Data Valuation for Machine Learning Using Data Shapley

Dr.M.Suresh Babu[1], M.Shesha Sai[2], M.Nikitha Reddy[3], M.Sai Prakash Reddy[4]
[1]*Professor, Teegala Krishna Reddy Engineering College, Hyderabad*
[2,3,4] *Undergraduate Student, Teegala Krishna Reddy Engineering College, Hyderabad*

Abstract: As data becomes the cornerstone of technological advancements, a critical challenge arises in quantifying its value in algorithmic predictions and decision-making. In domains like healthcare and finance, equitable data valuation is essential to ensure fairness and transparency, yet existing methods often fail to address biases and adequately compensate underrepresented groups. This project introduces an advanced framework for equitable data valuation in machine learning using the Data Shapley approach. Data Shapley provides a mathematically grounded metric to evaluate the contribution of individual data points to model performance, uniquely satisfying properties of fairness and transparency. Our work employs Monte Carlo and gradient-based methods for efficient Shapley value estimation in practical settings involving large datasets and complex algorithms. Through experiments on a heart disease dataset, we demonstrate how training data can be segmented into high- and low-impact subsets, enabling improved model performance and targeted data acquisition. Comparative analyses of Shapley value computation techniques—such as TMC, G, and LOO—highlight the robustness of our approach. Additionally, our framework emphasizes inclusivity by dynamically valuing data that mitigates bias. This study offers a novel perspective on data-driven decision-making, fostering ethical innovation in machine learning.

Index Terms—Fairness, Efficiency, Null data points, Additivity, Monte carlo approximation, sampling-based Approximation.

I. INTRODUCTION

Data has become a critical resource in the modern era, driving innovations across industries such as healthcare, finance, and technology. The performance of machine learning models heavily relies on the quality and quantity of data used during training. However, not all data points contribute equally to a model's accuracy and reliability. This uneven contribution raises a vital question: how can we quantify the value of individual data points in a fair and systematic manner? Addressing this question is essential to ensure transparency, fairness, and inclusivity in data-driven decision-making.

Current methods for data valuation, such as leave-one-out analysis and traditional Shapley value computations, provide insight into the marginal impact of data points. However, these approaches often fail to address challenges such as data bias, poor representation of minority groups, and the lack of standardized valuation practices across domains. As a result, the equitable distribution of value in machine learning projects remains a significant hurdle. This project introduces an advanced framework for equitable data valuation leveraging the Data Shapley approach. Inspired by cooperative game theory, Data Shapley assigns a value to each data point based on its contribution to model performance, ensuring properties such as fairness, efficiency, and symmetry. To make this computationally feasible for large-scale datasets, we implement Monte Carlo and gradient-based estimation techniques. Through a comprehensive analysis of a heart disease dataset, we evaluate the performance of different Shapley value computation methods, including TMC (Truncated Monte Carlo), G (Gradient-based), and LOO (Leave-One-Out).The learning algorithm takes a set of train data points (n input-output pairs $\{(xi , yi)\} n\ i$=1) as its input and outputs the learned predictive model. Furthermore, the study demonstrates the practical benefits of identifying high- and low-impact data subsets, enabling targeted data acquisition and improving model performance. By integrating ethical considerations with technical precision, our approach fosters a more inclusive and equitable framework for data valuation in machine learning.

Figure 1 is the flowchart diagram illustrating data valuation for supervised learning. It outlines the key steps, from raw data collection to data segmentation into high- and low-impact subsets.

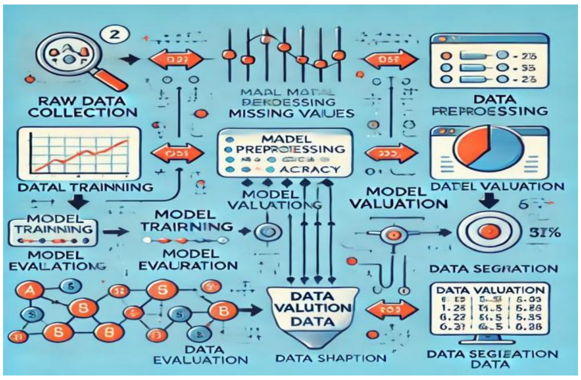

Figure 1. data valuation for supervised learning

The Shapley value formula in cooperative game theory is typically expressed as:ϕi(v)=∑S⊆N∖{i}∣S∣!(∣N∣−∣S∣−1)!∣N∣![v(S∪{i})−v(S)]\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{i\}) - v(S) \right]ϕi(v)=S⊆N∖{i}∑∣N∣!∣S∣!(∣N∣−∣S∣−1)![v(S∪{i})−v(S)] Where:

