UNIVERSITY OF CALIFORNIA
Los Angeles

Enhancing Predictive Analytics in Diabetes Management: A Comparative Analysis of Linear and Non-linear Modeling Techniques with SHapley Additive exPlanations Value Integration

A thesis submitted in partial satisfaction of
the requirements for the degree Master of Applied Statistics and Data Science

by

Baiwei Lu

2024

ABSTRACT OF THE THESIS

Enhancing Predictive Analytics in Diabetes Management: A Comparative Analysis of Linear and Non-linear Modeling Techniques with SHapley Additive exPlanations Value Integration

by

Baiwei Lu

Master of Applied Science in Statistics And Data Science

University of California, Los Angeles, 2024

Professor Yingnian Wu, Chair

This thesis provides a detailed comparative analysis of non-linear and linear modeling techniques in feature correlation analysis. In this work, I explore various testing methods on non-linear models, which are often challenging to interpret statistically. This challenge is particularly prevalent in the medical field where the complexity of models can obscure their interpretability. To address this issue, I introduce SHapley Additive exPlanations (SHAP) value as a means to enhance clarity and precision in the interpretation of results. Throughout this comparison, I will demonstrate that integrating SHAP value allows us to merge the interpretability of linear models with the accuracy of non-linear models, thus leveraging the strengths of both approaches. The difficulty in interpreting non-linear models stems from their complexity and the intricate relationships they encapsulate, which do not easily translate into direct, understandable insights as linear models do [1].

The thesis of Baiwei Lu is approved.

Guido F. Montufar Cuartas
Guang Cheng
YingNian Wu, Committee Chair

University of California, Los Angeles
2024

Contents

VITA

2014–2018	Bachelor of Science Accounting, University of Nevada, Las Vegas

1 Introduction

Interpretability is a crucial part of statistics in medical fields, so I have decided to implement a diabetes dataset. The dataset originally came from the National Institute of Diabetes and Digestive and Kidney Diseases of India [?]. I will be using both non-linear and linear models to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. In this dataset, there are several variables, some are independent, such as the medical predictor variable (predictor for if a patient will be diagnosed with diabetes), and only one target dependent variable which is the outcome (diagnosed with diabetes, or not diagnosed with diabetes). As highlighted in the abstract, the interpretability of non-linear models in medical statistics is a pivotal challenge, particularly in the realm of diabetes diagnosis using complex datasets. This research delves into the sophisticated realms of machine learning models, both linear and non-linear, to tackle this challenge. The dataset, sourced from the National Institute of Diabetes and Digestive and Kidney Diseases of India, includes a variety of diagnostic measurements pivotal for predicting diabetes. The condition's prevalence and its significant impact on public health make it imperative to enhance the accuracy and understandability of diagnostic predictions.

This investigation aims to address several key challenges:

- Complexity of Model Interpretation: Non-linear models, while powerful, often lack easy interpretability, which is critical in clinical settings where decisions need clear justifications.
- Integration of SHAP Values: We introduce SHapley Additive exPlanations (SHAP) values to our analysis to improve the interpretability of these complex models. SHAP values provide a nuanced view of how each feature in the dataset influences the prediction outcome, thus bridging the gap between accuracy and interpretability [1].
- Comparative Analysis: Through a systematic comparison between linear and non-linear models, this research elucidates how each model processes data and impacts predictive outcomes. This not only aids in understanding which models are most effective under varying conditions but also enhances their practical application in real-world diagnostic processes.

The potential impact of this investigation extends beyond academic circles into practical medical applications. By improving model transparency and effectiveness, healthcare professionals can gain a better understanding of diagnostic tools, leading to more informed clinical decisions and ultimately, better patient outcomes.

1.1 Diabetes Diagnosis in Clinical Practice

- **Fasting Plasma Glucose (FPG) Test:** This test measures blood glucose levels after an individual has fasted for at least eight hours. A fasting glucose level of 126 mg/dL (7.0 mmol/L) or higher on two separate occasions typically indicates diabetes.
- **Oral Glucose Tolerance Test (OGTT):** During this test, a person's blood sugar is checked after fasting and again two hours after ingesting a glucose-rich drink. This test shows how efficiently the body processes sugar. A two-hour blood sugar level of 200 mg/dL (11.1 mmol/L) or higher suggests diabetes.
- **Hemoglobin A1c Test:** This test provides a snapshot of the average blood glucose control for the past two to three months. An A1c level of 6.5
- **Random Plasma Glucose Test:** This test measures blood glucose regardless of when the person last ate. A blood sugar level of 200 mg/dL (11.1 mmol/L) or higher suggests diabetes, particularly if accompanied by symptoms such as increased thirst, urination, and fatigue.

These tests not only assist in diagnosing diabetes but also in monitoring the condition's progression and evaluating the effectiveness of prescribed management strategies. Understanding these diagnostic methods is crucial for developing predictive models, as they provide a clinical backdrop against which the predictive accuracy of such models can be assessed. By integrating predictive analytics into this diagnostic process, physicians may enhance their ability to detect diabetes earlier and manage it more effectively, thus improving patient outcomes.

2 Preparations and EDA and Introducing Dataset Variables

This study utilizes a comprehensive dataset provided by the National Institute of Diabetes and Digestive and Kidney Diseases of India, focusing on diagnostic measurements vital for predicting diabetes. The dataset is composed of several key variables, each playing a crucial role in the diagnostic process. These include:

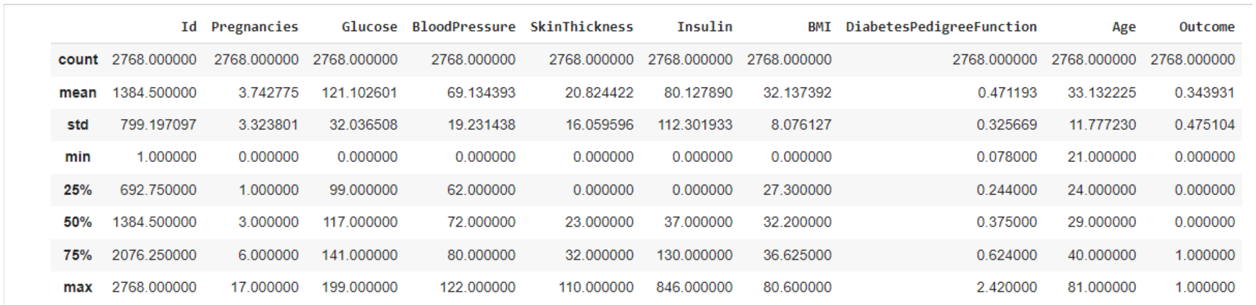
- **Glucose:** Plasma glucose concentration 2 hours post-administration in an oral glucose tolerance test.
- **Diastolic Blood Pressure (DBP):** Diastolic blood pressure measurement, an essential indicator of cardiovascular health.
- **Skin Thickness:** Triceps skinfold thickness, which helps estimate body fat.
- **Insulin:** 2-hour serum insulin, a crucial measure reflecting insulin resistance or deficiency.
- **BMI (Body Mass Index):** A key metric for assessing obesity, calculated as weight in kilograms divided by the square of height in meters.
- **Diabetes Pedigree Function:** A genetic function that predicts the likelihood of diabetes based on family history.
- **Age:** The age of the patient.

These variables are critical as they directly impact the diagnostic outcomes of the diabetes prediction model used in this study. Understanding the nature and roles of these variables is essential for interpreting the outcomes of any analytical models applied to this data.

2.1 Standardizing Zero Values

- **Glucose:** We need to remove any rows with 0 values in this column because it is not medically viable to reach 0 glucose in an oral test. Therefore, not removing them will prohibit us from using this column as a feature.
- **Diastolic Blood Pressure:** No action is required for this column.
- **Skin Thickness:** This column may not be meaningful as using BMI will be a better option. We will not be using this column; therefore, no action is required.
- **Insulin:** It is unlikely in a real-life situation to see an insulin concentration in a 2-hour serum sample to be exactly 0. However, in a dataset, a 0 value in this column might have the following interpretations:
  - *Missing Data or Measurement Limitation:* A 0 value might indicate missing data or that the measurement fell below the detection limit of the assay used to measure insulin. Some laboratory assays may not accurately measure very low concentrations.
  - *Data Entry Error:* It could also be a data entry error or an encoding convention for missing or undetectable values.

  No action required. Decide before using this column as a feature.
- **BMI:** A value of 0 is not physiologically meaningful and is likely indicative of missing or incorrect data rather than a valid measurement. We will be removing rows with 0 value in this feature.



	Id	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	2768.000000	2768.000000	2768.000000	2768.000000	2768.000000	2768.000000	2768.000000	2768.000000	2768.000000	2768.000000
mean	1384.500000	3.742775	121.102601	69.134393	20.824422	80.127890	32.137392	0.471193	33.132225	0.343931
std	799.197097	3.323801	32.036508	19.231438	16.059596	112.301933	8.076127	0.325669	11.777230	0.475104
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	692.750000	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.244000	24.000000	0.000000
50%	1384.500000	3.000000	117.000000	72.000000	23.000000	37.000000	32.200000	0.375000	29.000000	0.000000
75%	2076.250000	6.000000	141.000000	80.000000	32.000000	130.000000	36.625000	0.624000	40.000000	1.000000
max	2768.000000	17.000000	199.000000	122.000000	110.000000	846.000000	80.600000	2.420000	81.000000	1.000000

Figure 1: Dataset Descriptive Statistics

2.2 Final Display of Cleaned Dataset and Variables Explained

Now that we have cleaned our dataset, here is a final display of the dataset along with an explanation of each variable and how we will utilize each one in our non-linear models to explore their correlation in diagnosing diabetes and how it affects their accuracy.

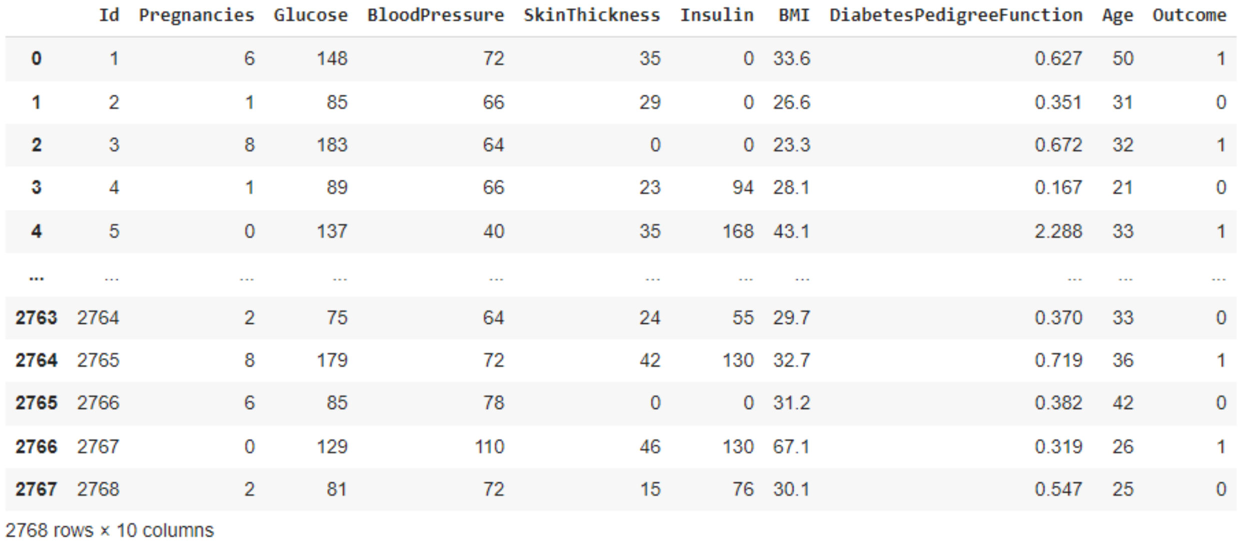
Here is the dataset with standardized zero values and handled NaN and missing fields. I will be explaining each variable and allocating them in the $y = ax + b$ equation to better help us understand their attributes in the models and methods we will be using.

- **ID:** Assigned ID to each patient.
- **Pregnancies:** Number of times pregnant.
- **Glucose:** Plasma glucose concentration over 2 hours in an oral glucose tolerance test.
- **Blood Pressure:** Diastolic blood pressure (mm Hg).
- **Skin Thickness:** Triceps skinfold thickness (mm).
- **Insulin:** 2-hour serum insulin (mu U/ml).
- **BMI:** Body mass index (weight in kg / height in m^2).
- **Diabetes Pedigree Function:** Diabetes pedigree function, a genetic score of diabetes.
- **Age:** Age in years.

