated using counterfactual reasoning, where the contribution of a feature is assessed by comparing the actual prediction with a counterfactual prediction made in a hypothetical scenario

where the feature's value is altered. This method emphasizes understanding the causal effect of changing a feature, highlighting the difference between mere correlation and causation.

4.4 Conditional Expectation-Based Approaches

These approaches calculate Shapley values based on the expected change in prediction when conditioning on different subsets of features, explicitly incorporating causal assumptions about feature interactions. This methodology requires a deep understanding of the model's structure and the data generation process.

5 Feature Interactions

1. **Polynomial Transformation**: Given a dataset with n features, the polynomial transformation generates new features by considering all possible combinations of the original features up to a specified degree d. For each original feature x_i, the transformation creates new features of the form x_i^j, where j varies from 1 to the specified degree d, and also includes interaction terms involving multiple features.

2. **Mathematical Representation**: Let's denote the original feature vector as $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ and the polynomial transformation of degree d as $\phi(\mathbf{x})$. The transformed feature vector $\mathbf{z} = [z_1, z_2, \ldots, z_m]$, where m is the total number of features after transformation, can be represented as:

$$\mathbf{z} = \phi(\mathbf{x}) = [1, x_1, x_2, \ldots, x_n, x_1^2, x_1x_2, \ldots, x_n^d, \ldots, \text{interactions}]$$

   Here, *interactions* represents all interaction terms involving multiple features, such as x_1x_2, x_1x_3, etc., up to degree d.

3. **Interpretation of Interaction Terms**: The interaction terms capture the combined effect of multiple features on the target variable. For example, if x_1 represents the age of a person and x_2 represents their income, the interaction term x_1x_2 could capture how the impact of age on the target variable (e.g., medical expenses) varies with income level. Similarly, higher-order interaction terms like $x_1^2x_2$ could capture more complex relationships.

4. **Importance of Interaction Features**: When fitting a model using the transformed features, the importance of each feature, including interaction terms, can be determined based on their contribution to the model's performance (e.g., predictive accuracy or error). Features with higher importance indicate stronger associations with the target variable, either individually or through interactions with other features.

5. **Limitations and Considerations**: While polynomial transformation allows capturing nonlinear relationships and interactions between features, it also introduces a larger feature space, which can lead to overfitting if not properly regularized. Moreover, interpretation of high-order interaction terms becomes increasingly complex, and domain knowledge may be required to make meaningful interpretations.

6 Interpretation of SHAP plots

SHAP (SHapley Additive exPlanations) offers several analysis techniques to interpret the predictions of machine learning models. Some of the main analysis techniques provided by SHAP include:

- **Summary Plot**: Aggregates the impact of each feature across all instances in a dataset, providing a global view of feature importance.

  Interpretation of Summary Plot
  Each feature is represented by a vertical line. Here's how to interpret the values and colors:

  - **Vertical Position**: Indicates the impact of the feature on model output. Features are sorted based on importance, with the most important at the top.

- **Horizontal Spread**: Represents the range of SHAP values across the dataset. A wider spread indicates a more significant impact.
- **Color**: Indicates the value of the feature. Higher values are in shades of red, while lower values are in shades of blue.

The Summary Plot provides a concise overview of feature importance and the relationship between feature values and model predictions.

- **Dependence Plot**: Visualizes the relationship between a specific feature and the model's output, showing how changes in the feature affect the predicted values.

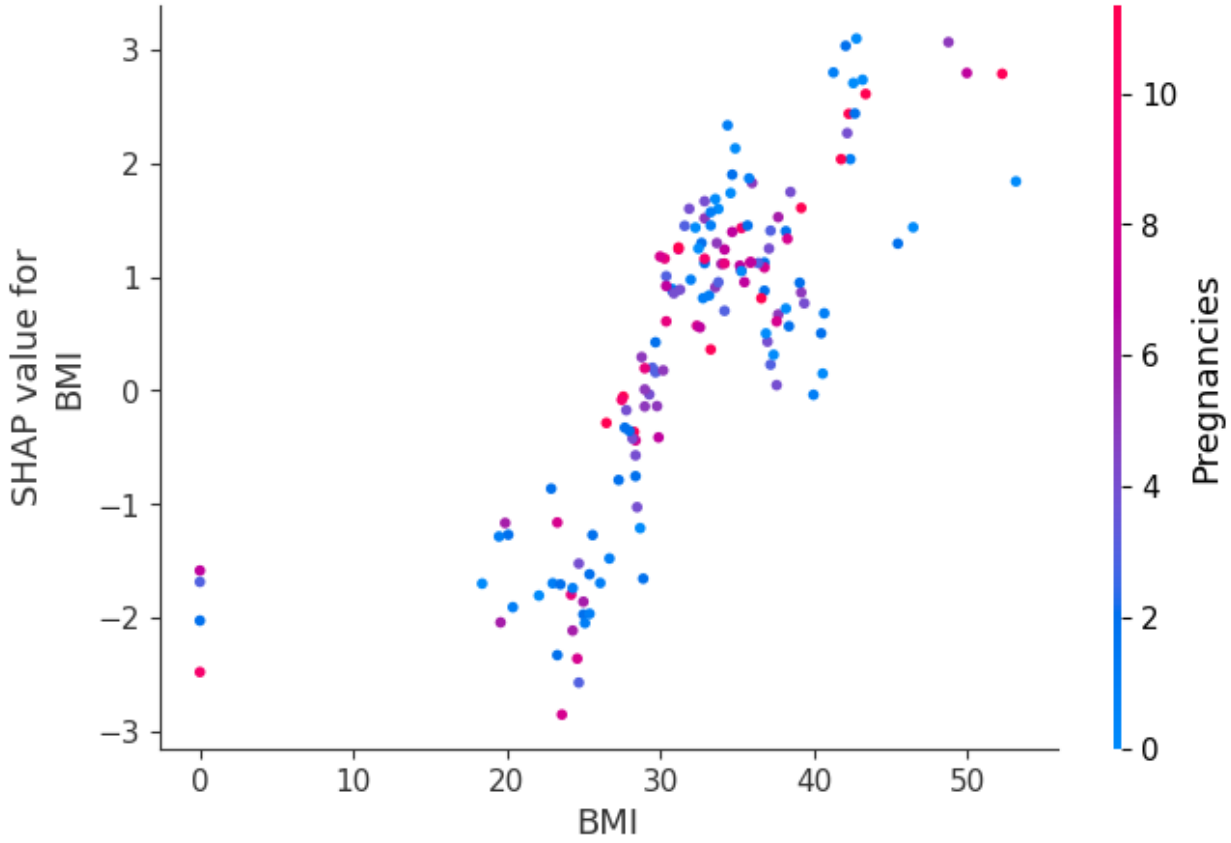


Figure 4: Dependence plot for classifier experiment

Interpretation of Dependence Plot
Each dot represents a single instance in the dataset. Here's how to interpret the values and colors:

- **Horizontal Axis**: Represents the value of the feature being analyzed.
- **Vertical Axis**: Indicates the SHAP value for the feature. It represents the impact of the feature on the model output for each instance.
- **Color**: Reflects the value of a second feature that may interact with the analyzed feature. The color scale indicates the value of the second feature, with higher values typically in shades of red and lower values in shades of blue.

Dependence Plots visualize how the model output changes as the value of a single feature varies, while considering potential interactions with another feature.

- **Interaction Values**: Quantifies the contribution of each feature pair to the difference between the predicted and baseline values.

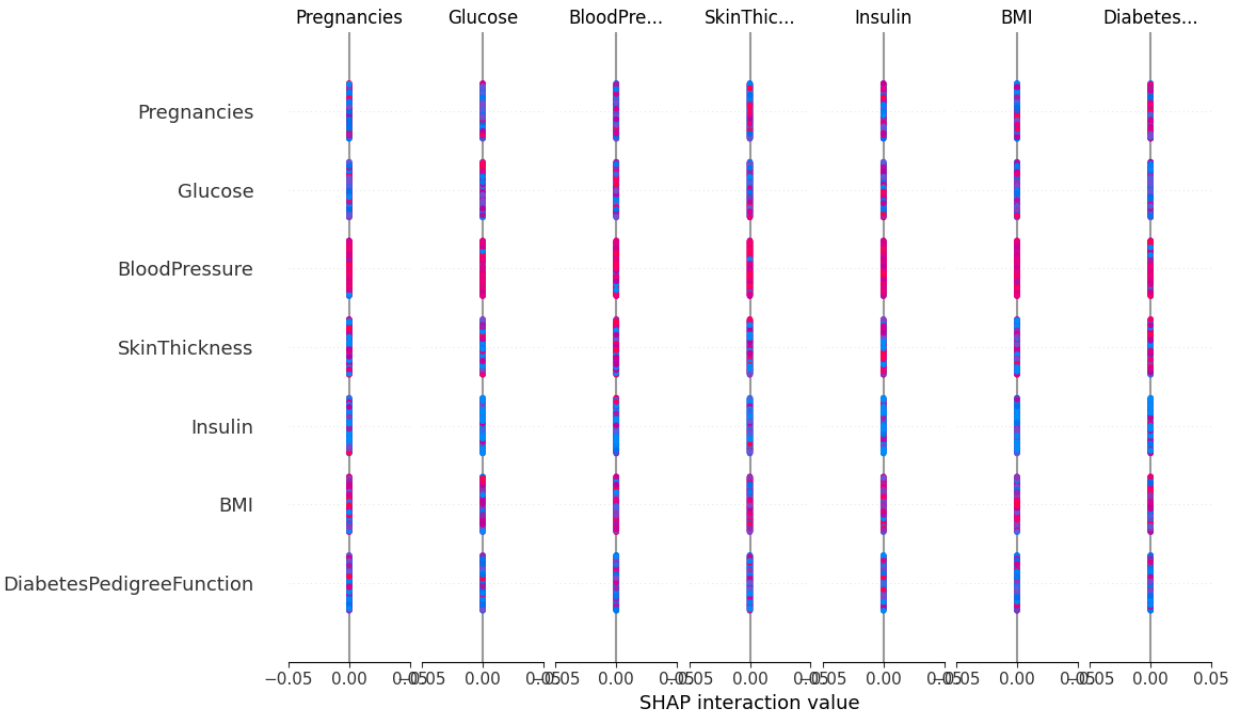


Figure 5: Interaction values for classifier experiment

Interpretation of Interaction Values
In SHAP interaction plots, each point represents a single instance in the dataset. Here's how to interpret the values and colors:

  - **Horizontal Axis**: Represents the SHAP value for the first feature.
  - **Vertical Axis**: Indicates the SHAP value for the second feature.
  - **Color**: Reflects the value of a third feature that may interact with both the first and second features. The color scale indicates the value of the third feature, with higher values typically in shades of red and lower values in shades of blue.

  Interaction Plots visualize the interaction effects between two features on the model output, while considering the influence of a third feature.

- **Force Plot**: Provides a detailed breakdown of how individual features contribute to a particular prediction, allowing for the interpretation of individual predictions. **Interpretation of Force Plots** SHAP force plots provide a detailed breakdown of the factors contributing to each individual prediction. Here's how to interpret the key elements:

  - **Base Value**: The model's average prediction across the dataset, represented by the horizontal dashed line. It serves as the reference point.
  - **Feature Contributions**: Vertical bars indicate how much each feature contributes to shifting the prediction away from the base value. Positive values indicate features that increase the prediction, while negative values represent features that decrease the prediction.
  - **Impact of Features**: The length of each bar represents the magnitude of the feature's contribution. Longer bars have a greater impact on the prediction.

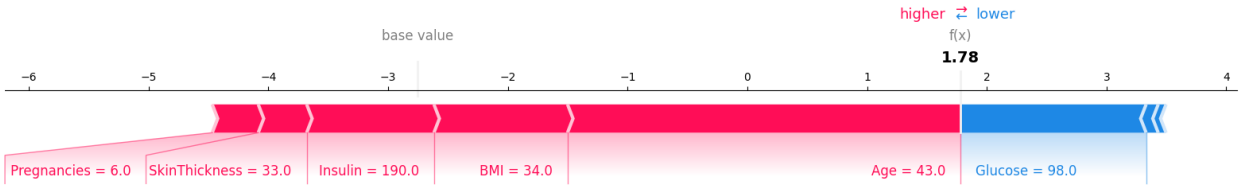


Figure 6: Force plot for classification experiment

– **Final Prediction**: The sum of the base value and the feature contributions yields the final prediction for the individual instance, indicated by the solid dot.

Force Plots provide insights into how individual features influence model predictions for specific data points, helping understand the model's decision-making process.

- **Waterfall Plot**: Illustrates the decomposition of the model's prediction for a single instance into contributions from each feature.