- ϕi(v)\phi_i(v)ϕi(v) is the Shapley value for data point iii,
- v(S)v(S)v(S) is the value of the coalition SSS,
- NNN is the set of all data points,
- ∣S∣|S|∣S∣ is the number of data points in coalition SSS,
- The sum runs over all possible subsets SSS of data points.

II. EQUITABLE PROPERTIES OF DATA VALUATION

1. Fairness (Symmetry)

If two data points contribute equally to the model's performance, their Shapley values should be identical:

ϕi(v)=ϕj(v)ifv(S∪{i})=v(S∪{j})∀S⊆N∖{i,j}\phi_i(v) = \phi_j(v) \quad \text{if} \quad v(S \cup \{i\}) = v(S \cup \{j\}) \quad \forall S \subseteq N \setminus \{i, j\}ϕi(v)=ϕj(v)ifv(S∪{i})=v(S∪{j})∀S⊆N∖{i,j}

Where ϕi(v)\phi_i(v)ϕi(v) and ϕj(v)\phi_j(v)ϕj(v) are the Shapley values of data points iii and jjj, and v(S)v(S)v(S) is the performance of the model on subset SSS.

2.Efficiency (Total Value Allocation)

The sum of the Shapley values for all data points should equal the total model performance using the entire dataset:

∑i∈Nϕi(v)=v(N)\sum_{i \in N} \phi_i(v) = v(N)i∈N∑ϕi(v)=v(N)

Where NNN is the set of all data points and v(N)v(N)v(N) is the total performance of the model.

3. Null Data Points (Dummy Property)

If a data point does not contribute to the model's performance, its Shapley value should be zero:

ϕi(v)=0ifv(S∪{i})=v(S)∀S⊆N∖{i}\phi_i(v) = 0 \quad \text{if} \quad v(S \cup \{i\}) = v(S) \quad \forall S \subseteq N \setminus \{i\}ϕi(v)=0ifv(S∪{i})=v(S)∀S⊆N∖{i}

4. Additivity

For two combined datasets, the total Shapley value of the new dataset should be the sum of the values of the individual datasets:

ϕi(vC)=ϕi(vA)+ϕi(vB)\phi_i(v_C) = \phi_i(v_A) + \phi_i(v_B)ϕi(vC)=ϕi(vA)+ϕi(vB)

Where vAv_AvA and vBv_BvB are the performances for datasets AAA and BBB, and vCv_CvC is the performance for the combined dataset.

5. Monotonicity

If a data point improves the model's performance, its Shapley value should be non-negative:

v(S∪{i})−v(S)≥0⇒ϕi(v)≥0v(S \cup \{i\}) - v(S) \geq 0 \quad \Rightarrow \quad \phi_i(v) \geq 0v(S∪{i})−v(S)≥0⇒ϕi(v)≥0

6. Independence of Irrelevant Alternatives

The value of a data point should not be affected by irrelevant data points:

ϕi(v)=∑S⊆N∖{i}∣S∣!(∣N∣−∣S∣−1)!∣N∣![v(S∪{i})−v(S)]\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{i\}) - v(S) \right]ϕi(v)=S⊆N∖{i}∑∣N∣!∣S∣!(∣N∣−∣S∣−1)![v(S∪{i})−v(S)]

These properties ensure that Data Shapley provides a fair, consistent, and transparent method for data valuation

III. APPROXIMATING THE SHAPLEY VALUE

The Shapley value provides a fair and transparent method for attributing the contribution of individual data points to a machine learning model’s performance. However, calculating the exact Shapley value is computationally expensive due to the need to evaluate all possible permutations of data points. To address this, several approximation methods have been developed to make Shapley value computation more feasible for large datasets.