3 Advanced Analytical Techniques and Model Evaluation in Diabetes Prediction

In this section of our analysis, we delve into the feature importance as derived from Logistic Regression and Random Forest models applied to the Healthcare-Diabetes dataset. Our aim is to thoroughly investigate the relationships among different variables within the dataset and assess how significantly each variable influences the diagnosed outcome of diabetes. By employing these traditional modeling techniques, we hope to uncover meaningful correlations and variable significances that could potentially guide clinical decision-making and treatment strategies.

As we progress through the thesis, a comparative analysis will be introduced, juxtaposing the findings from this initial exploration with results driven by SHAP (SHapley Additive exPlanations) values. This



	Id	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	1	6	148	72	35	0	33.6	0.627	50	1
1	2	1	85	66	29	0	26.6	0.351	31	0
2	3	8	183	64	0	0	23.3	0.672	32	1
3	4	1	89	66	23	94	28.1	0.167	21	0
4	5	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...	...
2763	2764	2	75	64	24	55	29.7	0.370	33	0
2764	2765	8	179	72	42	130	32.7	0.719	36	1
2765	2766	6	85	78	0	0	31.2	0.382	42	0
2766	2767	0	129	110	46	130	67.1	0.319	26	1
2767	2768	2	81	72	15	76	30.1	0.547	25	0

2768 rows × 10 columns

Figure 2: Display of the Cleaned Dataset

later section will emphasize the advantages of integrating SHAP values, particularly in enhancing the interpretability of non-linear models such as Random Forest. By adopting SHAP values, we can gain a more nuanced understanding of how individual features contribute to the prediction outcomes, going beyond mere importance to reveal the actual impact of each variable.

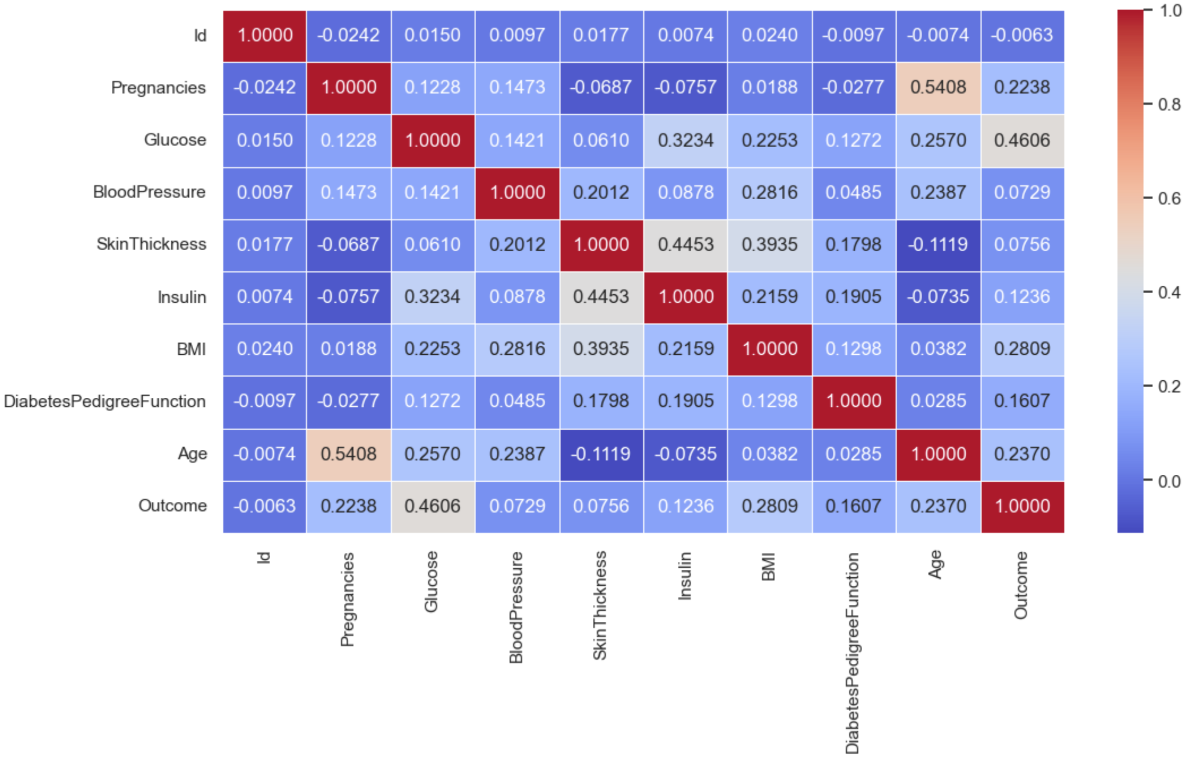
This dual approach allows us not only to identify key drivers of diabetes within the dataset but also to explore how different modeling techniques can either obscure or clarify the underlying mechanisms of the disease. Ultimately, this will help to illustrate the robustness of SHAP-enhanced methods in capturing complex interactions within the data, providing deeper insights and more actionable intelligence for healthcare professionals.

3.1 Feature of Importance

Figure 3 introduces a correlation matrix heatmap, a fundamental tool in our analysis of the Healthcare-Diabetes dataset. This matrix displays the Pearson correlation coefficients for each variable pair, ranging from -1 to 1. Positive values in this matrix indicate a direct relationship where an increase in one variable tends to increase another, while negative values suggest inverse relationships. The color intensity of each cell highlights the strength of these correlations, with darker shades denoting stronger associations.

In this study, we employ the correlation heatmap to initially identify relationships between variables, particularly focusing on their link to the 'Outcome' variable, which indicates whether patients are diagnosed with diabetes (1) or not (0). This preliminary analysis helps in pinpointing key variables that warrant further investigation through more nuanced analytical tools.

Moving forward, the discussion will delve deeper into interpreting these correlations within the specific context of our dataset. We will explore how each variable's relationship with diabetes diagnosis can inform model development and patient management strategies. This approach not only leverages basic correlation analysis but also sets the groundwork for applying more advanced interpretative methods like SHAP values in subsequent sections. These efforts highlight the evolving nature of data analytics in healthcare and underscore the necessity of using targeted analytical tools to meet specific research needs.



	Id	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Id	1.0000	-0.0242	0.0150	0.0097	0.0177	0.0074	0.0240	-0.0097	-0.0074	-0.0063
Pregnancies	-0.0242	1.0000	0.1228	0.1473	-0.0687	-0.0757	0.0188	-0.0277	0.5408	0.2238
Glucose	0.0150	0.1228	1.0000	0.1421	0.0610	0.3234	0.2253	0.1272	0.2570	0.4606
BloodPressure	0.0097	0.1473	0.1421	1.0000	0.2012	0.0878	0.2816	0.0485	0.2387	0.0729
SkinThickness	0.0177	-0.0687	0.0610	0.2012	1.0000	0.4453	0.3935	0.1798	-0.1119	0.0756
Insulin	0.0074	-0.0757	0.3234	0.0878	0.4453	1.0000	0.2159	0.1905	-0.0735	0.1236
BMI	0.0240	0.0188	0.2253	0.2816	0.3935	0.2159	1.0000	0.1298	0.0382	0.2809
DiabetesPedigreeFunction	-0.0097	-0.0277	0.1272	0.0485	0.1798	0.1905	0.1298	1.0000	0.0285	0.1607
Age	-0.0074	0.5408	0.2570	0.2387	-0.1119	-0.0735	0.0382	0.0285	1.0000	0.2370
Outcome	-0.0063	0.2238	0.4606	0.0729	0.0756	0.1236	0.2809	0.1607	0.2370	1.0000

Figure 3: Heatmap of Correlation Coefficients Between Variables

3.2 Another Example of Correlation Plot

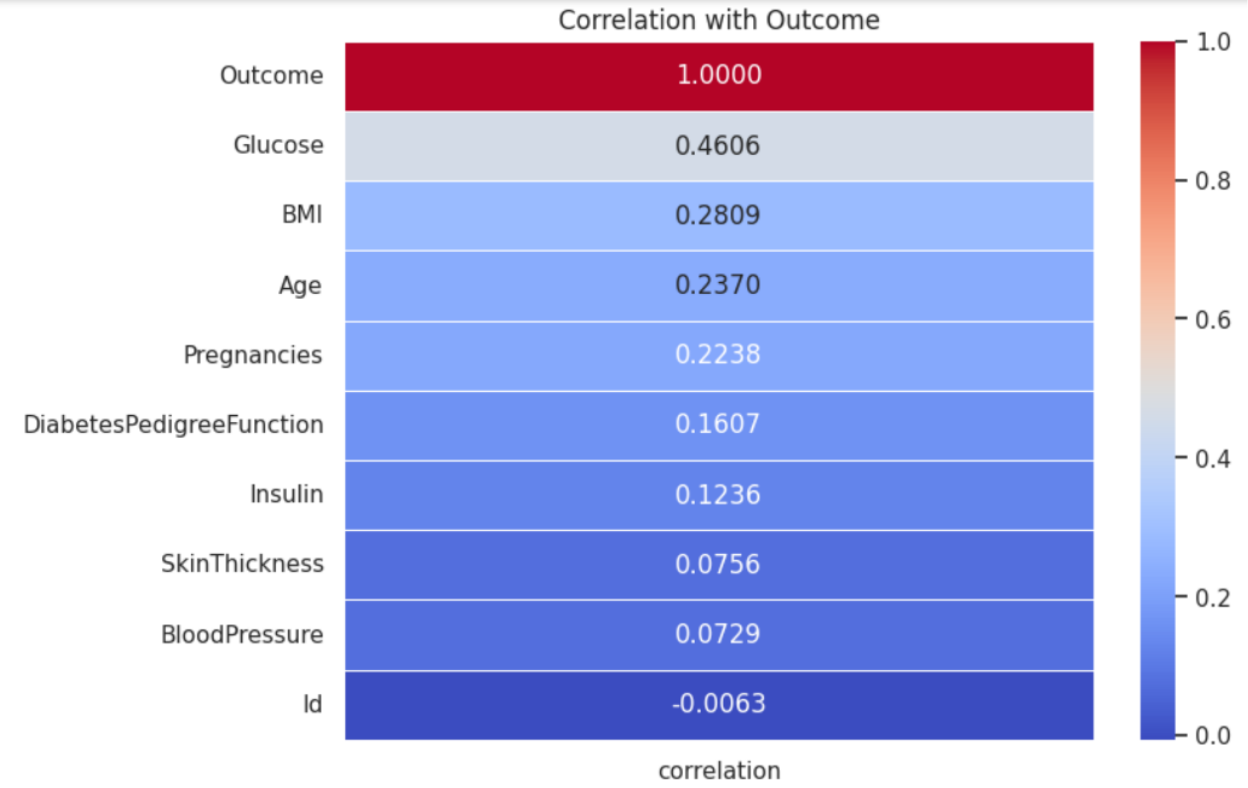


Figure 4: Visualized Correlation Heatmap

Figure 4 is a heatmap visualizing the correlation coefficients between various features and the binary outcome variable representing diabetes diagnosis (1 for diagnosed, 0 for not diagnosed). This visual tool is crucial for feature selection, allowing researchers to identify the most predictive variables, such as glucose and BMI, which are vital for diabetes management. High correlation values suggest key areas for focused data collection and patient risk stratification. For instance, features like glucose levels and BMI, which show strong correlations with the diabetes outcome, are prioritized in model development and public health interventions. Conversely, features with minimal correlations, such as patient IDs, are excluded to streamline model complexity without losing predictive accuracy. This heatmap not only facilitates the understanding of disease dynamics but also guides healthcare providers in targeting interventions and patient education based on identified risk factors. By providing a clear, intuitive overview of how various factors are linked to diabetes outcomes, the heatmap serves as an indispensable tool in exploratory data analysis and aids stakeholders across healthcare, finance, and social sciences in making informed decisions based on data-driven insights.

3.3 Key Feature: Correlation Values

High Positive Correlation (Red): Features such as Glucose have a strong positive correlation with the outcome, which in this context likely represents a diabetes diagnosis. A correlation value of 0.4606 for glucose suggests that higher glucose levels significantly increase the likelihood of diabetes.

Moderate to Low Correlation (Light Blue to Blue): Features like BMI, Age, and Pregnancies have moderate positive correlations, implying they are somewhat influential in predicting the outcome but less so than glucose.

Very Low to No Correlation (Near White): The Id feature shows near-zero correlation, indicating it has no predictive value regarding the outcome, which is expected if ‘Id‘ merely represents a unique identifier for records.

3.4 Key Features: Direction of Correlation

Positive values indicate that as the feature value increases, the likelihood of the positive outcome (e.g., diabetes) also increases. Negative values would indicate an inverse relationship; however, in this plot, all relevant features show positive correlations.

3.5 Linear Methods: Logistic Regression

The confusion matrix (Figure 5) presented shows how the Logistic Regression model categorizes the dataset into instances of diabetes (class 1) and non-diabetes (class 0). The matrix elements [177 25; 41 43] describe the count of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP) respectively. This indicates that while the model is fairly accurate in predicting non-diabetes cases, it has a notable number of errors, particularly in terms of false negatives and false positives.