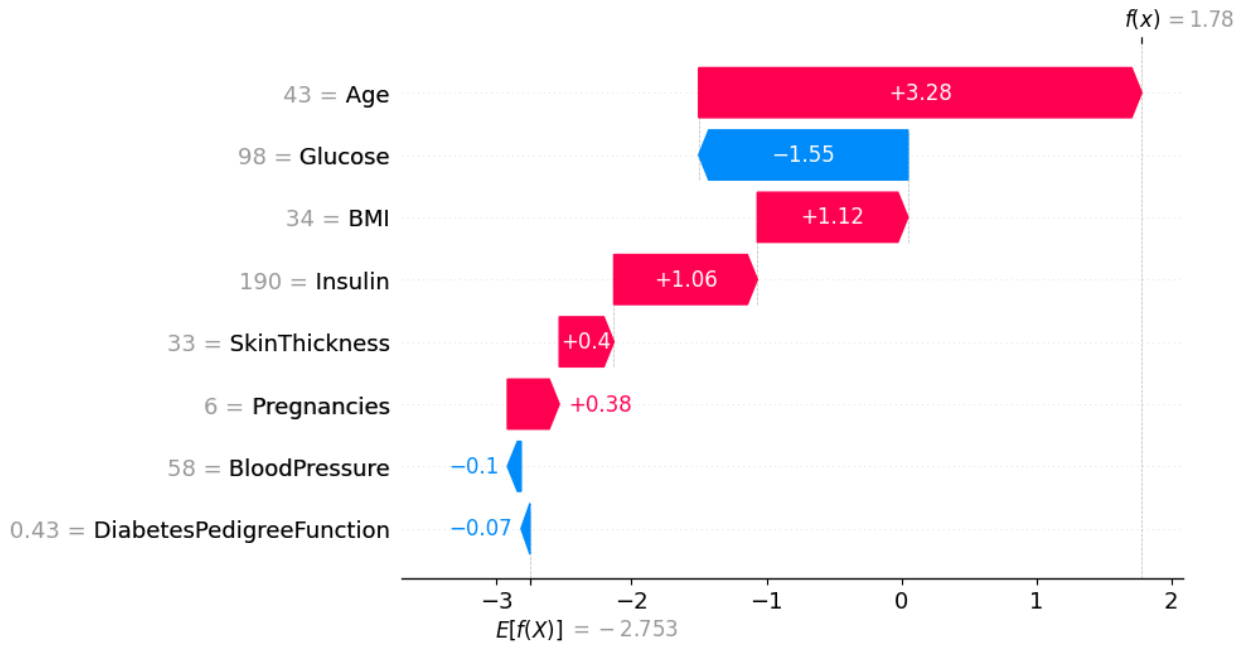


Figure 7: Waterfall plot from classification experiment

Interpretation of Waterfall Plots
SHAP waterfall plots provide a step-by-step breakdown of how features contribute to the difference between a specific prediction and the base value. Here's how to interpret the key elements:

– **Base Value**: The starting point of the plot, representing the model's average prediction across the dataset.

- **Step Contributions**: Each step in the plot represents the contribution of a feature to the prediction. Positive steps indicate features that increase the prediction, while negative steps represent features that decrease the prediction.
- **Impact of Features**: The length of each step indicates the magnitude of the feature's contribution. Longer steps have a greater impact on the prediction.
- **Cumulative Effects**: The cumulative sum of the steps represents the cumulative effect of all features on the prediction. It shows how each feature's contribution adds up to the final prediction.
- **Final Prediction**: The last step of the plot corresponds to the final prediction for the individual instance, obtained by adding the cumulative effects to the base value.

Waterfall Plots provide a clear visualization of how individual features contribute to model predictions for specific data points, aiding in the understanding of the model's decision-making process.

- **Summary Interaction Plot**: Summarizes the interaction effects between pairs of features across all instances in a dataset.

  Interpretation of Summary Interaction Plots
  SHAP summary interaction plots visualize the interactions between pairs of features and their impact on model predictions. Here's how to interpret the key elements:

  - **Feature Interaction Effects**: Each point on the plot represents the interaction effect between two features on model predictions. The x-axis and y-axis correspond to the SHAP values of the two interacting features.
  - **Color and Size**: The color and size of each point indicate the magnitude and direction of the interaction effect. Darker and larger points represent stronger interaction effects, while lighter and smaller points indicate weaker interactions.
  - **Diagonal Line**: The diagonal line represents no interaction effect between the two features. Points above the diagonal indicate positive interaction effects, where the joint presence of both features leads to higher predictions than expected based on their individual effects. Points below the diagonal represent negative interaction effects, where the joint presence of both features leads to lower predictions than expected.
  - **Trend and Patterns**: By examining the distribution of points and trends in the plot, you can identify which feature pairs have the most significant interaction effects and understand how the joint presence of features influences model predictions.

  Summary Interaction Plots help uncover complex interactions between features in predictive models, providing insights into how feature combinations affect predictions.

- **Decision Plot**: Visualizes the decision process of a model by displaying the path through a decision tree or other models for a specific instance.

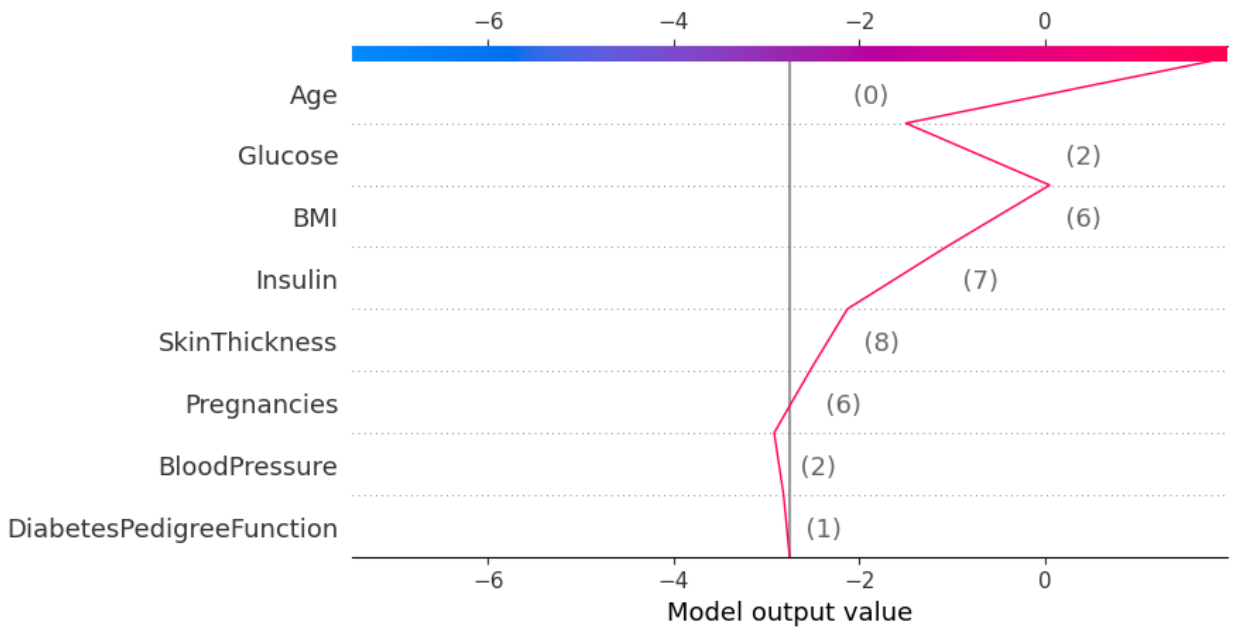


Figure 8: Decision plot for classification experiment

Interpretation of Decision Plots
Decision plots in SHAP visualize how model predictions change as the value of a specific feature varies. Here's how to interpret the key elements:

- **Feature Value Range**: The x-axis of the plot represents the range of values for the selected feature.
- **Predicted Output**: The y-axis shows the model's predicted output (e.g., probability, class probability, regression value) corresponding to each feature value.
- **Colored Line**: The colored line on the plot indicates the SHAP value of the selected feature across its value range. It represents how the feature's contribution to the model's prediction changes as its value varies.
- **Diverging Points**: Points above and below the colored line represent cases where the actual output diverges from the predicted output based on the feature's SHAP value. These points highlight instances where the model's prediction differs from what would be expected based solely on the feature's influence.
- **Influential Regions**: By examining the slope and shape of the colored line, you can identify influential regions of the feature's value range where its impact on model predictions is particularly strong or weak.

Decision Plots provide insights into how individual features contribute to model predictions and how their effects vary across different feature values.

- **SHAP Summary Plot**: Visualizes the impact of each feature on model predictions across the dataset.

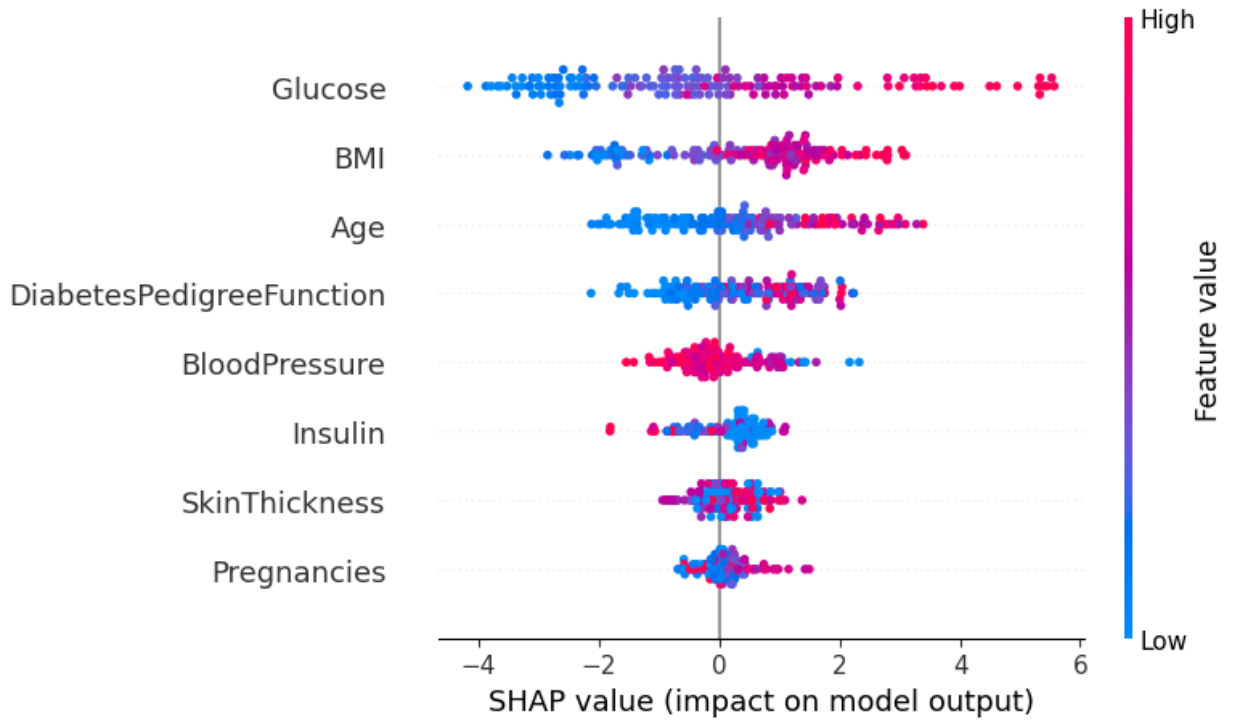


Figure 9: Summary plot for classification experiment

Interpretation of SHAP Summary Plot
The SHAP summary plot visualizes the impact of each feature on model predictions across the dataset. Here's how to interpret the key elements:

- **Feature Importance**: Features are ranked in descending order of importance based on the sum of the absolute SHAP values across all data points.
- **Horizontal Bars**: Each horizontal bar represents a feature, and its length corresponds to the magnitude of the average SHAP value for that feature across all data points.
- **Color Gradient**: The color of the bars represents the direction and magnitude of the SHAP values. Blue indicates lower feature values, while red indicates higher feature values.
- **Error Bars**: Error bars extending from each bar indicate the variability or uncertainty in the SHAP values for that feature across different data points.
- **Vertical Line**: A vertical line at zero indicates the baseline or reference point for SHAP values. Features to the right of the line have positive SHAP values, indicating they push predictions higher, while features to the left have negative SHAP values, indicating they push predictions lower.

The SHAP summary plot provides a concise overview of feature importance and their effects on model predictions, facilitating interpretation and decision-making in model analysis.

7 Comparison with Other Feature Attribution Methods

7.1 Traditional Feature Importance Scores

Causal Shapley values provide more nuanced insights by considering causal relationships between features, whereas traditional feature importance scores often lack this causal context.

7.2 Partial Dependence Plots (PDPs)

PDPs visualize the marginal effect of individual features on model predictions but do not explicitly consider causal relationships. Causal Shapley values offer a more comprehensive understanding of feature importance by accounting for causal dependencies.