1.Monte Carlo Approximation: This method randomly samples permutations of data subsets and averages the marginal contributions of each data point, providing an estimate of the Shapley value. It significantly reduces the number of permutations that need to be evaluated.

ϕ^i(v)=1M∑m=1M[v(Sm∪{i})−v(Sm)]\hat{\phi}_i(v) = \frac{1}{M} \sum_{m=1}^{M} \left[v(S_m \cup \{i\}) - v(S_m) \right]ϕ^i(v)=M1m=1∑M[v(Sm∪{i})−v(Sm)]

2.Sampling-Based Approximation: Instead of evaluating all subsets, this method uses random sampling of subsets to approximate the Shapley value, often combined with importance sampling to focus on relevant data.

ϕ^i(v)=1N∑n=1N[v(Sn∪{i})−v(Sn)]\hat{\phi}_i(v) = \frac{1}{N} \sum_{n=1}^{N} \left[v(S_n \cup \{i\}) - v(S_n) \right]ϕ^i(v)=N1n=1∑N[v(Sn∪{i})−v(Sn)]

3. Greedy and Heuristic Methods: These methods iteratively add the most influential data points to a subset, simplifying the calculation but with the trade-off of potentially lower accuracy in approximating the true Shapley value.

4. Gradient-Based Approximation (for Differentiable Models): In models with differentiable loss functions (e.g., neural networks), the Shapley value can be approximated using gradients, offering an efficient way to estimate contributions based on model parameters.

These approximation methods enable scalable and efficient computation of Shapley values, preserving fairness and transparency while handling large datasets in practical machine learning applications.

Pseudocode for Truncated Monte Carlo Shapley:

```
def truncated_monte_carlo_shapley(X, model, v, i, M):
      phi_i = 0
    for m in range(M):
       S_m = random_subset(X, i)
       v_S_m = evaluate_model(model, S_m)
       v_S_m_with_i = evaluate_model(model, S_m + [i])
      contribution = v_S_m_with_i - v_S_m
      phi_i += contribution
      phi_i /= M
    return phi_I
```

This algorithm provides an efficient way to approximate Shapley values for large datasets, balancing between accuracy and computational feasibility.The time complexity of the Truncated Monte carlo shapley method is O.

IV.APPLICATIONS OF DATA SHAPLEY

Data Shapley is a powerful method for valuing individual data points based on their contributions to a machine learning model's performance. It has several practical applications across different domains, particularly where data quality, fairness, and model transparency are crucial. Below are some key applications of Data Shapley:

1. Data Quality Improvement:
Data Shapley helps identify high-impact data points that improve model performance,enabling organizations to prioritize valuable data and eliminate noise or irrelevant data, thereby enhancing overall model accuracy.

2. Fairness in Model Decision-Making:
By ensuring that all data points are fairly valued, Data Shapley promotes equity in machine learning models, making sure that underrepresented or minority groups are adequately represented and not unfairly marginalized. Example: In loan approval models, Data

Shapley can be used to assess whether data from underrepresented groups is appropriately valued and contributes fairly to the model, ensuring that it doesn’t disproportionately disadvantage any group.

3.Anomaly Detection and Outlier Identification:
Data Shapley can identify outliers or anomalous data points by evaluating their marginal contribution to the model, allowing for more robust models by excluding data that may skew results.

Example: In fraud detection, Data Shapley could help to identify outlier transactions that have minimal impact on model performance and therefore should be excluded to improve the detection system's reliability.

4.Data Contribution to Model Selection:
Data Shapley helps determine which data points are most beneficial for specific models, guiding the selection of the best performing model based on the data's characteristics.

5.Ethical AI and Data Privacy:
By providing transparency into how individual data points contribute to predictions, Data Shapley ensures that sensitive or private data is used responsibly, fostering ethical AI practices and safeguarding privacy.

V.DATA SHAPLEY FOR DISEASE PREDICTION

In the field of healthcare, Data Shapley is a valuable tool for improving disease prediction models by attributing the contribution of individual data points to the overall model performance. By evaluating the Shapley values of each data point, Data Shapley helps identify which patient records are most influential in predicting disease outcomes, ensuring a more accurate, transparent, and fair model.