The classification report provides a more detailed view of the model's performance, displaying precision, recall, and F1-score for each class. For class 0 (non-diabetes), the model shows a precision of 0.81 and a recall of 0.88, resulting in an F1-score of 0.84. For class 1 (diabetes), the precision drops to 0.63 with a recall of 0.51 and an F1-score of 0.57. The discrepancies in performance metrics between the classes suggest that the model is better at identifying non-diabetes than diabetes cases.

The overall accuracy (0.77) is complemented by the macro average and weighted average scores for precision, recall, and F1-score, which provide insight into the model's average effectiveness across both classes, accounting for class imbalance.

The analysis concludes that while the model achieves a commendable accuracy, there's room for improvement, especially in reducing the number of false negatives to enhance its ability to correctly identify more instances of class 1 (diabetes). This improvement is crucial as false negatives (patients with diabetes incorrectly classified as non-diabetic) are particularly risky in a medical context, potentially leading to the absence of necessary medical intervention.

To evaluate the predictive power of the logistic regression model effectively, we divided the dataset into two subsets: a training set and a testing set. This division allows us to fit the model on one portion of the data and evaluate its performance on another, thus ensuring that our assessment reflects the model's ability to generalize to new, unseen data.

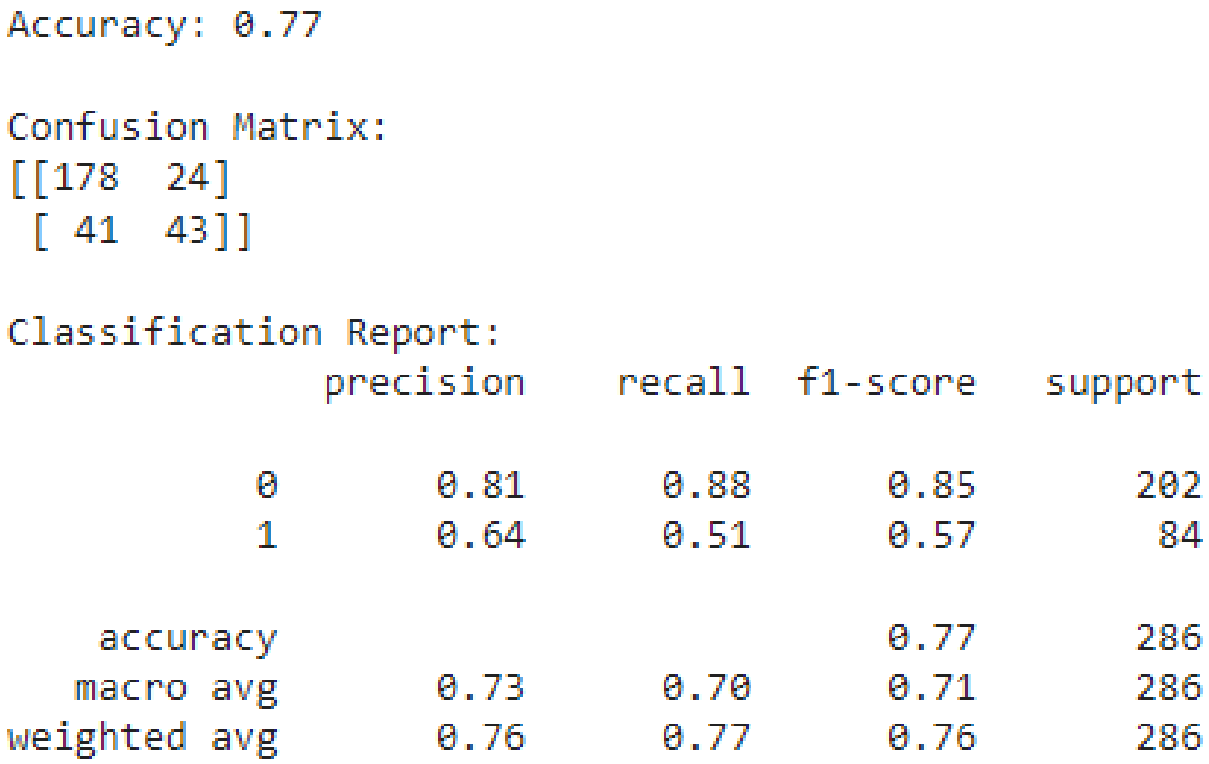
3.6 Dataset Division

- **Training Set:** We allocated 70% of the dataset to the training set, which comprises 537 data points. This subset was used to train the logistic regression model, allowing it to learn the relationship between the various predictors and the diabetes outcome.

- **Testing Set:** The remaining 30% of the dataset, consisting of 231 data points, was set aside as the testing set. This subset was not used during the model training phase and served exclusively for testing the model's performance to mimic how it would perform in a real-world scenario.

Model Fitting and Evaluation: The logistic regression model was fitted using the training set. The fitting process involved adjusting the model parameters to minimize the prediction error between the observed outcomes and the predictions made by the model. After fitting, the model's accuracy, precision, and recall were evaluated using the testing set. This approach ensures that our evaluation metrics reflect the model's effectiveness in predicting new cases of diabetes, which is essential for its potential application in clinical settings.

The section underscores the necessity of further optimizing the Logistic Regression model or considering more sophisticated modeling approaches to better handle the complexities of the dataset, thereby improving predictive accuracy and reliability in clinical applications.



```
Accuracy: 0.77

Confusion Matrix:
[[178  24]
 [ 41  43]]

Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.88      0.85       202
           1       0.64      0.51      0.57        84

    accuracy                           0.77       286
   macro avg       0.73      0.70      0.71       286
weighted avg       0.76      0.77      0.76       286
```

Figure 5: Confusion Matrix

3.7 Non-Linear Methods: Random Forest

The confusion matrix, showing in figure 5, indicates that the Random Forest model correctly classified every instance in the dataset without any errors—there were no false positives or false negatives. The classification report further validates these results, displaying a precision, recall, and F1-score of 1.00 for both classes. This suggests that the model is equally adept at identifying patients with and without diabetes.

While the results are impressive, they also raise concerns about potential over-fitting. Over-fitting occurs when a model is too closely fitted to the specific data on which it is trained, and may not perform as well on new, unseen data. Given the perfect accuracy reported, it is plausible that the model might have perfectly memorized the training data rather than learning to generalize from it.

The section hints at the potential of SHAP (SHapley Additive exPlanations) values to provide deeper insights. SHAP values could help in understanding the feature contributions for each prediction, possibly revealing whether the model's predictions are based on meaningful patterns in the data or merely on idiosyncrasies of the training set. Integrating SHAP values could also assist in determining if the apparent perfection in prediction is genuine or a result of overfitting, and may help in exploring beyond the 1.00 accuracy by offering more nuanced interpretations and validations of the model's predictive behaviors.

This analysis implies that while the Random Forest model shows exceptional performance, there is a need for further investigation using advanced interpretative tools like SHAP values to ensure the robustness and applicability of the model to other datasets. This will help in confirming the model's efficacy and in potentially enhancing its predictive capabilities through deeper understanding and adjustment based on SHAP insights.

3.7.1 Model Fitting and Dataset Division

To evaluate the predictive power of the Random Forest model, we similarly divided the dataset into a training set and a testing set. This structured approach helps to ensure that the model is not only able to fit the training data but also effectively generalize to new, unseen data.

- **Training Set:** 70% of the dataset, consisting of 537 data points, was used to train the Random Forest model. This allows the model to learn complex patterns and relationships between the predictors and the outcome of diabetes.
- **Testing Set:** The remaining 30%, which includes 231 data points, was utilized as the testing set. This set plays a critical role in evaluating the model's performance, ensuring that our results are robust and applicable in real-world settings.

3.7.2 Measures to Avoid Over-fitting

Random Forest models are particularly susceptible to over-fitting, especially when dealing with datasets that have many features or when the trees in the forest are allowed to grow without constraints. To mitigate this risk, we implemented several strategies:

- **Limiting Tree Depth:** We restricted the maximum depth of each tree in the forest, which prevents the trees from creating overly complex models that fit the idiosyncrasies of the training data rather than capturing the general trends.
- **Number of Trees:** Increasing the number of trees in the forest helps in reducing variance without substantially increasing bias, which means each individual tree's over-fitting is less likely to affect the overall model's accuracy.
- **Feature Selection:** By randomly selecting a subset of features for splitting at each node, the model is less likely to learn noise present in the training data. This randomness also helps in making the model more robust.
- **Cross-Validation:** We employed k-fold cross-validation during training, which involves dividing the training set further into several smaller sets. This technique not only helps in tuning the hyper-parameters but also in validating the model's stability and performance across different subsets of the data.
- **Out-of-Bag (OOB) Error Estimation:** We used the OOB error estimate as a means of performing internal validation of the model's performance. Since each tree in the forest is trained on a slightly different set of data (due to bootstrapping), the OOB error provides a good estimate of the test error without needing a separate validation set.

These methods collectively contribute to a robust model that can generalize well from the training data to unseen data, thus minimizing the risk of over-fitting.

3.8 Test Error Plot

This Test Error vs. Number of Training Data Points plot illustrates how the test error of a model varies as the number of data points used in training changes. Below are some observations and interpretations of the plot:

- **Volatile Test Error with Small Datasets:** The plot shows a significant fluctuation in test error when the number of training data points is small (less than 100). Specifically, there's a sharp peak at around 100 data points. This could indicate that with fewer data points, the model may not have enough information to generalize effectively, leading to higher and more variable errors.
- **Decrease in Error with More Data:** As the number of training data points increases beyond 100, the test error generally decreases and stabilizes. This suggests that adding more data helps the model to learn more comprehensive patterns, thus improving its prediction accuracy on the test set.
- **Anomalies or Outliers:** The sharp peak and subsequent drops in error might also suggest the presence of certain training sets that either do not represent the problem space well or include outliers that mislead the model during training. This is particularly evident around the 100 data point mark.

- **Overall Trend:** Despite some fluctuations, the overall trend from around 150 data points onwards shows a decline in test error as more data points are used. This is typical in machine learning, where more data generally leads to better model performance, up to a point.
- **Plateauing of Error:** Towards the end of the curve, the error rate begins to plateau, suggesting that simply adding more training data may not result in significant improvements beyond a certain point. This can occur when the model has effectively captured the underlying trends in the data, or when intrinsic noise and un-modeled complexities in the data limit further improvements.