7.3 Traditional Shapley Values

Traditional Shapley values provide a way to distribute "credit" among features based on their contributions to the model's output, without considering the causal relationships between features. Causal Shapley values extend this by incorporating causal inference, offering a more nuanced understanding that differentiates between correlation and causation.

7.4 LIME and SHAP

LIME creates interpretable models locally around predictions but does not inherently account for causal relationships between features. It is more focused on local approximations and may not capture the overall model complexity or causal effects.

SHAP(non-causal) utilizes the concept of Shapley values to explain the output of any model in a consistent and additive manner. While SHAP provides a robust framework for feature attribution, it primarily captures correlation rather than causal influence.

Causal Shapley values offer a deeper layer of explanation by identifying not just the importance of features, but also how changes in those features causally impact the model's predictions.

In summary, while traditional feature attribution methods like SHAP and LIME have significantly advanced the field of XAI by providing interpretable insights into complex models, causal Shapley values represent a further step towards understanding the "why" behind model predictions. This advancement addresses the critical need for discerning causal relationships in predictive modeling, thus offering more actionable and trustworthy explanations. However, the complexity and computational demands of calculating causal Shapley values, along with the requirement for robust causal assumptions, present ongoing challenges that are the focus of current research efforts.

8 Applications

8.1 Healthcare

Personalized Medicine: In personalized medicine, causal Shapley values can elucidate how different genetic and environmental factors contribute to the risk of diseases or the efficacy of treatments. By understanding the causal impact of each factor, clinicians can tailor treatments more effectively to individual patients, improving outcomes.

Disease Prediction and Prevention: AI models that predict disease outbreaks or patient risks can benefit from causal Shapley values by identifying not just correlations but causal factors that can be targeted for intervention. This has profound implications for preventive medicine, where interventions can be prioritized based on their causal impact on disease risk.

8.2 Finance

Credit Risk Modeling: In finance, causal Shapley values can improve the transparency of credit risk models by identifying the causal effect of various factors (e.g., income level, employment history, past credit behavior) on credit scores. This helps in making fairer lending decisions and in explaining those decisions to applicants, enhancing trust.

Algorithmic Trading: For models used in algorithmic trading, understanding the causal relationships between market indicators and stock movements is crucial. Causal Shapley values can help in identifying which factors genuinely drive market trends, leading to more robust trading strategies.

8.3 Environmental Modeling

Climate Change Analysis: In environmental science, causal Shapley values can be used to assess the impact of various factors (e.g., CO2 emissions, deforestation, urbanization) on climate change models. This can guide policy decisions by highlighting the most effective areas for intervention.

Pollution Control: For models predicting pollution levels, causal Shapley values can identify the primary contributors to pollution in different areas. This information is invaluable for designing targeted environmental policies that can effectively reduce pollution.

9 Impact on Explainability and Trustworthiness

The applications of causal Shapley values across these diverse domains significantly enhance the explainability and trustworthiness of AI systems in several ways:

Enhanced Decision-Making: By providing a clear understanding of how and why certain features influence model predictions, causal Shapley values enable better-informed decision-making, grounded in causality rather than correlation.

Increased Model Transparency: These applications demonstrate how causal Shapley values can open the "black box" of complex AI models, making their inner workings more transparent to users and stakeholders.

Improved Fairness and Equity: In domains like healthcare and finance, where fairness is critical, causal Shapley values can help identify and mitigate biases by distinguishing between causal effects and spurious correlations.

Fostering Trust: By offering explanations that align more closely with human intuition about causality, causal Shapley values build trust among users, facilitating the wider adoption and acceptance of AI systems.

Regulatory Compliance: As regulations around AI transparency and accountability become stricter, the ability of causal Shapley values to provide detailed causal explanations helps organizations comply with these regulatory requirements.

In summary, the application of causal Shapley values across various domains not only improves the explainability and interpretability of AI systems but also enhances their trustworthiness by providing insights that are actionable, fair, and grounded in causal understanding.

10 Advancements

Advancements in utilizing the Shapley value in various domains of machine learning have been significant. In feature selection tasks, the Shapley value serves to evaluate the importance of features within predictive models. It treats input features as players and considers model performance as the payoff, thereby enabling the quantification of individual feature contributions to overall model performance. Similarly, in data valuation scenarios, the Shapley value is applied by treating training set data points as players. Through the computation of Shapley values, this approach offers insights into the significance of individual data points in influencing model predictions. Furthermore, in the context of federated learning, modeling the collaborative process as a cooperative game allows for fair credit attribution to data owners based on their contributions to model performance. The Shapley value assists in assessing each data owner's impact on the collaborative model. Moreover, in explainable machine learning, employing the Shapley value at the instance level facilitates the provision of explanations for model predictions. This enhances model interpretability by quantifying the contributions of individual input features to predictions. Additionally, in multi-agent reinforcement learning scenarios, the Shapley value facilitates fair credit allocation to participating agents based on their contributions to achieving global rewards, thus promoting cooperative behavior. Lastly, in model valuation within ensembles, utilizing the Shapley value enables the assessment of individual machine learning models' contributions. This aids in evaluating each model's effectiveness in improving the ensemble's predictive performance.

In summary, this paper introduces rational Shapley values as a novel approach to address challenges in explainable artificial intelligence (XAI). By integrating feature attributions and counterfactuals, rational Shapley values offer a versatile method for delivering personalized explanations suited to individual

preferences and beliefs.The paper acknowledges potential concerns regarding scalability and susceptibility to confirmation bias but argues that advancements in computational capabilities, probabilistic modeling techniques, and domain-specific knowledge can mitigate these challenges. Moreover, emphasizing transparency in the explanation process helps counteract the risk of confirmation bias.Overall, rational Shapley values present a practical and intuitive solution for XAI, empowering users to comprehend and trust the outputs of complex machine learning models. Through the framework's alignment with expected utility principles, it furnishes concise and actionable insights that can guide decision-making across various applications.

FastSHAP consistently outperforms baseline methods in both qualitative and quantitative assessments. Visually, FastSHAP and GradCAM excel in highlighting important objects, while KernelSHAP often introduces noise. Quantitative metrics confirm FastSHAP's superiority, particularly in pinpointing informative image regions, as indicated by its high Exclusion AUC. Additionally, FastSHAP demonstrates competitive performance in Inclusion AUC, ranking second only to KernelSHAP-S. Moreover, FastSHAP's rapid processing speed makes it ideal for real-time applications, surpassing KernelSHAP in both training and explanation run-times for 1,000 images. Overall, FastSHAP's efficient Shapley value estimation offers promising prospects across various domains.[10][11][12][13][14][15]

11 Challenges in Implementation

11.1 Complexity of Implementation

Implementing Shapley values can be complex, especially for large-scale or high-dimensional datasets. The computation involves considering all possible coalitions of features, which can become computationally intensive.

11.2 Data availability and quality

Shapley values require access to the model predictions and feature values, which may not always be readily available or of high quality. Dealing with missing data, noisy features, or biased datasets can impact the reliability of Shapley value interpretations.

11.3 Model Specific Consideration

The interpretation of Shapley values can vary depending on the type of model used (e.g., linear models, tree-based models, neural networks). Adapting Shapley values to different model types and understanding their behavior across various models can be challenging.

11.4 Computational Resources

Computing Shapley values can be resource-intensive, especially for complex models or large datasets. It may require significant computational resources and time, which can be a limitation for real-time applications or resource-constrained environments

12 Conclusions

Shapley values stand as a powerful and versatile tool for interpreting and understanding the contributions of individual features in cooperative game theory and machine learning. Their ability to fairly attribute the value generated by a coalition of players or features provides valuable insights into feature importance, aiding in model interpretability, decision-making, and fairness considerations.

While Shapley values offer significant benefits, challenges remain in their practical implementation, scalability, and interpretability, particularly in the context of complex models and large datasets. However, ongoing advancements in algorithmic efficiency, scalability, and interpretability are addressing these challenges and expanding the applicability of Shapley values to real-world settings.

Moreover, the integration of causal inference methods with the Shapley value framework has led to the development of causal Shapley values, offering a causal interpretation of feature importance and uncovering the causal mechanisms underlying model predictions.

From experiments, it can be concluded that model's interpretability is inversely proportional to complexity. Having domain knowledge makes causal inferences easier to validate and compute.

13 Appendix

13.1 Experimental Setup

13.2 Content

PIMA INDIANS DIABETIS DATASET : The Pima Indians Diabetes Dataset includes attributes related to the onset of diabetes and its complications. Epidemiological evidence suggests that type 2 diabetes mellitus (T2DM) results from interactions between genetic and environmental factors, highlighting the importance of considering various attributes in predicting diabetes.

PROGRESSIVE DIABETIS DATASET: The dataset consists of 442 instances and includes ten baseline variables such as age, sex, BMI, blood pressure, and six blood serum measurements. Each instance also contains a target variable, "disease progression," representing a quantitative measure recorded one year after baseline.

Collected from 442 diabetes patients, the dataset provides comprehensive information on baseline characteristics and disease progression over time. It serves as the foundation for training a regression model to predict disease progression accurately, aiding in the management of diabetes in clinical settings.

13.3 Inspiration

The inspiration for this dataset comes from the urgent need to accurately predict diabetes mellitus (DM) to facilitate early diagnosis and intervention. Despite the availability of advanced diagnostic technologies, the complexity of the disease necessitates attention to multiple factors. Machine learning algorithms applied to the Pima Indians Diabetes dataset (PIDD) offer the potential to construct a prediction model with higher accuracy, aiding in better management and treatment of diabetes.

13.4 Related Information

The dataset is carefully curated, with patients selected based on specific criteria. The dataset comprises various medical predictor variables and a target variable labeled "Outcome." These predictor variables encompass parameters such as the number of pregnancies, BMI, insulin levels, and age. The dataset aims to contribute to understanding the factors contributing to diabetes and improving diagnostic and predictive capabilities in managing the disease.

Experiments have been conducted on the above datasets: One for classification and one for regression. XG boost classifier was used to classify. For regression, linear model and decision tree model were used and Shapley values were compared against accuracy.

13.5 Explaining classification model

13.5.1 Data

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

13.5.2 Content

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in $kg/(heightinm)^2$)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)
- Number of Instances: 768
- Number of Attributes: 8 plus class
- For Each Attribute: (all numeric-valued)
- Number of times pregnant
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skin fold thickness (mm)
- 2-Hour serum insulin (mu U/ml)
- Body mass index ($weightinkg/(heightinm)^2$)
- Diabetes pedigree function
- Age (years)
- Class variable (0 or 1)
- Missing Attribute Values: Yes
- Class Distribution: (class value 1 is interpreted as "tested positive for diabetes")

13.5.3 Domain Knowledge

Commonly known risk factors are high BMI, High glucose, low insulin and age. While other features contribute,, they are known to have indirect effect. Hence, a high Shapley value (positive) is expected.