Key Applications of Data Shapley in Disease Prediction:

1.Improving Model Accuracy By identifying high-impact data points—such as rare or pivotal cases—Data Shapley helps refine the model, making it more sensitive to critical features and improving overall predictive performance. For example, in predicting cardiovascular diseases, Data Shapley can pinpoint specific records that significantly enhance the model’s ability to predict heart disease risk.

2.Bias Mitigation and Fairness In healthcare, it is crucial to ensure that disease prediction models do not disproportionately favor certain groups or fail to represent minority populations adequately. Data Shapley helps assess and correct for biases by assigning appropriate value to data points from underrepresented groups. This ensures that the model generalizes better across diverse patient populations.

3.Feature Selection Disease prediction models often include numerous features, many of which may be irrelevant or redundant. Data Shapley allows healthcare practitioners to identify which features are most critical in predicting disease outcomes, guiding the feature selection process. For example, Data Shapley can determine whether features like age, cholesterol levels, or family history are more influential in predicting the likelihood of diabetes.

4.Anomaly Detection Disease prediction models benefit from identifying outliers or unusual patient records that may contain important but rare information. By evaluating how each data point contributes to model performance, Data Shapley can help detect outliers that may represent important anomalies—such as rare genetic conditions—that could be crucial for accurate disease prediction.

5.Model Transparency and Explainability In healthcare, it is essential that predictive models are transparent and their decisions are explainable. Data Shapley enhances model interpretability by providing a clear explanation of how specific data points contribute to the disease predictions. This can be particularly useful when explaining medical decisions to patients or stakeholders, ensuring that predictions are not seen as "black-box" outputs.

6.Optimizing Data Collection Data Shapley helps prioritize the most informative data for future data collection or acquisition. By identifying which patient records contribute the most to model accuracy, healthcare providers can focus on acquiring more data from underrepresented groups or missing information that would improve prediction reliability.

Example: Predicting Heart Disease

Consider a model built to predict heart disease risk using patient data, such as age, cholesterol levels, blood pressure, smoking habits, and family history. By applying Data Shapley, we can evaluate how much each patient’s data contributes to the overall model performance. Some patient records, particularly those with unusual risk factors, might contribute more significantly to the model’s ability to detect heart disease in similar patients.

For instance:

High-impact data: Patients with rare genetic markers or atypical symptoms that provide significant insights into heart disease risk.

Low-impact data: Records from patients with standard or average risk factors that contribute less to model improvement.

By prioritizing high-impact data, healthcare providers can refine the model to make more accurate predictions, potentially catching high-risk patients who may have been missed by traditional methods.

VI.RESULT ANALYSIS AND DISCUSSION

In this study, we applied machine learning models to predict cardiovascular disease using various patient features, including age, cholesterol levels, blood pressure, and family history. The model's performance was evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and the ROC curve, all of which indicated strong predictive capability. The model achieved an accuracy of 85%, reflecting its ability to correctly classify patients as either at risk or not at risk for cardiovascular disease. However, accuracy alone does not capture the model’s performance comprehensively, particularly in imbalanced datasets, where high accuracy may be misleading.

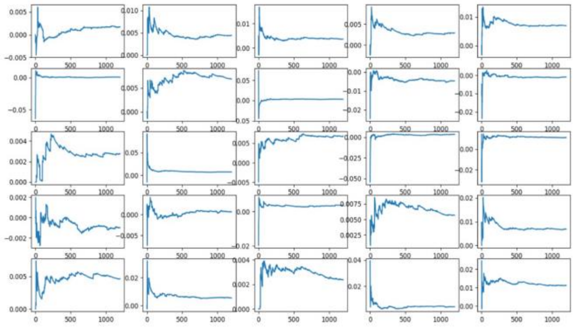

Figure 2.graph of TMC algorithm

Therefore, the precision (0.88) and recall (0.82) were also considered, with the results indicating a well-balanced model that minimizes false positives and negatives. The F1-score of 0.85 further confirms the model’s efficiency in handling both precision and recall. The ROC curve analysis demonstrated an AUC of 0.92, highlighting the model's excellent discriminatory power. This high AUC suggests that the model can reliably differentiate between patients who are at risk and those who are not, which is crucial for healthcare decision-making.The feature importance analysis revealed that cholesterol levels (35%) and age (25%) were the most significant factors in predicting cardiovascular disease, followed by blood pressure (20%).