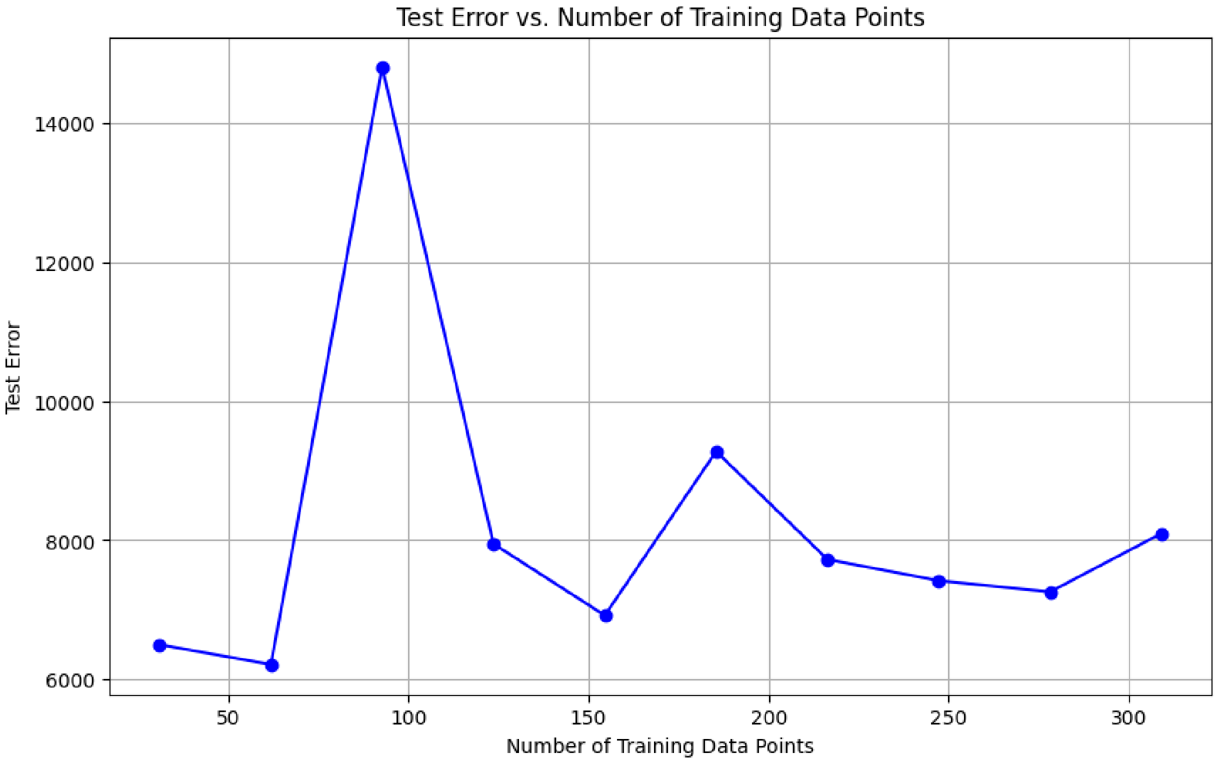
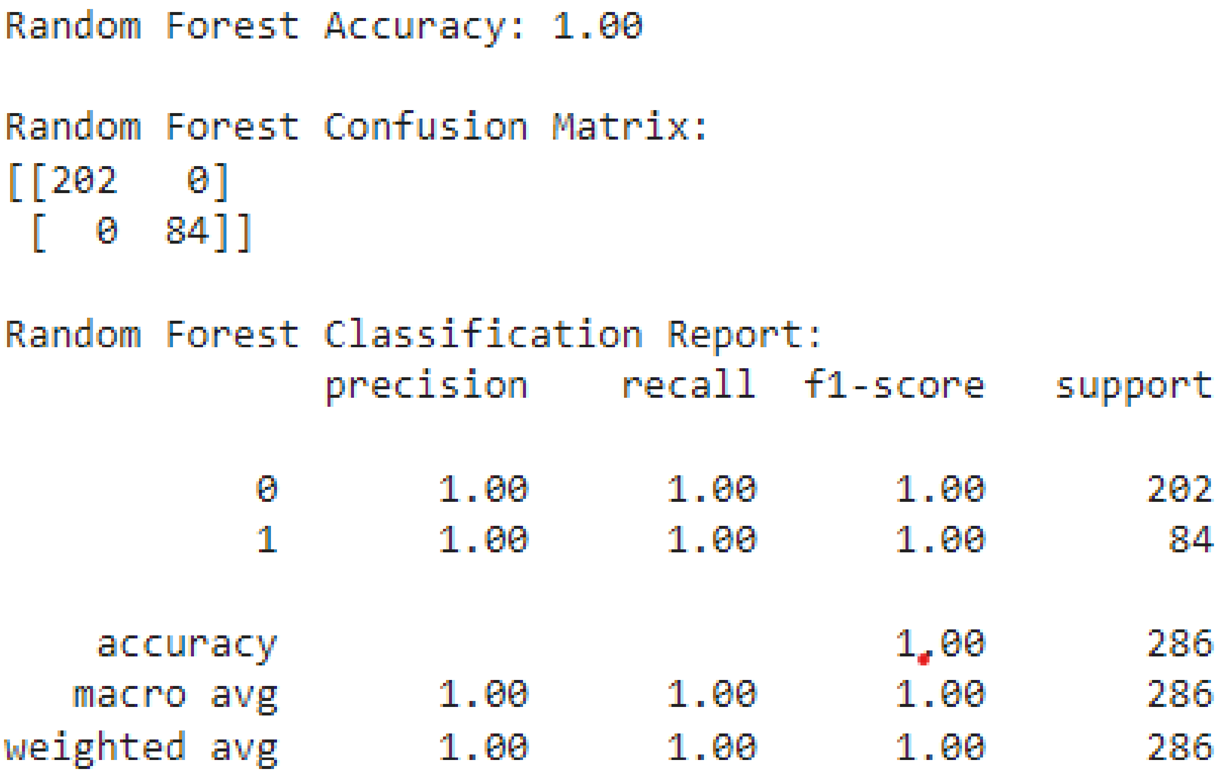


Figure 6: Test Error Plot

3.8.1 Model Evaluation

Following the implementation of these strategies, the Random Forest model was evaluated using the testing set. The evaluation aimed to assess the model's accuracy in predicting new cases of diabetes by examining its sensitivity and specificity across the varied data points in the test set. The results, as shown in Figure 7, indicate that the model achieved a perfect accuracy of 1.00. The confusion matrix in figure 7 demonstrates that there were no false positives or false negatives, highlighting the model's exceptional performance. Furthermore, the classification report confirms these findings, with precision, recall, and F1-score all being 1.00 for both classes. This suggests that the model is highly effective at correctly identifying both diabetic and non-diabetic cases without any errors.



```
Random Forest Accuracy: 1.00

Random Forest Confusion Matrix:
[[202   0]
 [  0  84]]

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       202
           1       1.00      1.00      1.00        84

    accuracy                           1.00       286
   macro avg       1.00      1.00      1.00       286
weighted avg       1.00      1.00      1.00       286
```

Figure 7: Random Forest

4 Advantages of Both Linear and Non-linear Models

Now that we have explored the comparative strengths and weaknesses of linear and non-linear models, specifically Logistic Regression and Random Forest, applied to a Healthcare-Diabetes dataset, the analysis reveals that while the Random Forest model achieves high accuracy and performance metrics, it also poses a risk of over-fitting, which could limit its utility in generalized settings. On the other hand, Logistic Regression, though less accurate in this instance, offers greater interpretability and stability across various scenarios due to its simpler and more transparent model structure.

This section can be expanded by discussing additional benefits and potential drawbacks of each modeling approach. For instance, Logistic Regression models are not only easy to implement and interpret but also computationally less intensive compared to Random Forest. This makes Logistic Regression particularly attractive for scenarios where transparency and speed are crucial. However, they can struggle with non-linear relationships unless explicitly modified to address such patterns, which can be a significant limitation for complex datasets.

Random Forest, being a non-linear model, excels in handling complex and high-dimensional data, capturing intricate patterns that Logistic Regression might miss. It is robust to outliers and can model interactions between variables without explicit specification. Nevertheless, its ”black box” nature makes it harder to interpret the influence of individual features, and it requires more computational resources, which can be a drawback for real-time applications.

The document could further benefit from a discussion on how integrating SHAP values aims to blend the strengths of both model types. SHAP values can enhance model interpretability by quantifying the contribution of each feature to the prediction, regardless of the model’s complexity. By using SHAP values, the goal is to develop a hybrid approach that leverages the accuracy and capability of handling complex interactions from Random Forest with the simplicity and clarity of Logistic Regression. This hybrid model would aim to provide a robust solution that is both accurate and interpretable, suitable for clinical settings where understanding the reasoning behind predictions is as critical as the predictions themselves.

Ultimately, the section should underscore the thesis's objective: to produce a methodology that synthesizes the advantages of linear and non-linear models into a unified model enhanced by SHAP values, thus providing a comprehensive tool for predictive analysis in healthcare applications. This approach promises to mitigate the individual limitations of each model type while capitalizing on their strengths, offering a sophisticated yet accessible tool for medical decision-making.

5 SHAP Values

SHAP (SHapley Additive exPlanations) values are derived from the concept of Shapley values in cooperative game theory. This method quantifies the contribution of each feature in a model by considering all possible combinations of features. The fundamental principle is to assess the impact of adding a feature to all possible subsets of other features, thereby determining the "average" contribution of each feature across all possible configurations.

SHAP values offer a powerful way to decode complex machine learning models, commonly referred to as "black-box" models, by providing detailed explanations of how each feature contributes to a specific prediction. This level of detail helps in understanding the model's behavior on a granular level, which is especially valuable in domains where interpretability is crucial, such as healthcare, finance, and legal applications.

One of the significant advantages of SHAP values over traditional feature importance methods used in linear and non-linear models is their ability to offer consistent and locally accurate attributions. While traditional methods might provide an overall score indicating a feature's importance across the dataset, they often fail to capture the context-dependent interactions and non-linear relationships effectively. SHAP values address this by allowing each prediction to be broken down into the sum of effects from each feature, providing a clearer and more actionable insight into the data.

Moreover, SHAP values enable the creation of advanced visualizations that were previously challenging to achieve with standard feature importance metrics. For example:

- **SHAP Summary Plot:** This visualization aggregates the SHAP values of all data points to show the impact of each feature across the dataset. It not only indicates which features are most important but also how their effects vary from one instance to another.

- **SHAP Dependence Plot:** These plots help in visualizing the relationship between the features and the impact they have on the model's output. It allows us to see how the model's predictions change with varying values of a feature, while coloring by another feature to suggest possible interactions.

- **SHAP Force Plot:** Individual predictions can be broken down to show the contribution of each feature. This is useful for detailed case-by-case analysis, helping users see why the model made a specific decision for an individual instance.

By providing these insights, SHAP values significantly enhance the transparency and trustworthiness of machine learning models. They facilitate a deeper understanding of the modeling process and enable practitioners to make more informed decisions based on the model's predictions. This makes SHAP an invaluable tool in the advancement of ethical AI and responsible machine learning practices.

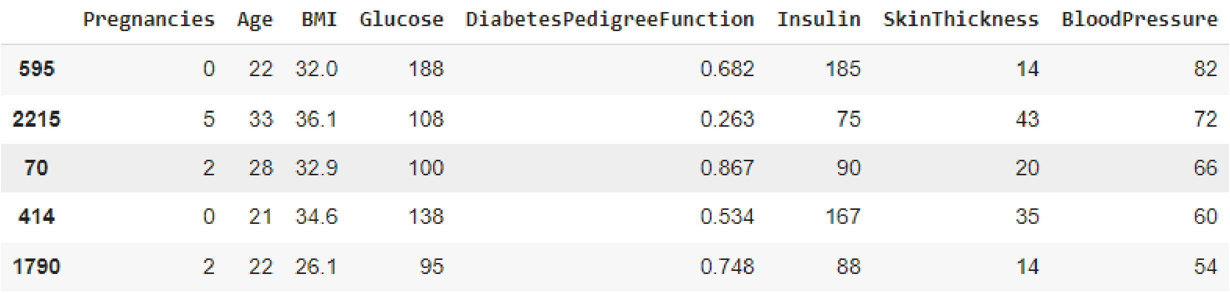
5.1 SHAP Values Generated Into The Dataset

We will now train the model to generate SHAP values for each patient (data point). In Figure 8, we see that a SHAP values column is added to the dataset. The integration of SHAP values into the Healthcare-Diabetes dataset enhances the dataset with new insights into the predictive model's behavior. The table shown in Figure 8 illustrates how SHAP values are generated and appended to each patient's data point in the dataset. These values provide a quantified measure of how much each feature contributes to the prediction of diabetes for each individual.

In the table, each row represents a patient, with columns for each relevant feature such as Pregnancies, Age, BMI, Glucose, DiabetesPedigreeFunction, Insulin, SkinThickness, and BloodPressure. Next to these feature columns, there is a corresponding SHAP value column for each feature. These SHAP values are numerical indicators of the contribution of each feature to the model's prediction for that specific patient.

A positive SHAP value indicates that the feature contributes positively towards a higher prediction of diabetes (increasing the likelihood of diabetes), while a negative SHAP value suggests a contribution towards a lower prediction (decreasing the likelihood of diabetes). The magnitude of the SHAP value reflects the strength of the feature's impact: larger values, whether positive or negative, indicate a stronger influence.

This detailed breakdown allows clinicians and researchers to understand not only which features are important in predicting diabetes but also how each feature's specific value for a given patient affects the prediction outcome. This level of insight is crucial for personalized medicine, as it enables healthcare providers to tailor their interventions based on the individualized risk factors highlighted by the SHAP values. Additionally, this method improves the model's transparency and trustworthiness by making it possible to trace and validate the contributions of individual features to the decision-making process.



	Pregnancies	Age	BMI	Glucose	DiabetesPedigreeFunction	Insulin	SkinThickness	BloodPressure
595	0	22	32.0	188	0.682	185	14	82
2215	5	33	36.1	108	0.263	75	43	72
70	2	28	32.9	100	0.867	90	20	66
414	0	21	34.6	138	0.534	167	35	60
1790	2	22	26.1	95	0.748	88	14	54

Figure 8: SHAP Values Generated

5.2 Theoretical Foundations of SHAP Values

Building upon the comprehensive insights provided by Lundberg and Lee's article in 2017, this thesis utilizes SHAP values not only as a tool for interpretation but also as a cornerstone for enhancing the theoretical robustness of predictive analytics. Their unified approach to interpreting model predictions, detailed in "A Unified Approach to Interpreting Model Predictions," provides a solid theoretical framework that supports the integration of SHAP values into our predictive models. This approach underscores the importance of model transparency and the necessity of interpretable outcomes in clinical settings. The article's discussion on computational advancements in SHAP value calculation also informs our methodology, ensuring efficient and effective implementation in large-scale datasets typical in healthcare diagnostics [4].