13.5.4 Statistics

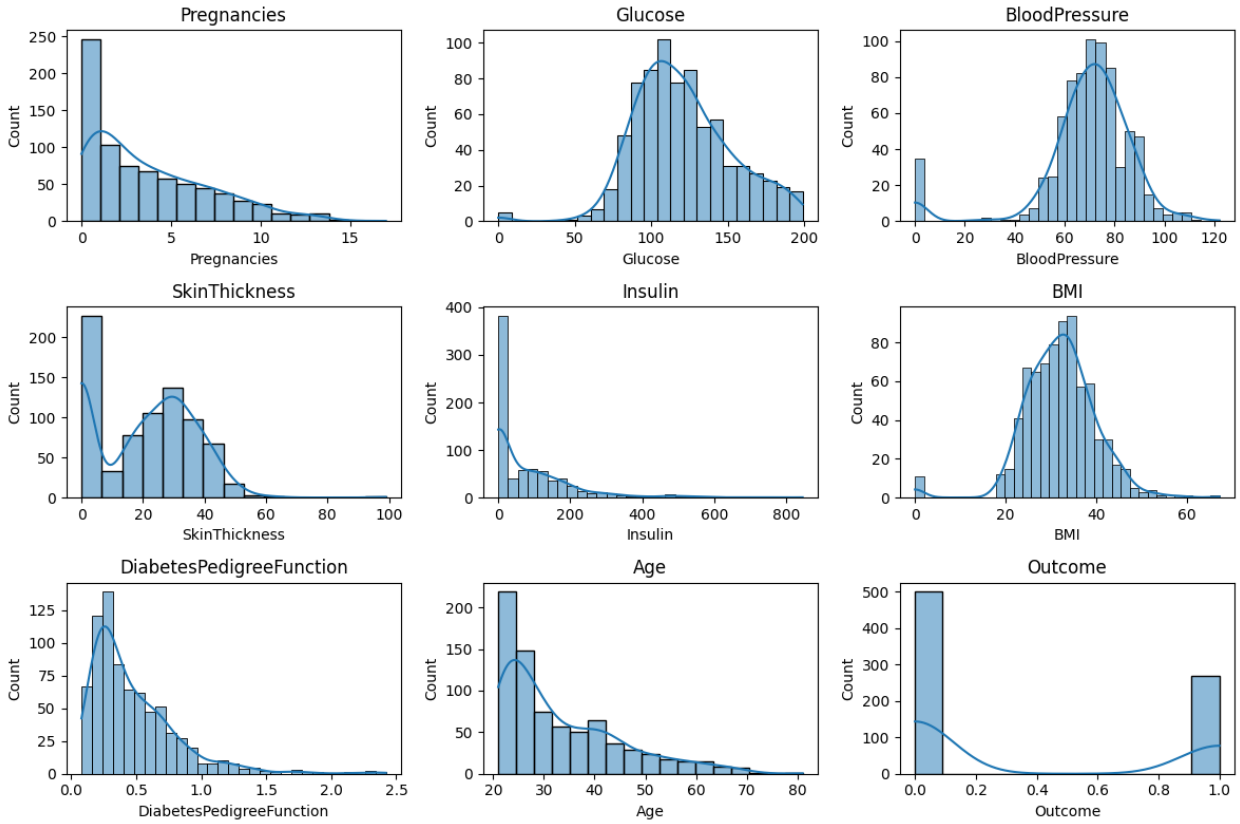


Figure 10: Bar Plot Statistics

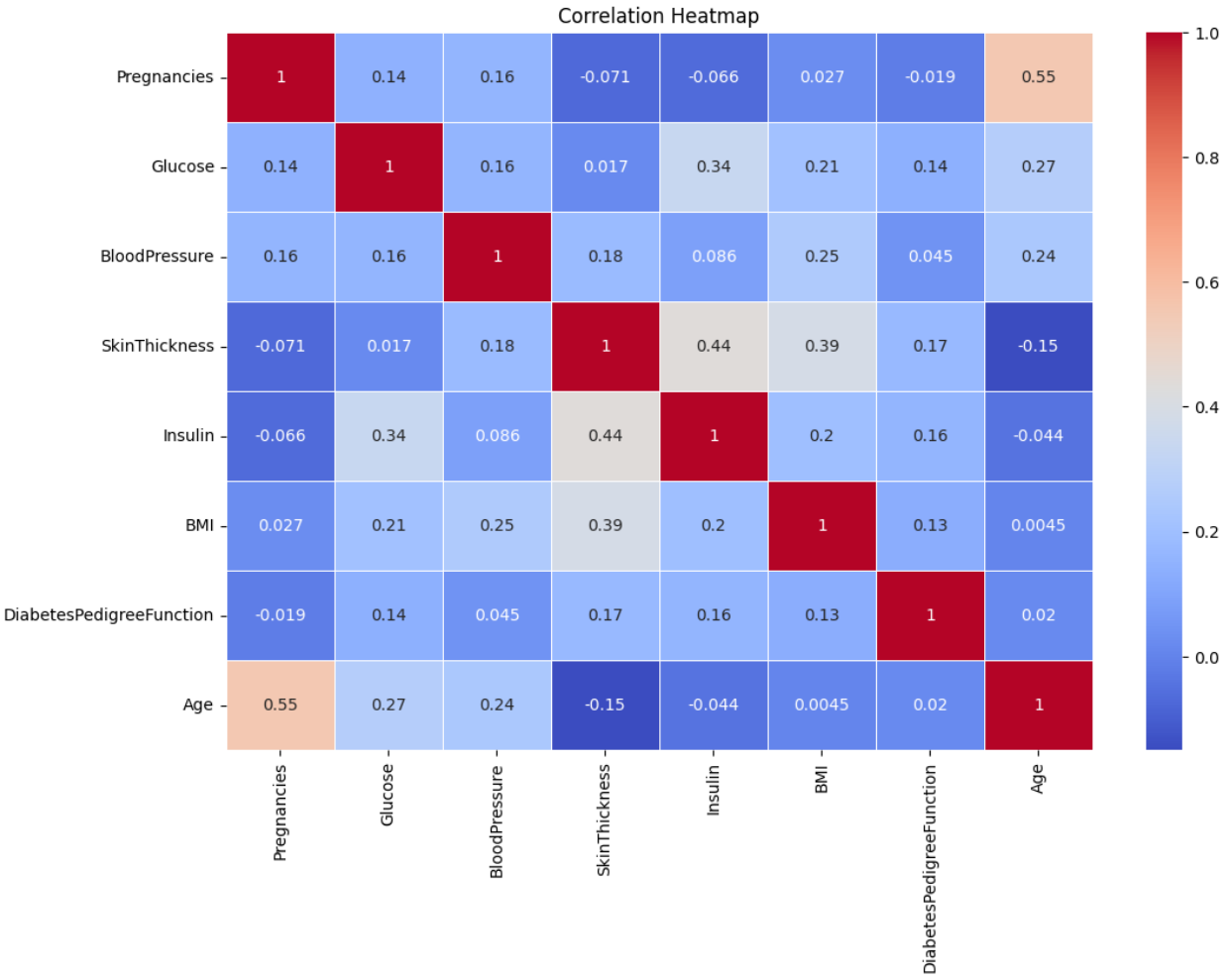


Figure 11: Correlation map for Regression data

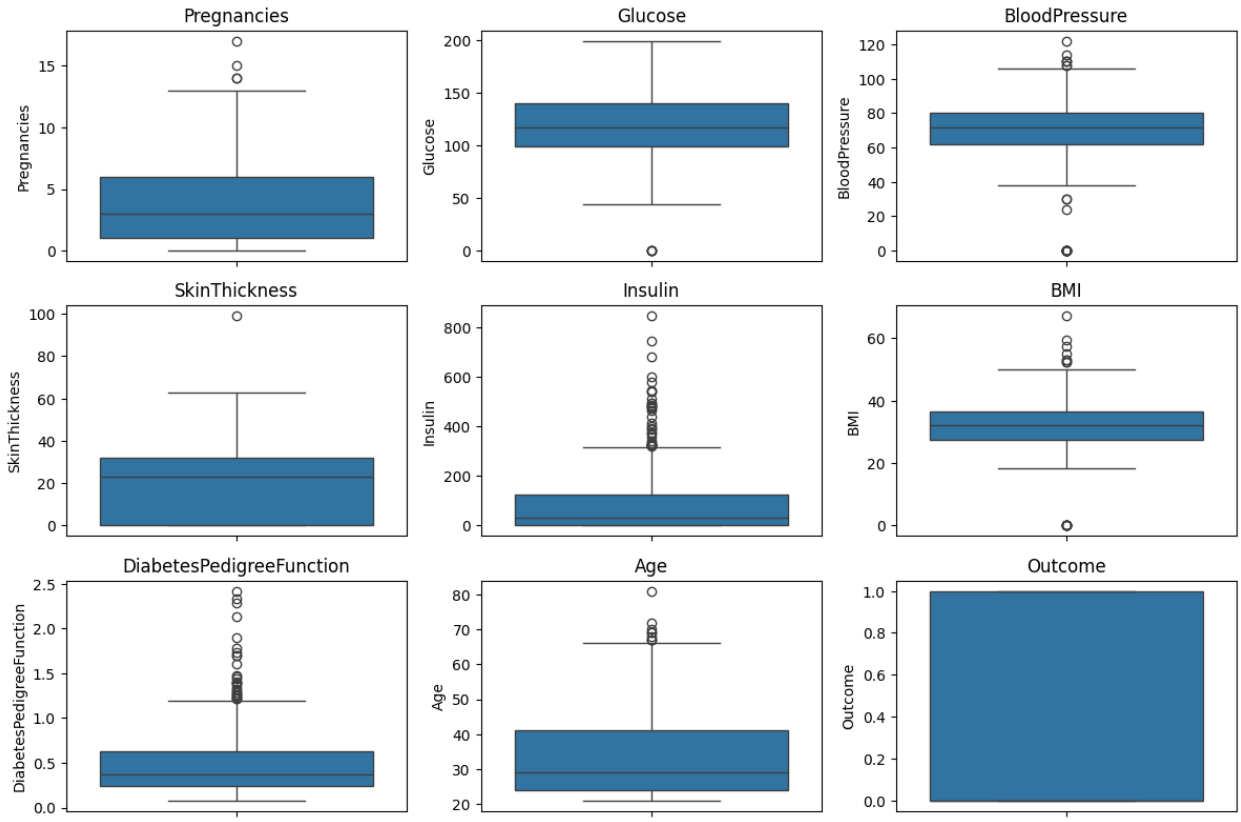


Figure 12: Outliers for classification datasets

13.5.5 Preprocessing

13.5.6 Modified data with feature interaction

Number of features increased from 8 to 44 as:

- Pregnancies
- Glucose
- BloodPressure
- SkinThickness
- Insulin
- BMI
- DiabetesPedigreeFunction
- Age
- Pregnancies^2
- Pregnancies Glucose
- Pregnancies BloodPressure
- Pregnancies SkinThickness

- Pregnancies Insulin
- Pregnancies BMI
- Pregnancies DiabetesPedigreeFunction
- Pregnancies Age
- Glucose^2
- Glucose BloodPressure
- Glucose SkinThickness
- Glucose Insulin
- Glucose BMI
- Glucose DiabetesPedigreeFunction
- Glucose Age
- BloodPressure^2
- BloodPressure SkinThickness
- BloodPressure Insulin
- BloodPressure BMI
- BloodPressure DiabetesPedigreeFunction
- BloodPressure Age
- SkinThickness^2
- SkinThickness Insulin
- SkinThickness BMI
- SkinThickness DiabetesPedigreeFunction
- SkinThickness Age
- Insulin^2
- Insulin BMI
- Insulin DiabetesPedigreeFunction
- Insulin Age
- BMI^2
- BMI DiabetesPedigreeFunction
- BMI Age
- DiabetesPedigreeFunction^2
- DiabetesPedigreeFunction Age
- Age^2

13.5.7 Shap values with and without interaction to explain XG Boost model

Feature	Shapley Value
Pregnancies	0.36055277
Glucose	1.87260587
BloodPressure	0.29626976
SkinThickness	0.40656895
Insulin	0.38597399
BMI	0.12497229
DiabetesPedigreeFunction	0.97530462
Age	1.18491613

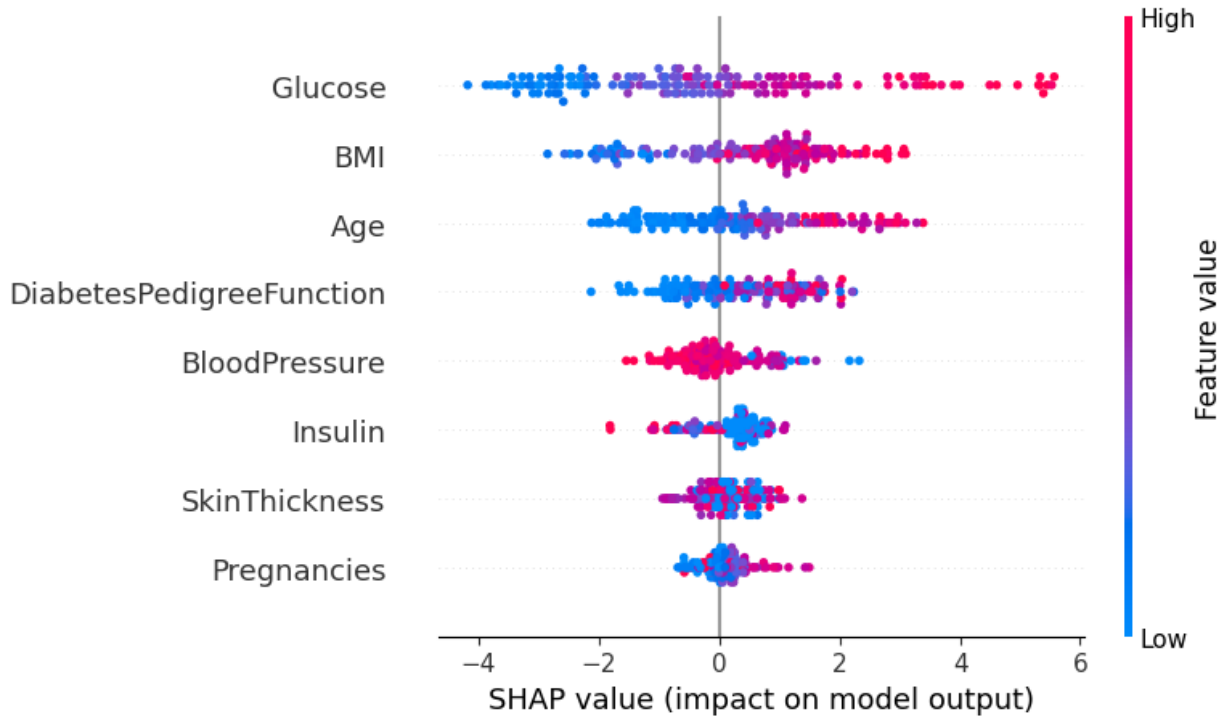


Figure 13: Shap values without interaction

13.5.8 Inferences

Based on the provided Shapley values for each feature in the diabetes dataset, here are some inferences we can make:

Glucose: The highest Shapley value is associated with the "Glucose" feature, indicating that it has the most significant impact on the model's predictions. This suggests that blood glucose levels play a crucial role in determining the likelihood of diabetes.