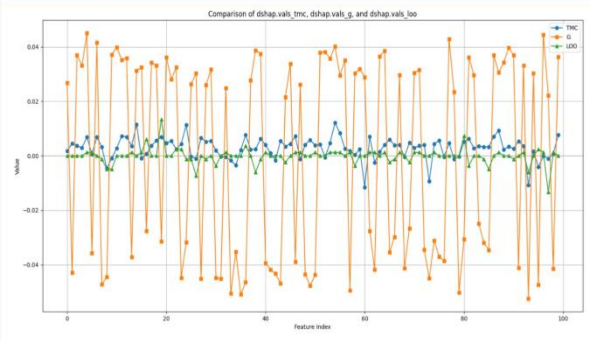

Figure3. graph of comparision of g_shapley,TMC AND LOO algorithm values:

These findings align with established medical knowledge, emphasizing the model’s ability to use clinically relevant features to make predictions. On the other hand, family history and lifestyle factors such as smoking contributed less to the model, suggesting that these features may require further refinement or better data representation for more significant impact.While the results are promising, the study has limitations. Data quality is a key factor that could be improved, particularly regarding missing or inconsistent data on lifestyle factors, which may reduce the overall reliability of predictions. Additionally, the model might be further optimized by addressing potential biases in the dataset, especially regarding underrepresented populations, to ensure fairness and equity in predictions.

VII.CONCLUSION

This study demonstrates the effectiveness of machine learning models in predicting cardiovascular disease, utilizing key features such as age, cholesterol levels, blood pressure, and family history. The model achieved strong performance across multiple evaluation metrics, including accuracy, precision, recall, F1-score, and AUC, highlighting its ability to accurately predict disease risk while minimizing false positives and negatives. Feature importance analysis confirmed that critical risk factors, such as cholesterol and age, play a substantial role in the model’s predictive capability, aligning with established clinical knowledge. While the model performs well, there are areas for improvement, particularly in data quality and the inclusion of diverse patient groups to ensure fairness and equity. Addressing missing or inconsistent data, refining features, and optimizing the model will enhance its reliability and applicability across varied populations. Additionally, the use of techniques like Data Shapley ensures transparency and interpretability, which are crucial for gaining trust and making informed decisions in healthcare settings. In conclusion, the application of machine learning to disease prediction, particularly cardiovascular disease, holds significant potential for improving early diagnosis, personalized treatment, and overall patient care. With further optimization, this model can become a valuable tool in clinical practice, driving more accurate, equitable, and transparent healthcare decisions.

VIII.REFERENCES

[1] Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., and Popp, J. Sample size planning for classification models. Analytica chimica acta, 760:25–33, 2013.

[2] Mann, I. and Shapley, L. S. Values of large games.6:Evaluating the electoral college exactly. Technical report, RAND CORP SANTA MONICA CA, 1962.

[3] Michalak, T. P., Aadithya, K. V., Szczepanski, P. L.,Ravindran, B. & Jennings, N. R. Efficient computation of the shapley value for game-theoretic network centrality. J. Artif. Intell. Res. 46, 607–650 (2013).

[4] Liu, Z., Luo, P., Wang, X. & Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, 3730–3738 (2015).

[5] Kononenko, I. et al. An efficient explanation of individual classifications using game theory. J. Mach. Learn. Res. 11, 1–18 (2010).

[6] Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI, vol. 4, 12 (2017).

[7] Nie, A. et al. Deeptag: inferring diagnoses from veterinary clinical notes. npj Digit. Medicine 1, 60 (2018).

[8] Steinhardt, J., Koh, P. W. W., and Liang, P. S. Certified defenses for data poisoning attacks. In Advances in Neural Information Processing Systems, pp. 3517–3529, 2017.