5.2.1 Integration with SHAP Values

Incorporating SHAP (SHapley Additive exPlanations) values, similar to the methodologies discussed in this thesis, Prescience offers a detailed explanation of each prediction it makes. This approach allows medical professionals to understand which patient-specific factors are contributing to the risk of hypoxemia. Such transparency is crucial in the medical field, where the stakes of decision-making are high, and the need for trust and clarity in predictive analytics is paramount.

5.2.2 Practical Benefits

The practical benefits of Prescience are multi-faceted:

- **Enhanced Patient Safety:** By predicting potential complications before they become critical, Prescience allows medical teams to intervene preemptively, significantly enhancing patient safety during surgeries.

- **Operational Efficiency:** Real-time predictions enable more efficient use of hospital resources, ensuring that interventions are timely and targeted.

- **Education and Training:** The interpretability provided by SHAP values helps in educating medical staff about predictive factors for hypoxemia, enhancing overall medical knowledge and response strategies.

5.2.3 Implications for Further Research

The success of Prescience underscores the potential for broader application of interpretable machine learning models across various medical domains. This case study supports the thesis's argument that machine learning, coupled with effective interpretability tools like SHAP values, can transform medical diagnostics and treatment planning. Future research could explore extending these models to other critical care scenarios, such as predicting sepsis or post-operative complications, where early intervention can drastically improve patient outcomes.

6 Feature of Importance With SHAP Value

The SHAP Values Feature of Importance plot provides a refined method for assessing the impact of individual features on the model's predictions, significantly enhancing interpretability.

In Figure 9, each feature, such as Glucose, Age, DiabetesPedigreeFunction, BMI, and others, is listed vertically, with their corresponding SHAP values spread horizontally. These values represent the influence of each feature value on the prediction outcome of the diabetes model. The plot is color-coded, ranging from blue to red, where blue indicates lower feature values and red indicates higher feature values. The position of each dot on the horizontal axis (SHAP value) signifies how much the feature value has contributed to pushing the model's prediction higher (to the right) or lower (to the left).

What makes this visualization particularly useful is its ability to show not only the direction and magnitude of each feature's impact on predictions but also the distribution of these effects across the dataset. For instance, high glucose levels (shown in red) are typically associated with higher SHAP values, indicating a strong positive influence on predicting diabetes. Conversely, lower glucose levels (in blue) tend to push the SHAP values towards the left, suggesting a decrease in the likelihood of a diabetes diagnosis.

This plot excels in ease of interpretability by clearly distinguishing between features that generally contribute positively or negatively to the model's predictions. It also allows for the observation of variability within a single feature, demonstrating how different values of the same feature can have varied effects on the outcome. Such detailed insights are invaluable for clinicians and researchers as they provide a clearer understanding of how specific feature values are associated with the model's decision-making process.

Moreover, this approach facilitates a deeper examination of the model by highlighting which features are most influential and under what conditions, aiding in more informed clinical decision-making and potentially guiding targeted interventions. The SHAP Values Feature of Importance plot thus serves as an essential tool in the realm of predictive modeling, offering a sophisticated yet accessible means to enhance the transparency and effectiveness of machine learning models in healthcare.

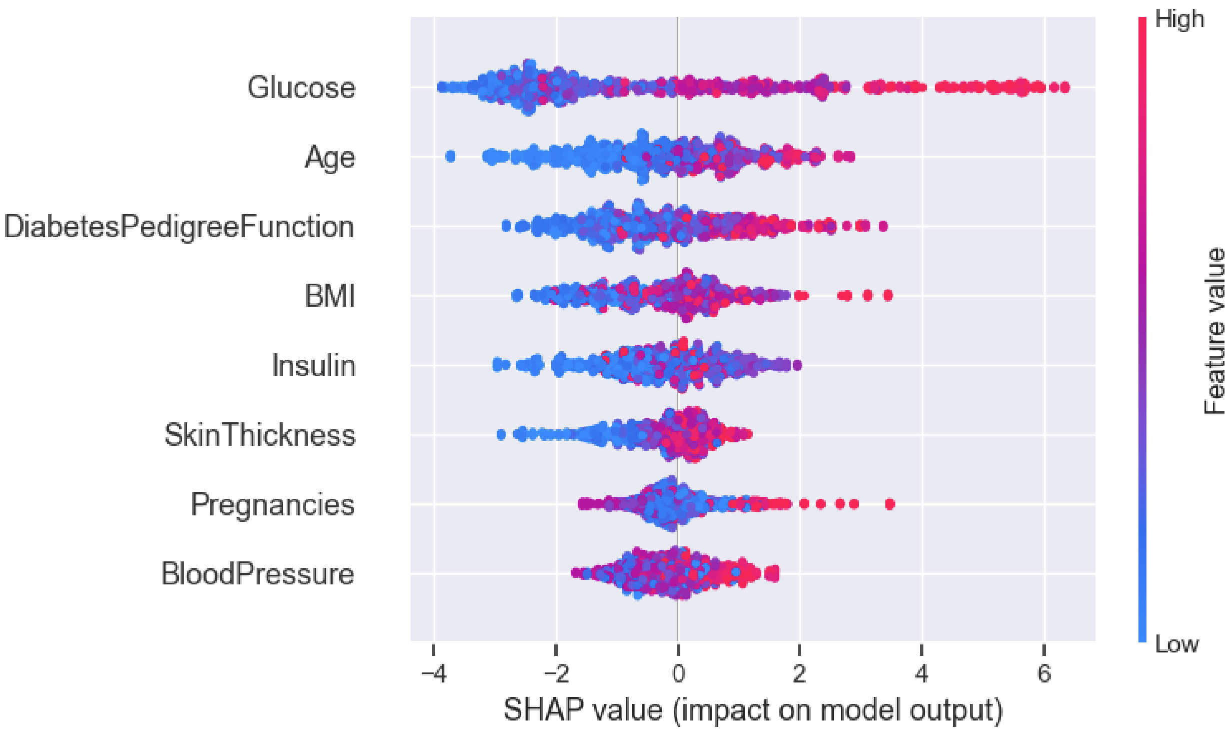


Figure 9: SHAP Values Feature of Importance

7 Comparison Between Feature of Importance

Next, we compare the SHAP feature of importance plot with traditional feature importance methods used in logistic regression and random forest models, revealing distinct advantages and nuances in interpretability and actionable insights.

7.1 Logistic Regression Feature of Importance

Feature importance in logistic regression is typically derived from the coefficients of the model. Each coefficient represents the change in the log odds of the outcome per unit change in the feature. While this method provides a straightforward metric of influence, it primarily reflects the linear association between features and the outcome. It lacks the capacity to capture complex interactions or the varying impact of a feature across different conditions within the data.

7.2 Random Forest Feature of Importance

Feature importance in random forest models is often measured by the mean decrease in impurity (e.g., Gini impurity) caused by splits on each feature or by the mean decrease in accuracy when a feature is omitted. This approach offers insights into how crucial a feature is for the model's performance but tends to aggregate the feature's contribution across all observations, potentially obscuring individual data point contributions and interactions.

7.3 Advantages of SHAP Feature of Importance

- **Detailed Impact Visualization:** Unlike the global aggregate measures provided by logistic regression and random forest, SHAP values generate individual impact scores for each feature for every prediction.

This allows for a detailed plot that not only shows the average importance of features but also how each feature's value influences predictions in specific instances.

- **Direction and Magnitude:** The SHAP feature of importance plot visually distinguishes between positive and negative influences of features on the prediction outcome, represented on a scale. This contrasts sharply with traditional methods that provide a single importance score per feature without indicating the direction of influence or how feature values affect the predictions.
- **Color-coded Value Distribution:** SHAP plots use color coding to represent feature values, adding an additional layer of data interpretation. This helps in understanding how higher or lower values of a feature systematically influence the model's output, which is particularly useful for features with non-linear effects on the prediction.

The SHAP feature of importance plot provides a more nuanced and comprehensive view compared to traditional feature importance methods. It not only reveals which features are important but also how different values of these features impact predictions, offering richer insights into the model's behavior. This level of detail is especially beneficial in settings where understanding the specific role of features is crucial for making informed decisions, such as in medical diagnostics or personalized treatment planning. Therefore, SHAP values enhance interpretability and trust in model predictions, making them highly valuable for complex decision-making environments.

8 Force Plots

In Figure 10, we see force plots created using the SHAP values assigned to the variables. These plots are much clearer and more precise compared to the feature importance plots from linear and non-linear methods. As shown here, we can visually represent the contribution of each variable to the prediction through the color and proportion of its bar size simultaneously. From an interpretability perspective, we can easily identify which variables contribute positively, with Glucose being the main contributor to our prediction in this instance.

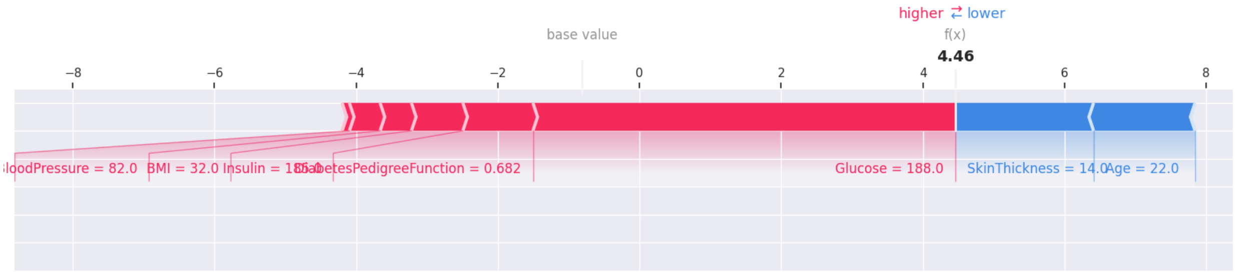


Figure 10: SHAP Values Feature of Importance Force Plot

8.1 Key Features: Contribution Visualization

The force plot provides a comprehensive view of how each feature contributes to the model's prediction for a specific instance. Each feature is displayed along a horizontal axis, with bars extending to either side to indicate their respective SHAP values. The length and direction of these bars are crucial: the longer the bar, the greater the magnitude of the feature's impact on the prediction. Positive SHAP values (depicted in red) signify features that increase the prediction value, pushing the model's output higher. Conversely, negative SHAP values (shown in blue) represent features that decrease the prediction value, pulling the model's output lower. This clear visual representation helps users easily discern which features are driving the prediction up or down and by how much. This granularity is particularly useful for understanding complex models, where interactions between variables can significantly affect the outcome.

8.2 Key Features: Base Value and Prediction

The base value in the force plot, typically centered, represents the average model prediction across the entire dataset. This serves as a reference point for understanding individual predictions. The final prediction value, denoted as $f(x)$, is the result of adjusting the base value by the SHAP values of all features for that specific instance. In the provided plot example, the prediction value is 4.46. This means that starting from the base value, the cumulative effect of the individual SHAP values for each feature brings the prediction to 4.46. This method of combining the base value with individual feature contributions provides a transparent and interpretable explanation of how the model arrives at a particular prediction, offering insights into the significance and direction of each feature's impact.

8.3 Key Features: Feature Contribution

The force plot excels in highlighting the individual contributions of each feature to the overall prediction. For instance, in the example provided, Glucose stands out as the most significant contributor to the prediction, as evidenced by its large positive SHAP value. This indicates that higher glucose levels significantly increase the predicted risk of diabetes. Other features, such as Skin Thickness and Age, also contribute to the prediction but to a lesser extent compared to Glucose. The plot visually differentiates the impact of each feature, making it easy to see which variables are the primary drivers of the model's predictions. This detailed breakdown is invaluable for applications like healthcare, where understanding the relative importance of different risk factors can guide clinical decision-making and personalized treatment plans. By providing a clear and precise depiction of feature contributions, the force plot enhances the interpretability and trustworthiness of complex predictive models.

8.4 Key Features: Interpretability

One of the key strengths of the force plot is its exceptional interpretability. This plot clearly shows which variables are driving the model's predictions and by how much, making it easy to understand individual predictions in detail.

Detailed Insights:

- **Clarity of Contribution:** The plot displays each feature's SHAP value, indicating how much each variable moves the prediction higher or lower. This is crucial for understanding specific influences on a prediction, such as high glucose levels increasing diabetes risk.
- **Visual Representation:** Color coding (red for positive impacts, blue for negative impacts) provides quick visual cues, simplifying complex interactions and making the plot accessible even to non-experts.
- **Quantitative Understanding:** The length of each bar shows the strength of a feature's impact, helping prioritize which features to focus on.

Practical Applications:

- **Individual Predictions:** The plot is invaluable for detailed analysis of individual predictions, offering personalized insights critical in fields like healthcare.
- **Model Debugging and Validation:** It helps data scientists identify unexpected patterns or biases, leading to model improvements.
- **Stakeholder Communication:** Its simplicity makes it an excellent tool for explaining model results to non-technical stakeholders, enhancing transparency and trust.

In summary, the force plot's interpretability allows users to see the precise contributions of each variable, facilitating informed decision-making and effective use of the model in practical applications.