Age: The second-highest Shapley value corresponds to the "Age" feature, indicating that age also plays a significant role in predicting diabetes. This aligns with common knowledge that older individuals are more prone to developing diabetes.

Diabetes Pedigree Function: The Shapley value for the "DiabetesPedigreeFunction" feature is also relatively high, suggesting that family history of diabetes may contribute significantly to the predictive power of the model.

Insulin: The "Insulin" feature has a moderate Shapley value, indicating that insulin levels also influence the model's predictions to some extent.

Skin Thickness, Blood Pressure, BMI, and Pregnancies: These features have relatively lower Shapley values compared to Glucose, Age, and Diabetes Pedigree Function. However, they still contribute to the model's predictions to varying degrees.

Interactions: It's important to note that the Shapley values provide insights into the individual impact of each feature on the model's predictions. Interactions between features may also influence the predictions, but these are not explicitly captured by the Shapley values for individual features.

From the above figure, we can interpret that a person with high blood sugar and BMI is prone to diabetes than just high blood glucose as seen for non interacting features. Based on these observations, features were transformed to a 2nd degree polynomial to capture interaction between two features and then trained XG boost model. We observed that it gave betetr accuracy than considering individual features. Therefore, Shapley values can be used as a feedback to improve feature engineering and model's outcome.

Table 1: Performance Metrics

	Precision	Recall	F1-Score	Support
Performance on New Data				
Class 0	0.84	0.76	0.80	99
Class 1	0.63	0.75	0.68	55
Accuracy		0.7532		
Performance on Raw Data				
Class 0	0.79	0.78	0.78	99
Class 1	0.61	0.62	0.61	55
Accuracy		0.7208		

The diabetis data for classification (8 features)

Metric	**Value**
Accuracy	0.7078
Precision	0.5806
Recall	0.6545
F1 Score	0.6154

Table 2: Evaluation Metrics

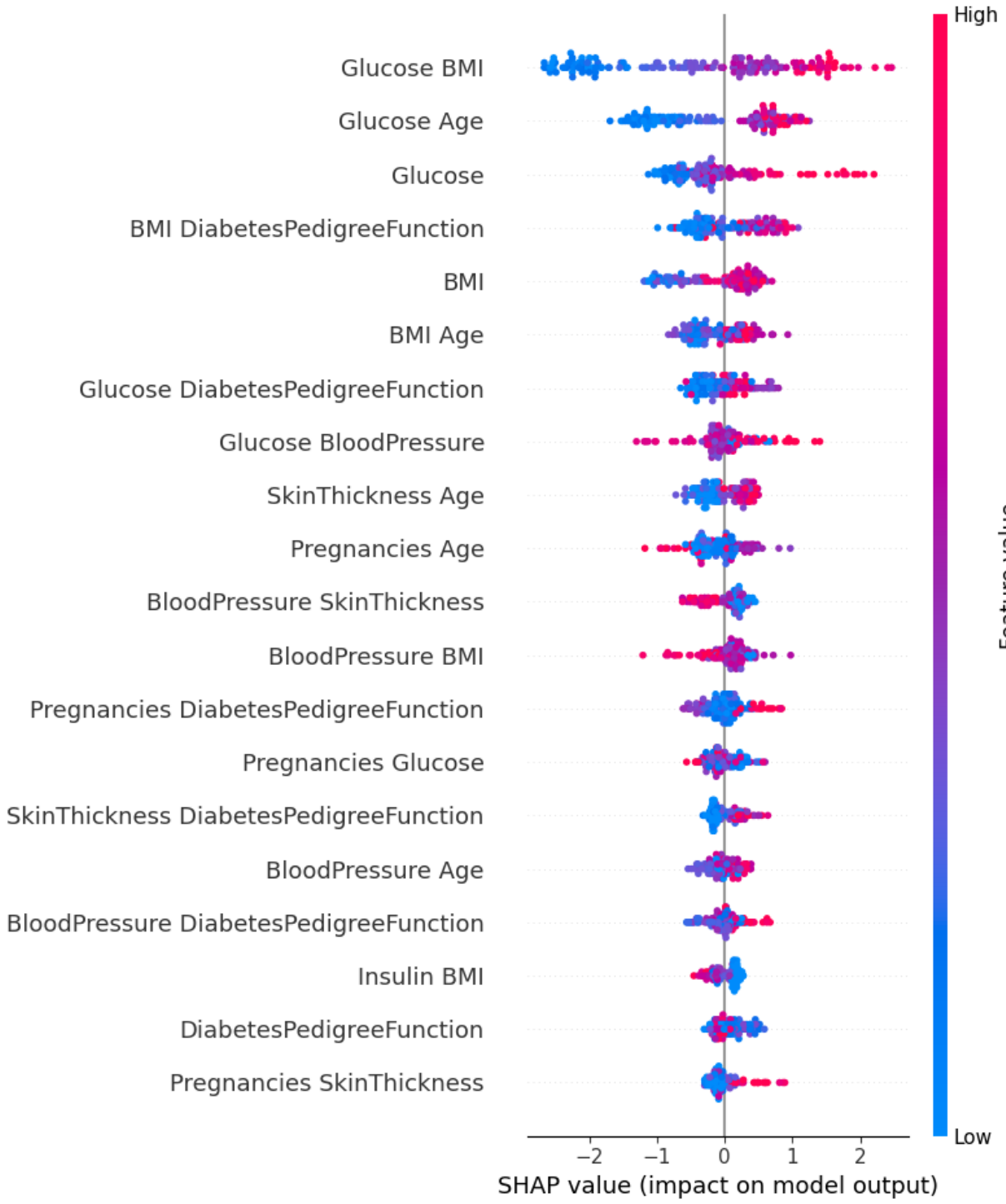


Figure 14: Shap values with interaction

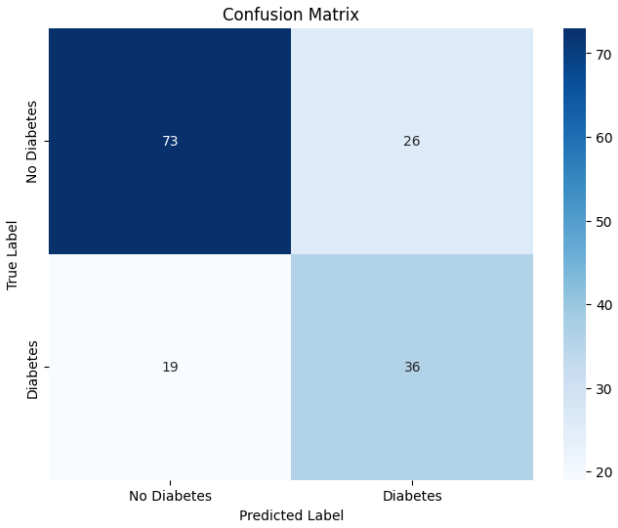


Figure 15: Confusion matrix

Table 3: Feature Importance by XG Boost classifier

Feature	**Importance**
Glucose	0.234045
BMI	0.142424
Age	0.124511
DiabetesPedigreeFunction	0.120475
Pregnancies	0.116584
Insulin	0.098942
SkinThickness	0.092796
BloodPressure	0.070222