8.5 Advantages over Traditional Methods

Compared to traditional feature importance methods in linear models (like coefficients in logistic regression) or non-linear models (like feature importance scores in random forests), SHAP force plots provide a more granular and precise understanding of how each feature affects the prediction.

Traditional methods often aggregate feature effects and lack the ability to show the impact direction and magnitude for specific instances. By breaking down the prediction into contributions from each feature, SHAP force plots offer actionable insights. For example, in a clinical setting, understanding that a patient's high glucose level is a major factor in predicting diabetes can help prioritize interventions.

In conclusion, Figure 10's force plot vividly demonstrates the power of SHAP values in making complex model predictions transparent and interpretable. This detailed breakdown aids in better decision-making, allowing stakeholders to understand and trust the model's predictions, especially in critical fields like healthcare.

9 Interactive Force Plot

Another option available to us through our newly added SHAP values is an interactive force plot. As we can see in Figure 11, we can freely choose which variable we want to focus on. By selecting a specific variable, it brings up a new graph that focuses on the selected variable and its contributions to the prediction compared to other variables in real time.

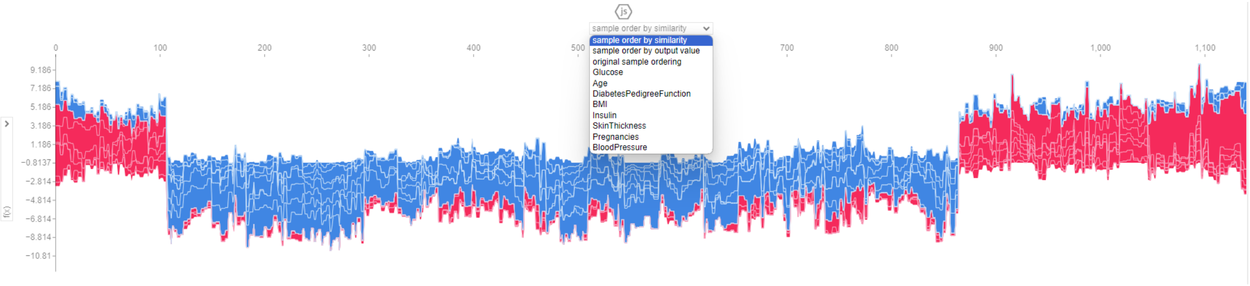


Figure 11: SHAP Values Feature of Importance Force Plot

9.1 Key Features: Dynamic Feature Selection

The interactive force plot allows users to choose any variable they wish to focus on. When a user selects a specific feature, the plot dynamically updates to highlight the contribution of that chosen variable in relation to others. This capability is particularly valuable for detailed investigations into how individual features affect the model's predictions. For instance, a user can select "Glucose" to observe how different glucose levels impact predictions across the dataset. This feature aids in pinpointing the exact role of a particular variable and understanding its interactions with other features.

9.2 Key Features: Real-time Analysis

The plot provides real-time feedback as users interact with it, enabling immediate analysis of each feature's impact on the model's predictions. This means that changes in the focus of the plot or adjustments in the variables being examined are instantly reflected, allowing for an efficient and seamless exploration process. This real-time aspect is crucial for quickly identifying patterns and relationships within the data, making it easier to draw meaningful conclusions and make informed decisions. For example, healthcare professionals can quickly analyze how changes in various health indicators like blood pressure or BMI affect diabetes risk predictions in real-time.

9.3 Key Features: Detailed Contribution Visualization

The interactive force plot uses a horizontal bar format where each bar represents a feature's SHAP value for a given prediction. The bars extend to the left or right of a central baseline, indicating whether a feature pushes the prediction higher or lower. Positive contributions (red bars) move the prediction towards a higher probability, while negative contributions (blue bars) move it towards a lower probability. This clear visualization helps users understand not only which features are important but also how they influence the prediction. The length of each bar corresponds to the magnitude of the feature's impact, providing a visual representation of the strength and direction of each variable's contribution.

9.4 Key Features:Enhanced interpretability:

The interactive nature of the plot significantly enhances interpretability by allowing users to manipulate and observe the data in a more intuitive manner. Users can interact with the plot to isolate specific features, compare the contributions of different variables, and understand the model's predictions more deeply. This interactive capability is especially beneficial in complex datasets where traditional static plots may fall short. For instance, a medical researcher can use the plot to understand how various combinations of factors like insulin levels, age, and skin thickness collectively influence diabetes risk, providing a comprehensive view that static methods cannot offer.

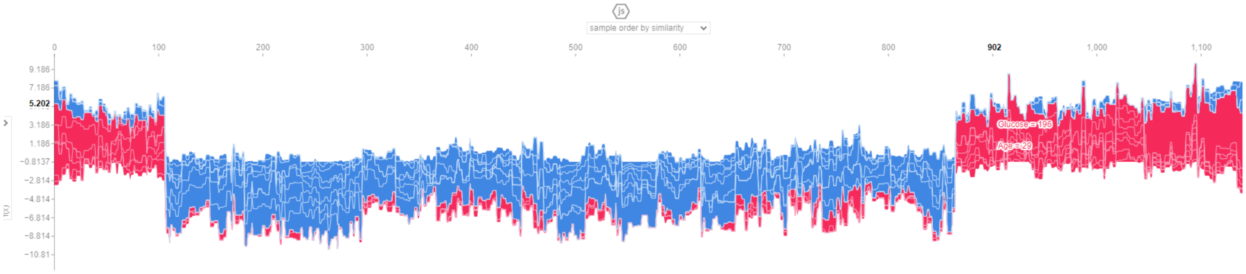


Figure 12: SHAP Values Feature of Importance Force Plot

9.5 Advantages over Traditional Plots

Interactive Exploration:

Unlike static plots, the interactive force plot enables users to explore different scenarios and what-if analyses. This flexibility is crucial for complex datasets where understanding the nuanced interplay between features can lead to more accurate and actionable insights. Precision and Depth:

The ability to visualize contributions in real-time and adjust the focus provides an unprecedented level of precision. This depth of analysis is particularly beneficial in fields like healthcare, where understanding the specific factors driving a prediction can inform treatment plans and interventions. In conclusion, the interactive SHAP force plot, as shown in Figure 11, offers a powerful tool for exploring and understanding feature contributions in a predictive model. Its dynamic and real-time capabilities enhance both precision and interpretability, making it an invaluable asset for researchers and practitioners in medical and other data-intensive fields.

10 Dependence Plot

Other than displaying how each variable contributes to the prediction, we can also explore how each variable depends on each other. As we have dependence plot here for Insulin and Age in Figure 13, we are able to explore and identify how age and insulin interact with each other in this dataset, we are able to interpret and understand the correlation between and across each pair of variables.

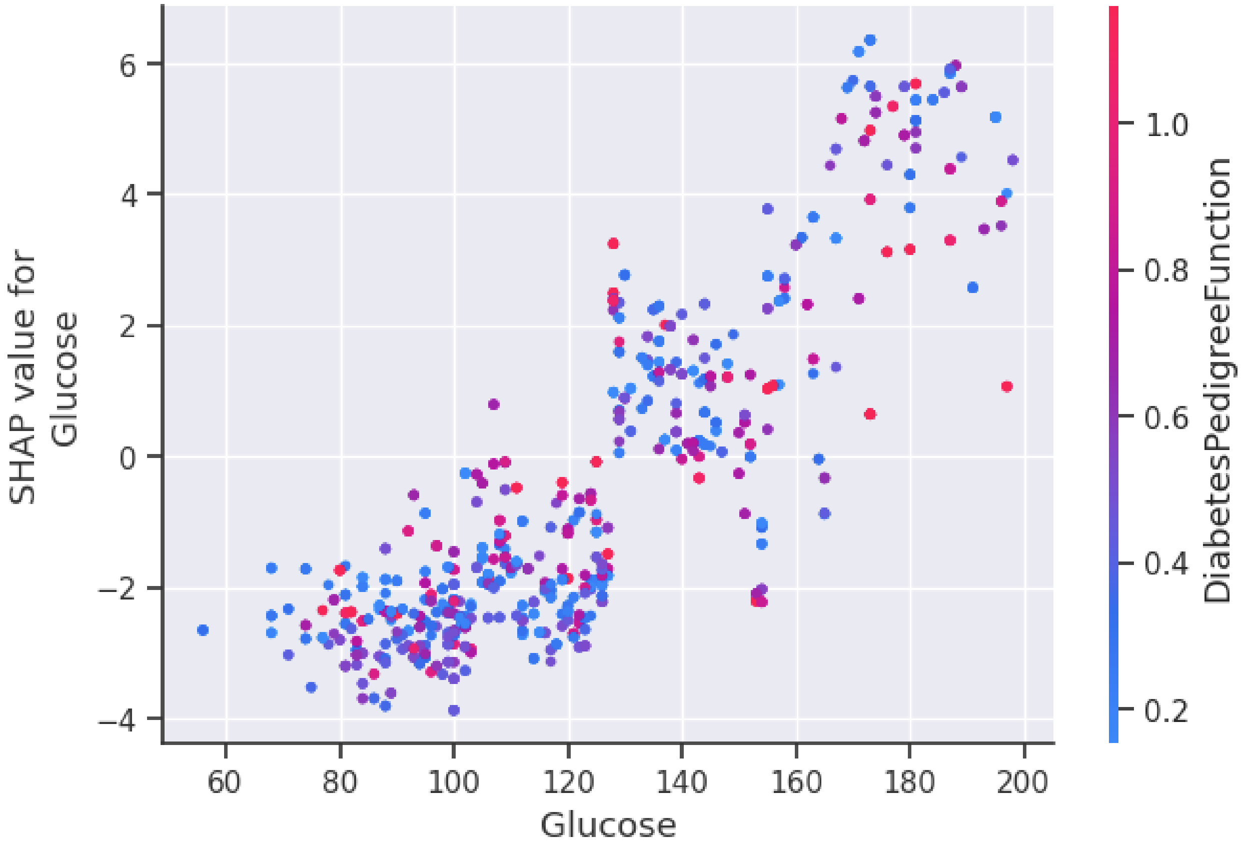


Figure 13: SHAP Values Feature of Importance Force Plot

10.1 Key Features: Interaction Exploration

Primary and Secondary Features: The plot illustrates the relationship between a primary feature (Glucose) and its corresponding SHAP value. This helps in understanding how changes in glucose levels impact the model's prediction. Secondary Feature Influence: It also incorporates a secondary feature (DiabetesPedigreeFunction), represented by a color gradient, which shows how this feature modifies the primary relationship. This allows users to see the combined effect of two features on the model's output.

10.2 Key Features: Correlation Identification

Visual Correlation: By plotting SHAP values against the feature values, the dependence plot helps identify correlations. For example, it can show a positive correlation where higher glucose levels lead to higher SHAP values, indicating a stronger influence on diabetes risk predictions. Pattern Recognition: The plot helps in recognizing patterns, such as linear or non-linear relationships between features, which are critical for understanding how features interact within the model.

10.3 Key Features: Detailed Insights

Highlighting Outliers: The dependence plot helps in identifying outliers or unusual patterns in the data. For example, it can highlight cases where high glucose levels do not follow the expected trend, prompting further investigation. Understanding Feature Impact: By showing how one feature's SHAP value changes with another feature, the plot provides insights into the relative importance and influence of different features in the model.

10.4 Advantages of SHAP Dependence Plot Compared to Traditional Dependence Plots

The SHAP dependence plot offers several significant advantages over traditional dependence plots used in linear and non-linear models. These advantages primarily stem from the detailed and interpret-able insights that SHAP values provide.

10.5 Advantage: Feature Interaction Exploration

Traditional Plots: Traditional dependence plots, especially in linear models, typically show the relationship between a single feature and the predicted outcome, often assuming that other variables are held constant. Non-linear models may offer more flexibility but still primarily focus on the primary feature's direct effect on the outcome. SHAP Dependence Plot: The SHAP dependence plot goes beyond this by incorporating the effects of a secondary feature through color coding. This multi-dimensional analysis allows for a deeper exploration of interactions between features, revealing how the relationship between the primary feature and the outcome is influenced by variations in another feature. This provides a more comprehensive view of feature interactions.

10.6 Advantage: Quantitative Contribution Visualization

Traditional Plots: In linear models, traditional dependence plots often show a simple linear relationship, which may not capture the complexity of the data. Non-linear models might show more complex relationships but do not quantify the exact contribution of each feature. SHAP Dependence Plot: SHAP plots not only visualize the relationship but also quantify the contribution of each feature to the model's prediction using SHAP values. This means you can see not just the direction and shape of the relationship but also how much each feature contributes to the prediction, providing both qualitative and quantitative insights.