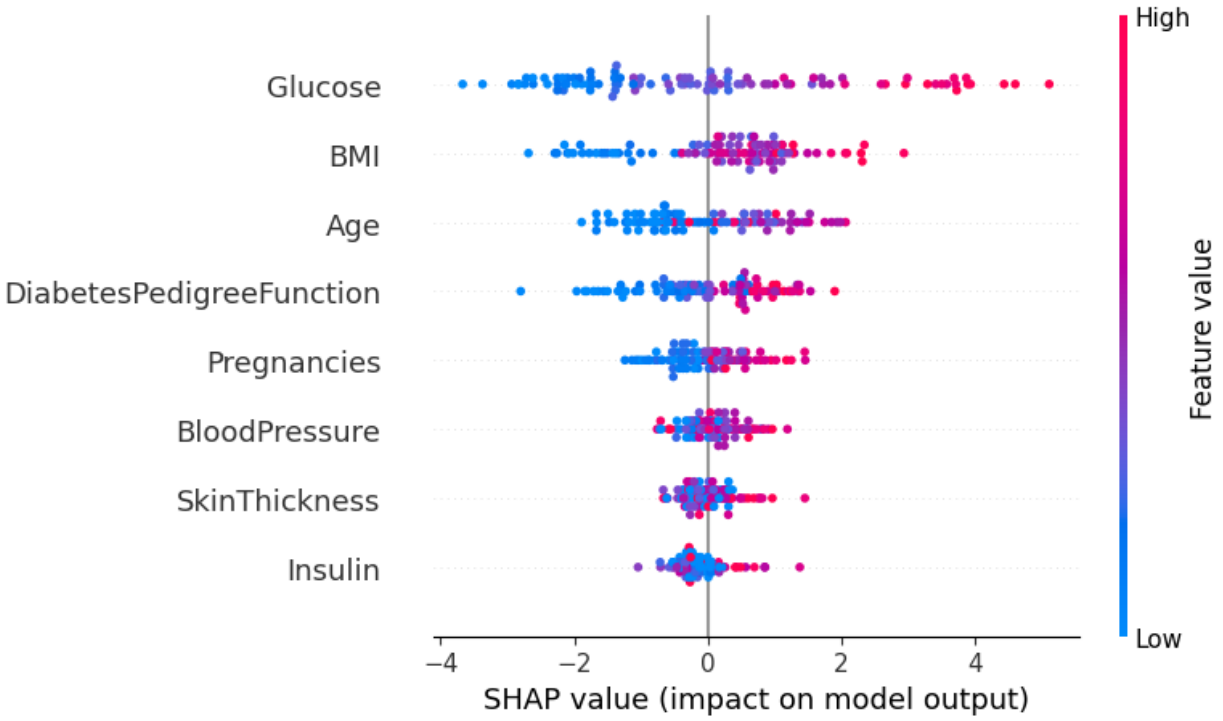


Figure 16: Shap plot for correctly classified data

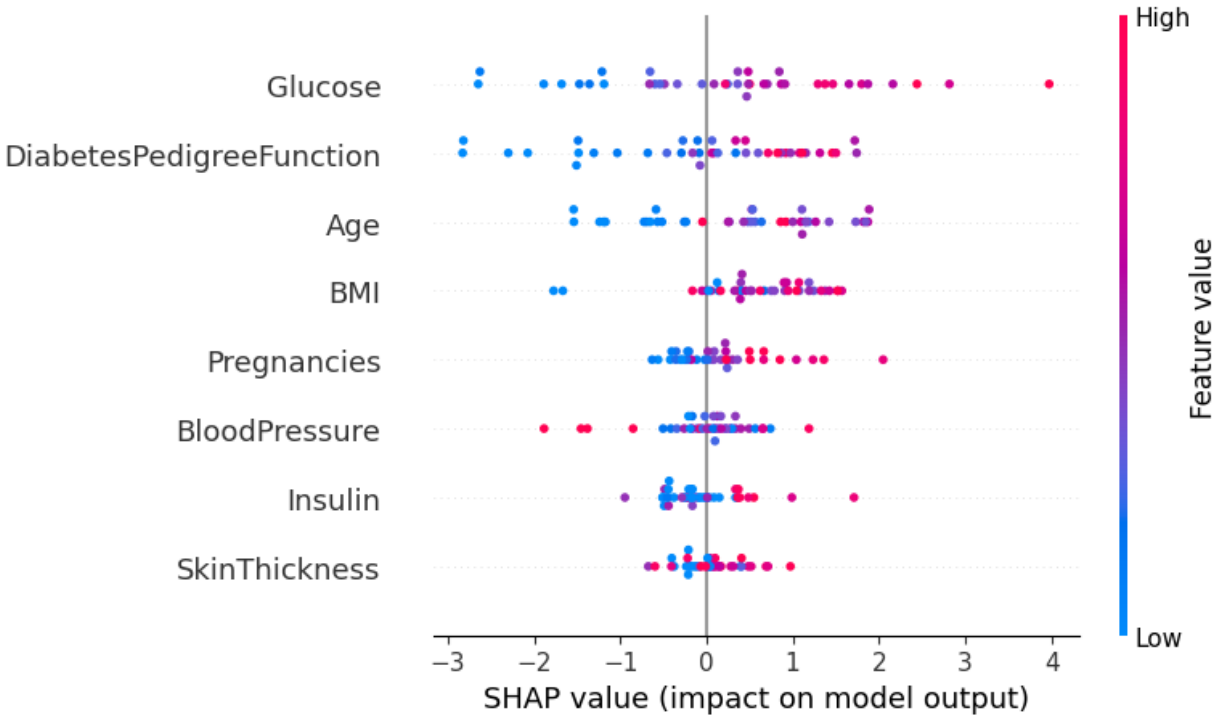


Figure 17: Shap plot for misclassified data

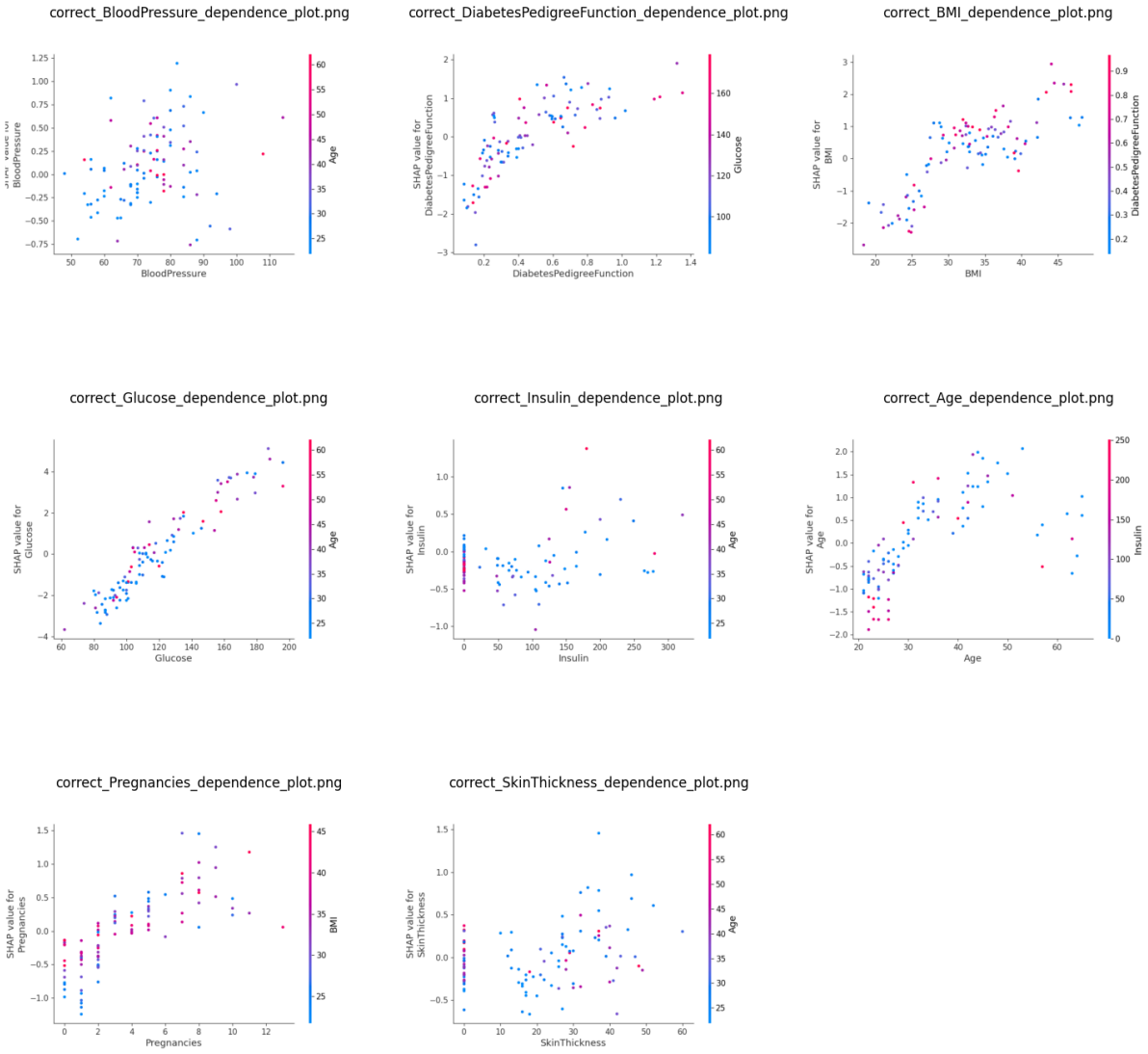


Figure 18: Dependence plots for correctly classified data

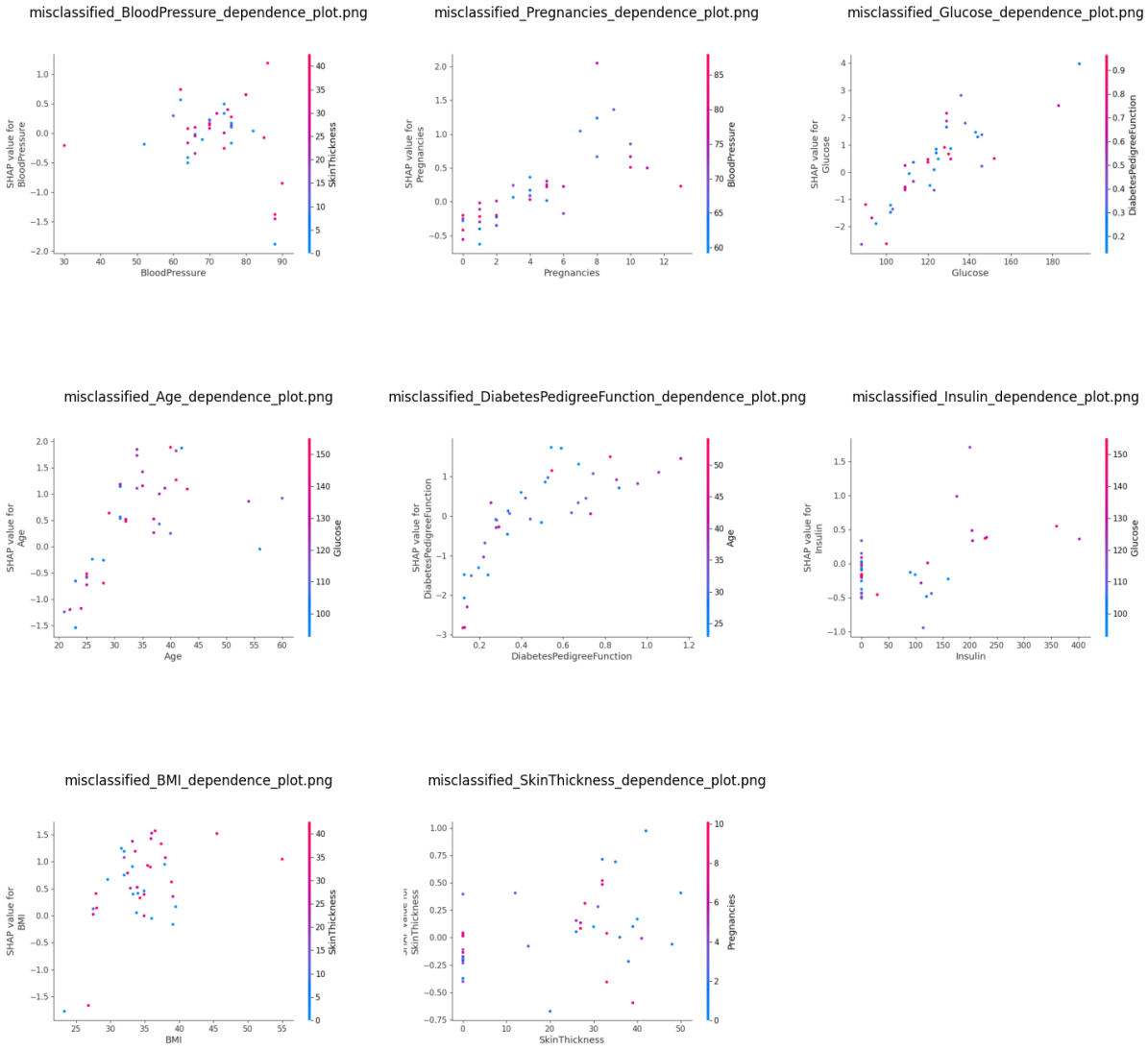


Figure 19: Dependence plots for misclassified data

The dependence plot (And shapley values) gives more importance to BMI to contribute directly to the model output, XG boost classifier considered Glucose as the highest contributer. Also, glucose and age dependence plot is mostly linear, indicating high correlation.

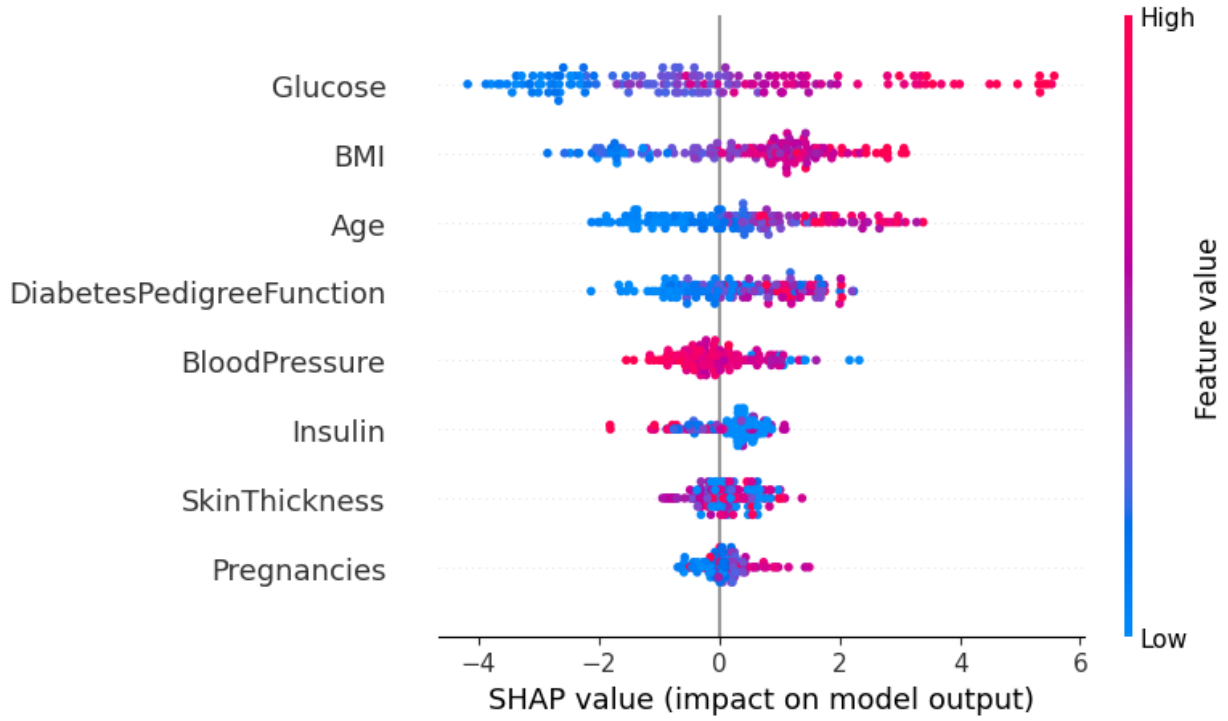


Figure 20: Summary plot of Shapley values

13.6 Diabetes progression

Table 4: Attribute Information

Attribute	**Description**
age	Age in years
sex	Sex (1 = male; 0 = female)
bmi	Body mass index
bp	Average blood pressure
s1	Total serum cholesterol (tc)
s2	Low-density lipoproteins (ldl)
s3	High-density lipoproteins (hdl)
s4	Total cholesterol / HDL (tch)
s5	Possibly log of serum triglycerides level (ltg)
s6	Blood sugar level (glu)