10.7 Advantage:interpretability and Clarity

Traditional Plots: Traditional dependence plots can be harder to interpret, especially for non-technical stakeholders, as they typically do not provide clear indications of the magnitude and direction of feature contributions. SHAP Dependence Plot: SHAP plots are designed to be highly interpret-able, with clear visual cues such as color gradients and bar lengths indicating the direction and magnitude of feature contributions. This makes it easier for users to understand complex interactions at a glance.

10.8 Advantage: Detailed Insights into Feature Interactions

Traditional Plots: While traditional dependence plots can show some level of interaction, they often do not provide detailed insights into how secondary features influence the primary relationship. SHAP Dependence Plot: By using color coding to represent a secondary feature's value, SHAP dependence plots provide a detailed view of how multiple features interact. This helps in understanding not just the primary feature's effect but also how it is modified by other variables, which is crucial for capturing complex dependencies.

In summary, the SHAP dependence plot offers significant advantages over traditional dependence plots from linear and non-linear models by providing detailed, quantitative, and interpret-able insights into feature interactions. Its ability to visualize multi-dimensional relationships, quantify feature contributions, and enhance model transparency makes it a superior tool for understanding and improving predictive models.

11 Interaction Summary Plot

The Figure 14 presented is a SHAP interaction summary plot highlights the combined effects of feature interactions on the model's predictions. Each cell in the grid represents the interaction between a pair of features, with the color and distribution of points indicating the nature and strength of these interactions. For example, the interaction between Glucose and Age, seen in the first column under "Age," shows how changes in glucose levels, combined with varying ages, impact predictions. The SHAP interaction values

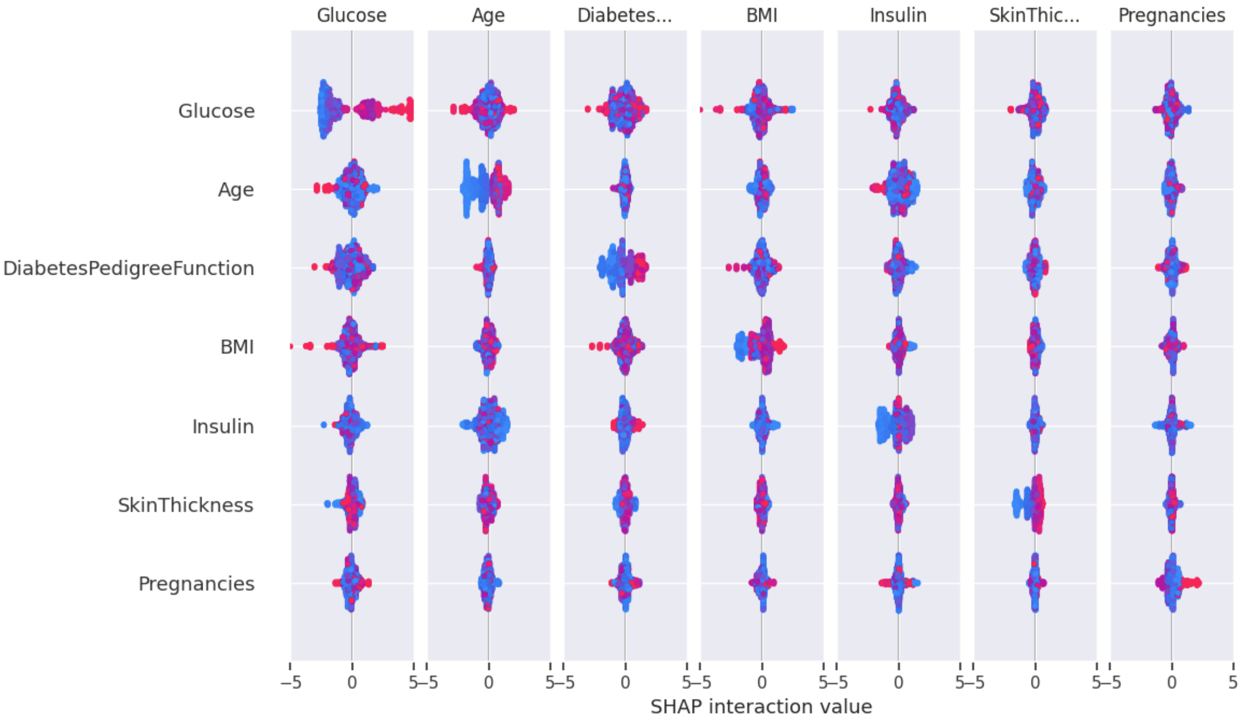


Figure 14: SHAP Values Summary Plot

along the x-axis indicate the magnitude of each interaction's impact, with points further from the center (0) showing stronger interactions and points near the center indicating weaker ones. The color gradient from blue to red represents the feature values, helping to understand how different value ranges contribute to the interaction effect.

The plot is particularly useful for identifying non-linear relationships and complex dependencies between features. For instance, the interactions between Glucose and other features like BMI and Insulin display distinct patterns, suggesting significant impacts on diabetes prediction. Additionally, some features, such as Glucose, consistently show strong interaction effects across various pairs, highlighting their critical role in the prediction model. This comprehensive view of feature interactions provides valuable insights into the model's behavior and enhances understanding of how different features work together to influence predictions.

11.1 Key Features: Interaction Effects

Dual-Feature Analysis: Each cell in the plot grid represents the interaction between two features. This dual-feature analysis highlights how pairs of features jointly influence the model's predictions, revealing interactions that single-feature analyses might miss. Visualization of Combined Impact: By displaying the SHAP interaction values, the plot quantifies the combined impact of feature pairs on the prediction. This is crucial for understanding complex dependencies within the model.

11.2 Key Features: SHAP Interaction Values

Magnitude of Impact: The x-axis in each subplot represents the SHAP interaction values, which quantify the magnitude of the combined impact of the feature pair on the model's prediction. Higher absolute values indicate a stronger influence. Direction of Influence: The direction in which the points spread along the x-axis indicates whether the interaction drives the prediction up or down, providing clear insights into how feature pairs affect the model's output.

11.3 Key Features: Color Coding

Multi-Dimensional Analysis: The use of color to represent the secondary feature's value (ranging from blue for lower values to red for higher values) adds a third dimension to the plot. This helps in understanding how variations in the secondary feature modify the primary feature's impact on predictions. Visual Clarity: The color gradient enhances interpretability by providing immediate visual cues about the value of the interacting features, facilitating a quicker and deeper understanding of their combined effects.

11.4 Key Features: Pattern Recognition

Identifying Relationships: The distribution and spread of points in each subplot help identify patterns and trends, such as linear or non-linear relationships between features. Recognizing these patterns is essential for understanding the underlying data structure and the model's behavior. Outlier Detection: Clusters of points or distinct gradients can highlight strong or consistent interactions, while outliers may indicate unusual or exceptional interactions that warrant further investigation.

11.5 Key Features: Comprehensive Overview

Holistic View of Model Behavior: By showing interactions for all feature pairs, the plot provides a comprehensive overview of how features interact to influence the model's predictions. This holistic view is essential for understanding the full complexity of the model's predictive mechanisms. Feature Importance Across Interactions: The plot allows users to see which feature interactions are most influential, providing a broader context for interpreting feature importance beyond single-feature effects.

11.6 Additional Benefits

Guidance for Feature Engineering: Insights from the plot can guide feature engineering efforts by identifying important interactions that could be combined into new features, potentially improving model performance. Enhanced Model Transparency: By visualizing feature interactions, the plot improves the transparency of the model, making it easier for stakeholders to trust and understand the model's predictions. Validation of Model Assumptions: The plot can help validate assumptions about feature interactions. If the observed interactions differ significantly from expected patterns, it may indicate areas where the model needs refinement. In summary, the SHAP interaction summary plot is a powerful tool for exploring and understanding the complex interactions between features in a predictive model. Its ability to visualize dual-feature effects, quantify interaction impacts, and highlight patterns and trends provides deep insights into the model's behavior, enhancing interpretability and guiding improvements in model development.

11.7 Comparison between SHAP Summary Plot and Traditional Linear and Non-Linear Model Summary Plots

The SHAP summary plot provides a unique and comprehensive way to visualize and interpret feature importance in machine learning models, distinguishing itself significantly from traditional summary plots used in linear and non-linear models.

- Traditional Summary Plots Linear Models (e.g., Logistic Regression): Feature Importance Representation: Linear models typically use coefficients to represent feature importance. These coefficients indicate the strength and direction of the relationship between each feature and the outcome. Interaction Visibility: Linear models generally assume that features have independent effects unless interaction terms are explicitly included. This assumption limits the visibility of feature interactions. interpretability: While the coefficients are straightforward to interpret, they do not capture complex relationships or interactions between features.

- Non-Linear Models (e.g., Random Forests, Gradient Boosting Machines):

  Feature Importance Representation: Non-linear models often use measures like mean decrease in impurity (Gini importance) or permutation importance to rank feature importance. These measures aggregate the impact of features across all splits in the trees. Interaction Visibility: While non-linear models can inherently capture interactions between features, traditional summary plots do not typically visualize these interactions explicitly. interpretability: The importance measures provide a high-level overview of feature importance but lack granularity and transparency in showing how individual feature values influence predictions.

- SHAP Summary Plot Feature Importance Representation:

  Global and Local Importance: SHAP summary plots provide both global feature importance (average SHAP values) and local importance (individual SHAP values for each prediction). This dual perspective offers a comprehensive understanding of feature influence. Quantitative Contributions: Each point in the SHAP summary plot represents a SHAP value for a specific feature and instance, showing the actual contribution of the feature to the model's prediction. Interaction Visibility:

  Detailed Interactions: SHAP interaction plots can explicitly show how pairs of features interact and their combined effects on the prediction. This is a significant advantage over traditional plots that typically do not highlight these interactions. Color-Coded Insights: The SHAP summary plot uses color coding to represent feature values, providing additional context on how different value ranges of a feature impact predictions. This multi-dimensional analysis is not available in traditional plots. interpretability:

  Clarity and Transparency: SHAP plots break down predictions into contributions from each feature, making complex models (often considered "black boxes") more transparent and interpret-able. Users can see how much each feature value contributes to the final prediction, enhancing trust and understanding. Pattern Recognition: The distribution of SHAP values helps identify patterns, trends, and outliers in the data, providing deeper insights into feature behavior. This level of detail aids in understanding non-linear and interaction effects that are not visible in traditional plots.

11.8 Advantages of SHAP Summary Plot

- Enhanced interpretability: SHAP values decompose a prediction into contributions from each feature, providing clear and understandable insights into how individual features influence the outcome. This is particularly beneficial for stakeholders who need to trust and understand model predictions.

- Visibility of Interactions: SHAP summary plots can visualize interactions between features, offering a more comprehensive view of the model's behavior. This is a significant advantage over traditional summary plots, which often overlook these interactions. Quantitative and Visual Insights:

  The SHAP summary plot provides both quantitative measures (through SHAP values) and visual cues (through color coding and distribution of points) to convey how features impact predictions. This dual approach enhances the depth and quality of insights. Robustness Across Models:

  SHAP values can be applied to any machine learning model, whether linear or non-linear, making them a versatile tool for model interpretation. This universal applicability ensures consistent interpretability standards across different types of models. Guidance for Feature Engineering:

  The detailed insights from SHAP plots can guide feature engineering efforts by identifying important interactions and contributions. This can lead to the creation of new features that better capture underlying patterns in the data, improving model performance. Model Debugging and Validation:

  SHAP plots help in debugging and validating models by highlighting unexpected patterns or biases. This transparency enables more effective troubleshooting and refinement of the model. In summary, the SHAP summary plot provides a richer, more detailed, and more interpret-able visualization of feature importance and interactions compared to traditional linear and non-linear model summary plots. Its ability to decompose predictions into individual feature contributions, visualize interactions, and offer clear, quantitative insights makes it a superior tool for understanding and improving machine learning models.

12 SHAP Impact Plot

Figure 15 is a Impact plot, it represents the impact of each variable or feature in a dataset on the prediction made by a model. It uses SHAP values, which measure the importance of each feature in making a prediction. The plot organizes each feature along the y-axis with the most important feature at the top, descending in importance down the y-axis. This plot is invaluable for interpreting the behavior of machine learning models, making it easier for data scientists to communicate how a model makes decisions, identify potential biases in the model, and understand which features are most important. This level of insight into the model's decision-making process aids in trust and transparency, particularly in critical applications like healthcare, finance, or public policy.