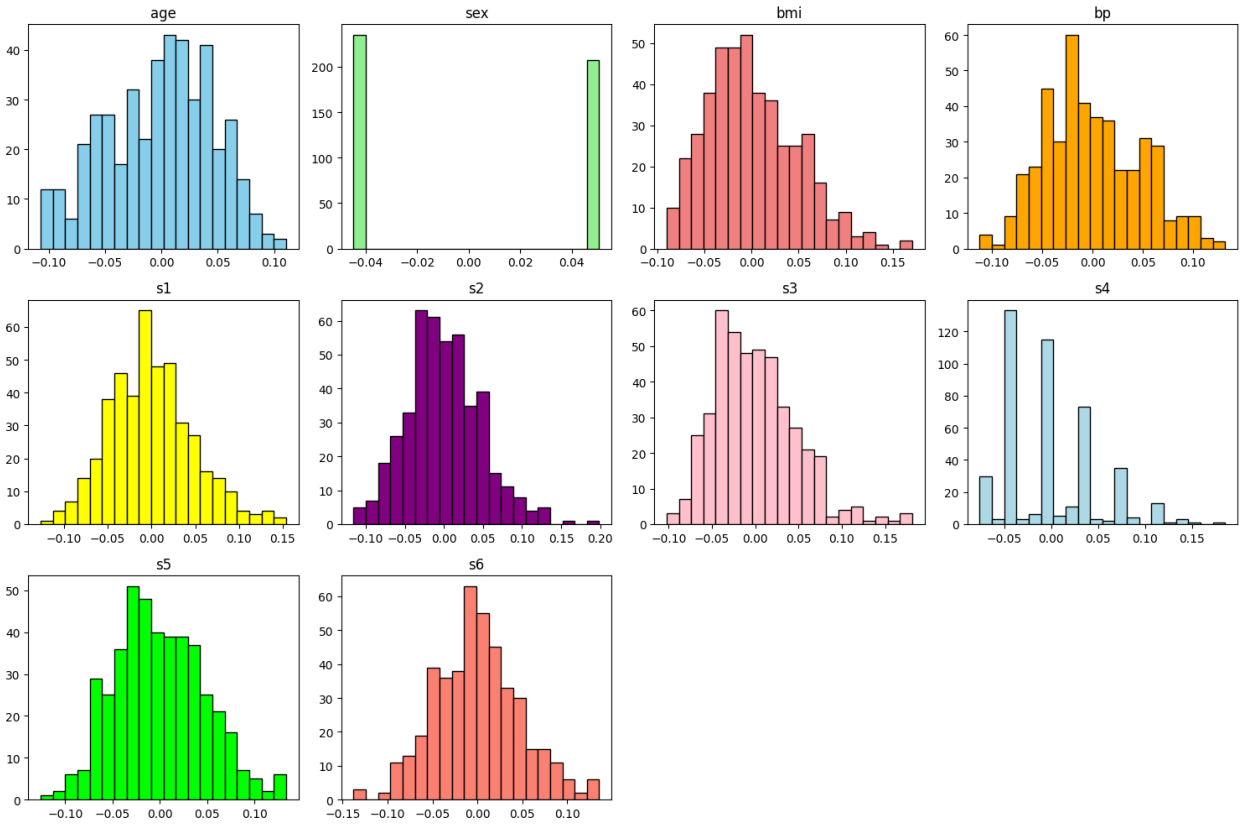


Figure 21: Statistics for regression data

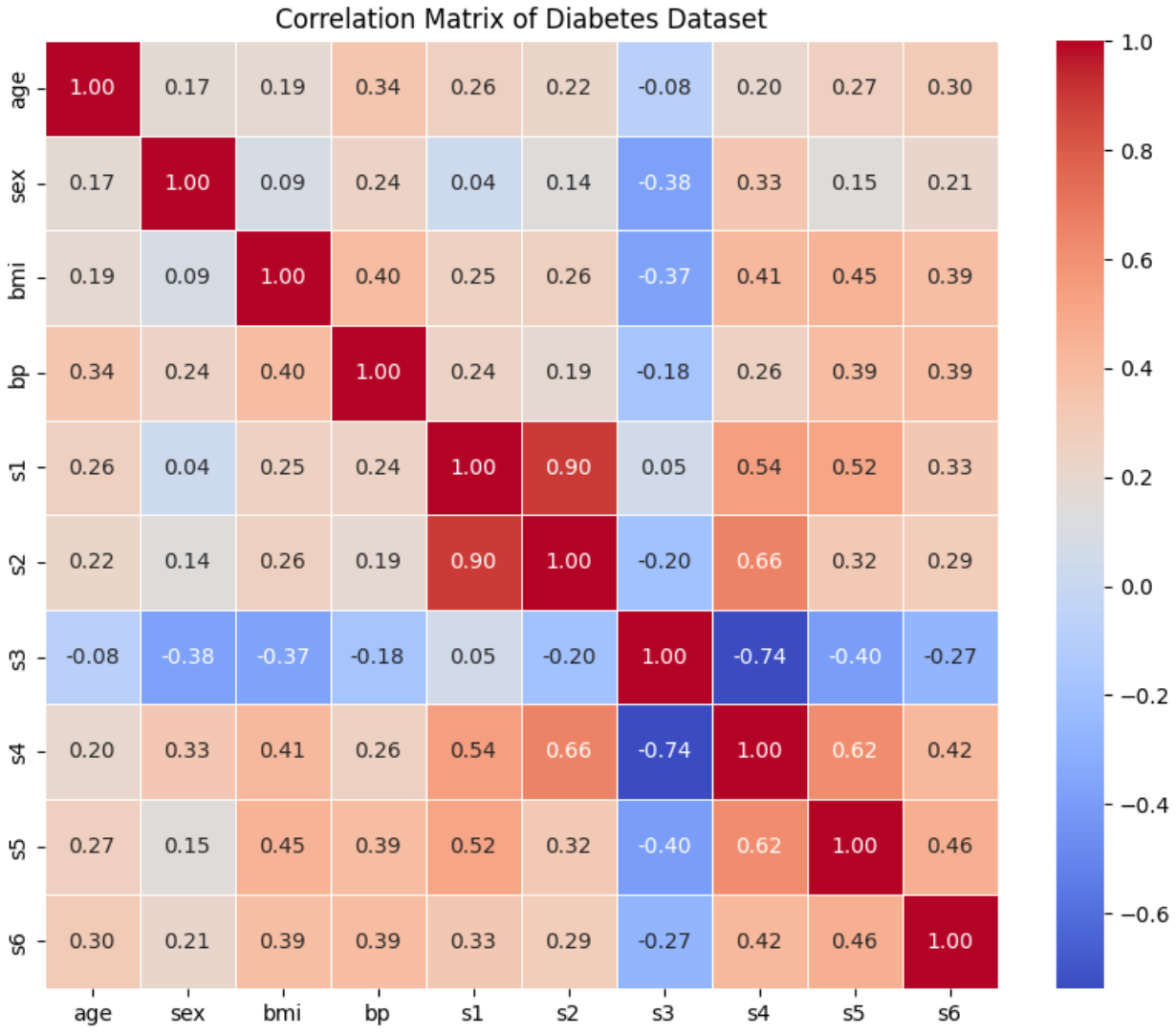


Figure 22: Correlation matrix

13.6.1 Statistics

13.6.2 Shapley values

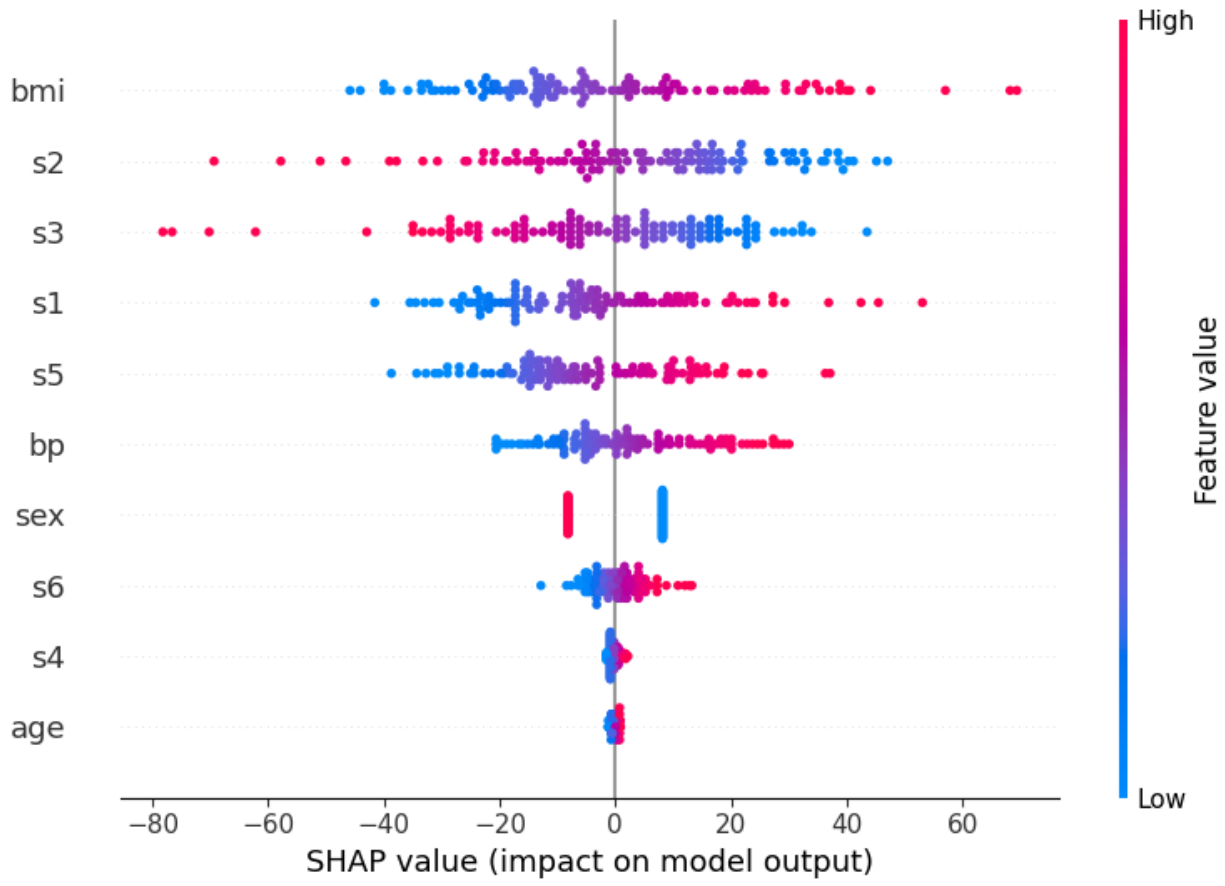


Figure 23: Summaery plot of shap values for linear regression model

Feature	Coefficient
age	11.44
sex	−171.43
bmi	545.70
bp	262.67
s1	367.93
s2	−472.49
s3	−435.02
s4	17.86
s5	378.70
s6	96.94

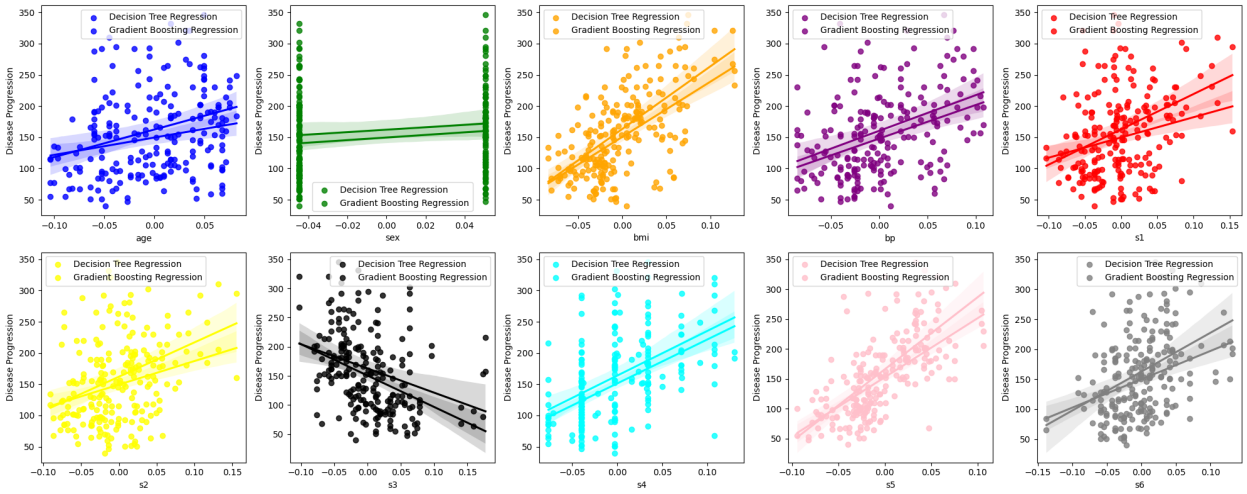


Figure 25: Feature dependance plots for Decision trees anf gradient boosting regression

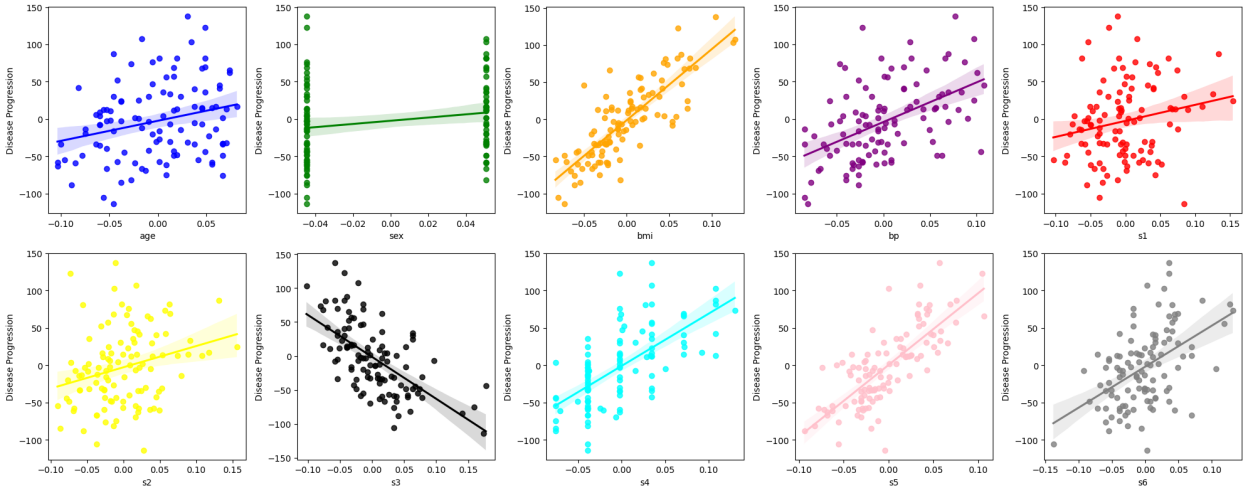


Figure 24: Feature dependance plots for linear regression model

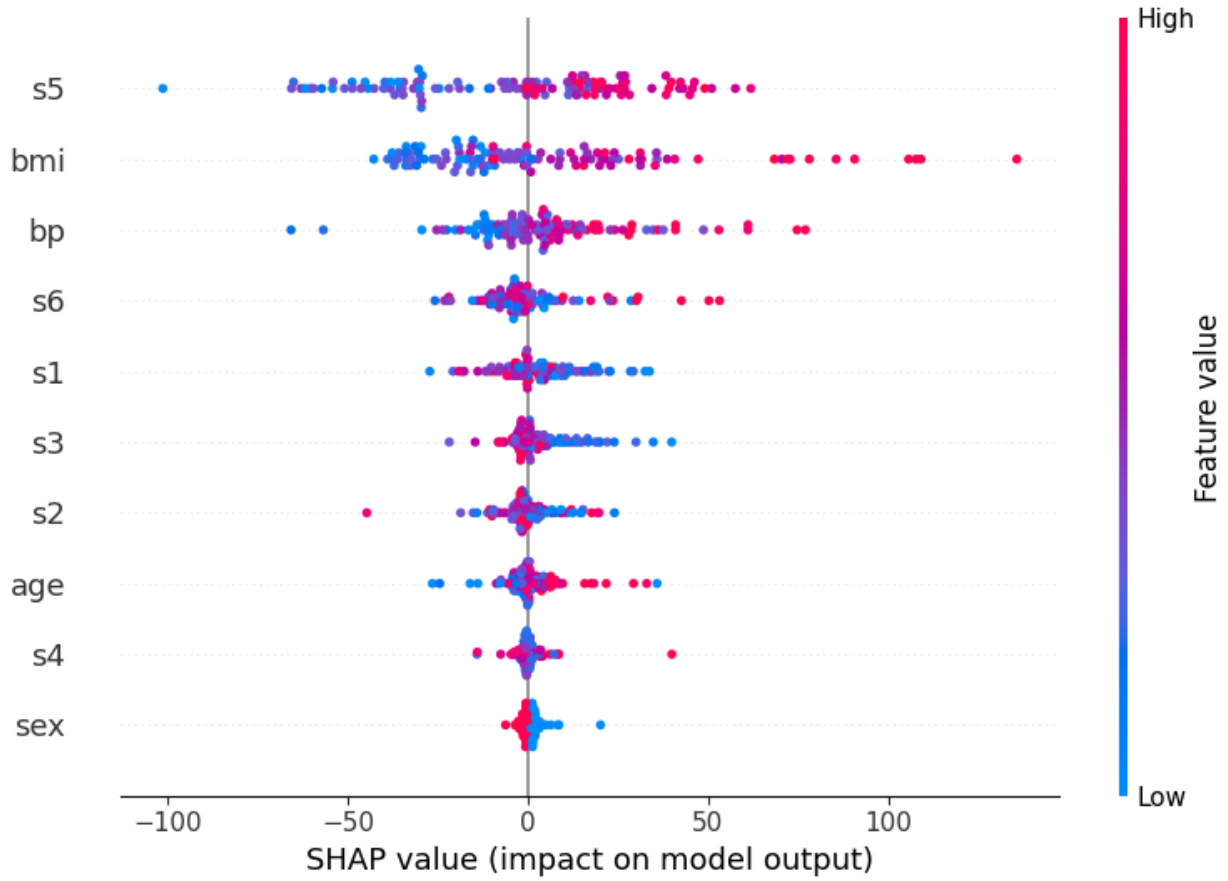


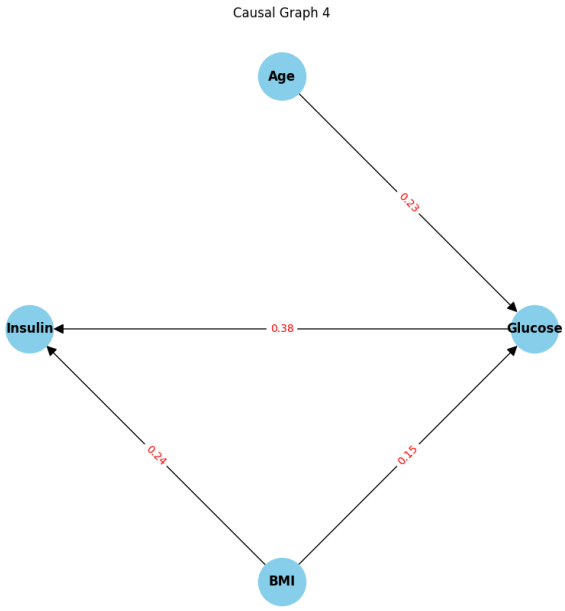
Figure 26: Shap summary plot for decision trees model

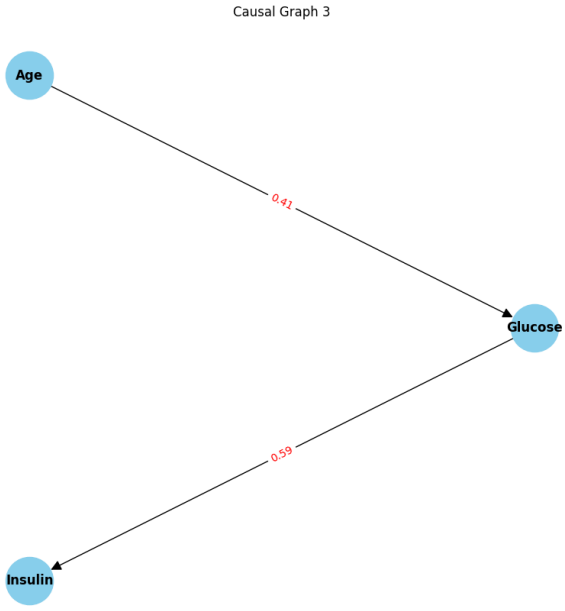
13.6.3 Inferences

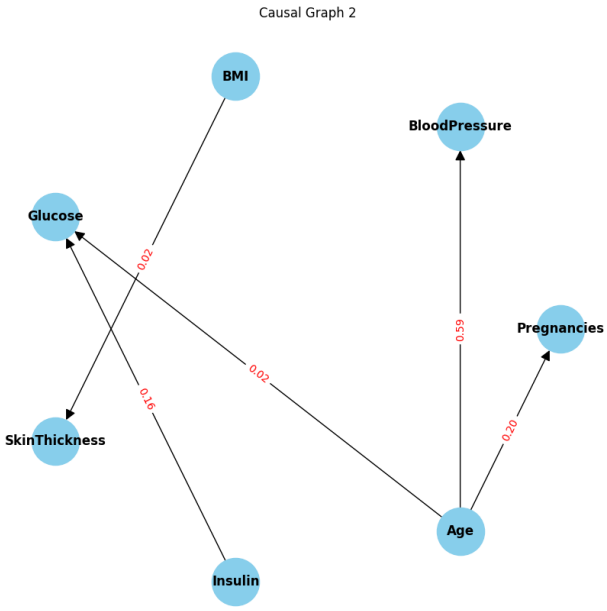
Based on domain knowledge, BMI has more contribution to the diabetes prediction than S5. Hence, linear model can better explain the feature importance, though complex decision tree is able to fit better. Decision Tree Regression Mean Squared Error: 6592.29729 Linear Regression Mean Squared Error: 26239.47

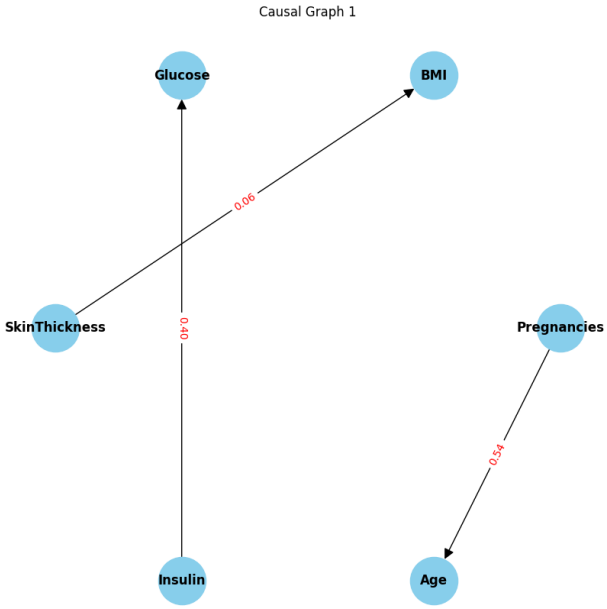
13.7 Causal inference: Ongoing and future work

We tried gaining domain knowledge and generated causal relationships as per Gao, XE., Hu, JG., Chen, B. et al. Causal discovery approach with reinforcement learning for risk factors of type II diabetes mellitus. BMC Bioinformatics 24, 296 (2023). They use an RL based algorithm to obtain causal graphs. Below are some generated graphs:


Causal Graph 4
Age
0.23
Insulin
0.38
Glucose
0.24
0.15
BMI


Causal Graph 3
Age
0.41
Glucose
0.59
Insulin





We wish to extend Shapley and interacted Shapley values to Causal shapley values by using methods like Propensity Score Matching, Directed Acyclic Graph(DAG), Counterfactual methods, Do-Calculus, etc.

References

[1] Christopher Frye, Ilya Feige, and Colin Rowat. Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability. arXiv preprint arXiv:1910.06358, 2019.

[2] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, pages 4765–4774, 2017.

[3] Lloyd S Shapley. A value for n-person games. Contributions to the Theory of Games, 2(28):307– 317, 1953.

[4] Judea Pearl. Causal diagrams for empirical research. Biometrika, 82(4):669–688, 1995. [Link]

[5] Judea Pearl. The do-calculus revisited. arXiv preprint arXiv:1210.4852, 2012.[Link]

[6] Hu, Jie, et al. "Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Machine Learning Models." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 11364-11371. 2020.[Link]

[7] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, Explainable AI: a brief survey on history, research areas, approaches and challenges, Nat. Lang. Process. Chin. Comput. (2019) 563–574,[Link]

[8] C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. Frontiers in Genetics, 2019.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you: Explaining the predictions of any classifier. In International Conference on Knowledge Discovery and Data Mining, 2016.

[10] Chen, H., Lundberg, S. M., Lee, S. I. (2022). Explaining a series of models by propagating Shapley values. Nature communications, 13(1), 4512.

[11] Ali, M. H., Le Biannic, Y., Wuillemin, P. H. (2023, May). Interpreting Predictive Models through Causality: A Query-Driven Methodology. In The International FLAIRS Conference Proceedings (Vol. 36).

[12]Carballo Castro, A. (2022). Explainability and Causality in Machine Learning through Shapley values. Li, M., Sun, H., Huang, Y., Chen, H. (2024).

[13] Shapley value: from cooperative game to explainable artificial intelligence. Autonomous Intelligent Systems, 4(1), 1-12.

[14] Jethani, N., Sudarshan, M., Covert, I. C., Lee, S. I., Ranganath, R. (2021, October). Fastshap: Real-time shapley value estimation. In International Conference on Learning Representations.

[15]Watson, D. (2022, June). Rational shapley values. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1083-1094).