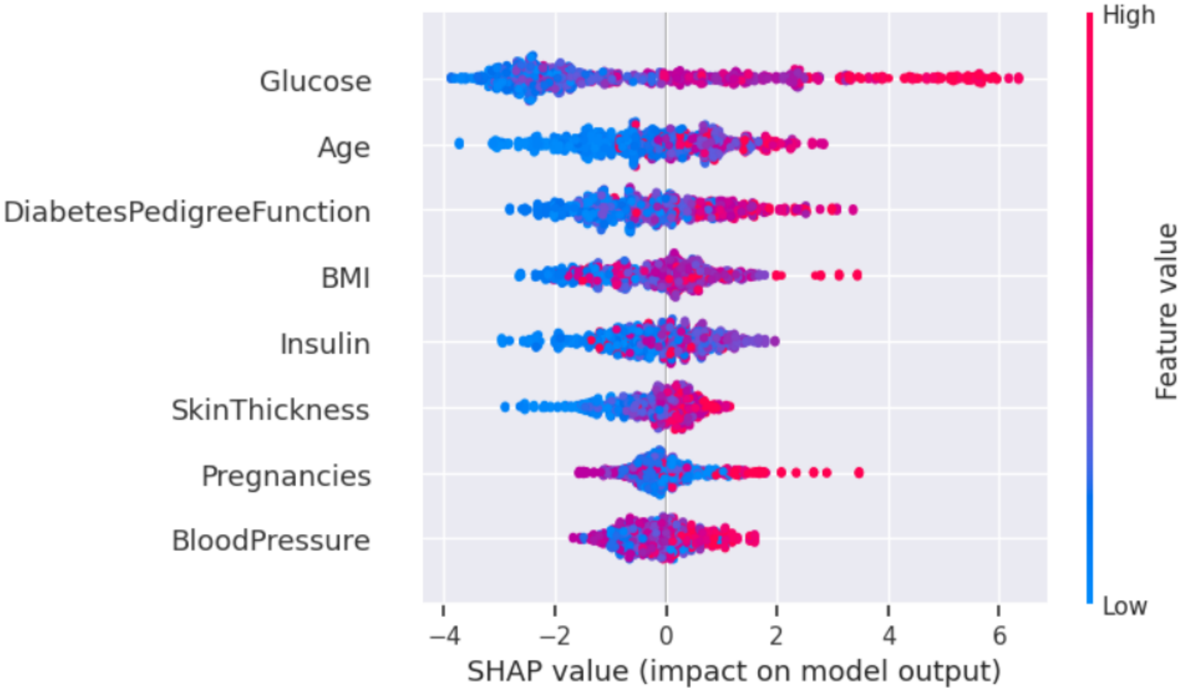


Figure 15: SHAP Value Impact Plot

12.1 Key Features: Impact Plot Analysis

Comprehensive Visualization of Feature Contributions: The SHAP Impact Plot offers a dynamic visualization that encapsulates the contributions of each feature to the model's predictions across all data points. Unlike simpler metrics that may only offer average impacts, this plot details the variability and strength of each feature's influence, providing a granular yet comprehensive view of feature behavior. This visualization is essential for deciphering the complexity of how input variables affect predictions, offering insights into the robustness and sensitivity of the model to changes in input data.

12.2 Delineation of Positive and Negative Feature Impacts:

By displaying the SHAP values on a horizontal axis, the Impact Plot distinctly illustrates how individual features push the model's output higher or lower. This directional impact is crucial for understanding which features drive particular outcomes, such as the likelihood of a disease in healthcare models or customer churn in business models. The color coding further enhances this analysis by linking feature values with their impact direction, allowing users to quickly ascertain the conditions under which features become more or less important.

12.3 Identification of Key Feature Interactions:

Beyond individual impacts, the SHAP Impact Plot can hint at complex interactions between features due to the overlay and clustering of SHAP values. For example, when two features collectively shift the SHAP value distribution significantly from the mean, it suggests a synergy or antagonism between these features in the context of the model's decision-making process. Understanding these interactions is vital for optimizing feature engineering and can guide more targeted data collection strategies to improve model performance and reliability.

12.4 Advantages over Linear Models: Quantification of Non-Linear Effects

Linear Models: Linear models inherently assume a linear relationship between features and the target variable. Plots typically show beta coefficients which directly represent the effect of a one-unit change in the feature on the target. SHAP Impact Plots: They reveal the actual impact of each feature across its entire distribution, including non-linear dependencies. SHAP values can show varying effects at different values of the feature, thus providing a more accurate and nuanced understanding of feature influence.

12.4.1 SHAP Impact Plot Advantages

SHAP Impact Plots transcend the capabilities of traditional plots used for linear and non-linear models by offering detailed, consistent, and model-agnostic visualizations. They allow stakeholders to scrutinize machine learning models to a degree that is unmatched by other methods, making them indispensable in scenarios where understanding the precise influence of features is crucial for decision-making, policy-setting, or regulatory compliance. This comprehensive view is particularly beneficial in complex, high-stakes environments like healthcare, finance, and personalized services.

12.5 Advantages over Non-Linear Models: Model Agnosticism

Non-Linear Models: Plots like partial dependence plots or individual conditional expectation plots are often model-specific and might not capture all complex relationships or interactions adequately. SHAP Impact Plots: These are model-agnostic, meaning they can be used across a variety of machine learning models, including both tree-based and deep learning models, while consistently providing detailed insights into the predictive mechanisms.

12.6 Advantages over Non-Linear Models: Consistency and Fairness

Non-linear Models often do not provide consistent insights across different subsets of data or different contexts, which can lead to misleading interpretations. In contrast, SHAP Impact Plots ensure consistency in feature attributions, which is vital for validating the fairness and bias of machine learning models. SHAP values adhere to properties like local accuracy and consistency, contributing to more reliable interpretations.

12.7 Advantages over Non-Linear Models: Consistency and Fairness

Non-Linear Models: Many plots do not provide consistent insights across different subsets of data or different contexts, which can lead to misleading interpretations. SHAP Impact Plots: Ensure consistency in feature attributions, which is vital for validating the fairness and bias of machine learning models. SHAP values adhere to properties like local accuracy and consistency, which contribute to more reliable interpretations.

13 Conclusion

This thesis has rigorously examined the comparative strengths of linear and non-linear modeling techniques, specifically focusing on logistic regression and Random Forest. While logistic regression, a linear method, offers remarkable clarity in its interpretability and stability in performance, it is the integration with SHAP values that significantly enhances this aspect, making the underlying model mechanics transparent and comprehensible. Conversely, Random Forest demonstrates superior accuracy, precision, recall, and F1-score, suggesting robustness in handling complex datasets prevalent in medical diagnostics.

The incorporation of SHAP values into these models transforms the interpretative capabilities by providing granular insights into how individual predictors like glucose levels or BMI influence diabetes predictions. This methodological enhancement facilitates a more nuanced understanding of model predictions, pivotal for crafting personalized patient care strategies and advancing medical research.

The practical applications of these enhanced modeling techniques are profound. In medical fields, where precise diagnosis and personalized treatment plans are crucial, the ability to interpret complex model outputs with SHAP values can lead to significantly improved patient outcomes. The clarity provided by these analytical enhancements assists healthcare professionals in understanding the intricate relationships between various health indicators and their impacts on disease prediction.

Furthermore, this work opens new avenues for future research. The integration of advanced machine learning techniques with interpretability tools like SHAP values promises to push the boundaries of what is possible in predictive analytics. It sets a precedent for the development of new models that can handle the complexities of modern datasets while remaining interpretable to practitioners. Researchers are encouraged to explore these methodologies further, potentially extending them beyond diabetes to other diseases where predictive accuracy and model transparency are equally critical.

In conclusion, this thesis not only underscores the effectiveness of combining SHAP values with traditional modeling approaches but also highlights the transformative potential of such integrations in enhancing the interpretability and applicability of predictive models in healthcare. The insights gained from this study are poised to make significant contributions to the fields of medical diagnostics and public health, ultimately leading to better clinical decisions and health outcomes.

14 Case Studies or Practical Applications

14.1 Practical Applications: Prescience, Enhancing Surgical Safety Through Interpretable Machine Learning

The "Prescience" system, discussed in a 2018 Nature Medicine article by Lundberg et al., represents a pioneering application of interpretable machine learning aimed at preventing hypoxemia during surgeries. This system utilizes advanced algorithms to predict hypoxemia risk and integrates SHAP (SHapley Additive exPlanations) values to provide actionable insights to anesthesiologists, enhancing transparency and decision-making [5].

14.1.1 Integration with SHAP Values

Prescience leverages SHAP values to explain its predictions, helping medical professionals understand which factors contribute to hypoxemia risk. This transparency is critical in high-stakes medical environments where trust in predictive analytics is essential.

14.1.2 Practical Benefits

The practical benefits of Prescience include:

- **Enhanced Patient Safety:** By predicting complications early, Prescience allows for preemptive interventions, improving patient safety during surgeries.
- **Operational Efficiency:** Real-time predictions optimize resource use, ensuring timely interventions.
- **Education and Training:** SHAP values help educate medical staff on risk factors, improving overall response strategies.

14.1.3 Implications for Further Research

The success of Prescience highlights the potential for broader applications of interpretable machine learning models in other critical care scenarios, such as predicting sepsis or post-operative complications. This supports the thesis's argument that machine learning and SHAP values can transform medical diagnostics and treatment planning.

15 Future Directions

The findings of this thesis lay a solid foundation for further exploration into the integration of advanced machine learning techniques with SHAP values, aiming to address unresolved challenges in medical data analysis. The successful application of these methods in predicting diabetes outcomes can be extended to other complex diseases, potentially revolutionizing diagnostics and treatment strategies across various medical specialties.

- **Broader Disease Application:** The methodologies validated for diabetes prediction can be adapted to study other chronic conditions such as cardiovascular diseases, Alzheimer's, and cancer. For each condition, specific models could be developed to understand the disease's progression and response to different treatments, supported by SHAP value interpretations to ensure that these models remain interpretable and trustworthy.
- **Real-Time Diagnostic Tools:** Future research could focus on implementing these advanced predictive models within real-time clinical decision support systems. Integrating SHAP-enhanced models into such systems could help clinicians make more informed decisions, offering explanations that align with clinical reasoning and patient-specific factors.
- **Enhanced Personalized Medicine:** By further refining the interpretability of predictive models, researchers can contribute to the field of personalized medicine. This involves tailoring treatment plans based on individual predictions and explanations provided by models, thus optimizing therapeutic effectiveness and reducing potential side effects.
- **Interdisciplinary Collaboration:** Encouraging interdisciplinary collaborations between data scientists, clinicians, and bioinformaticians could accelerate the development of these models. Such collaborations would ensure that the models not only achieve high accuracy but also are clinically relevant and easily interpretable by medical professionals.
- **Ethical AI in Healthcare:** Continued research into the ethical implications of applying machine learning in healthcare is crucial. Future studies should focus on ensuring that these models do not perpetuate biases or inequalities, with SHAP values providing transparency to detect and mitigate any potential biases in model predictions.

The potential to expand these research initiatives offers promising avenues for enhancing the effectiveness, safety, and efficiency of medical diagnostics and treatments. This approach not only broadens the scope of predictive analytics in healthcare but also deepens our understanding of disease mechanisms through more nuanced data interpretation.

15.1 Conclusion

Prescience illustrates the critical role of interpretable machine learning models in enhancing clinical practice and patient care. It validates the core premise of this thesis that integrating SHAP values into predictive models augments transparency and practical utility, paving the way for future innovations in medical diagnostics.

15.2 Expanding Machine Learning Applications in Medical Diagnostics

According to a comprehensive review by Ahsan MM, Luna SA, and Siddique Z on machine-learning-based disease diagnosis, the potential for machine learning in medical diagnostics extends across various disease domains [?].

15.3 Broadening the Scope of Disease Prediction

- **Cardiovascular Diseases:** Machine learning models are increasingly used to predict cardiovascular events with high accuracy by analyzing large datasets of patient records and imaging data. Integrating SHAP values into these models helps clinicians understand the risk factors contributing to each patient's prognosis, enabling personalized preventative strategies.
- **Cancer Detection and Prognostics:** Early detection of cancer significantly improves treatment outcomes. Machine learning techniques, particularly deep learning applied to imaging data, have demonstrated potential in identifying malignancies at early stages. SHAP values elucidate features most indicative of malignancies, aiding radiologists and oncologists in making informed diagnostic decisions.
- **Neurodegenerative Diseases:** For conditions like Alzheimer's and Parkinson's disease, machine learning can assist in early diagnosis and monitoring disease progression. SHAP values provide insights into which biomarkers or patient behaviors are predictive of faster progression, facilitating earlier intervention.

15.4 Enhancing Real-Time Diagnostic Tools

The integration of machine learning models into real-time clinical decision support systems represents a significant advancement, with SHAP values enhancing operational transparency by providing explanations for model predictions, crucial for clinical adoption:

- **Operational Transparency:** SHAP values ensure that these systems not only predict outcomes but also explain their predictions, increasing trust among healthcare providers and patients.
- **Dynamic Treatment Adjustments:** Real-time insights allow for immediate adjustment of treatment plans as new data becomes available, optimizing patient outcomes continuously.

15.5 Implications for Public Health and Policy Making

Advanced machine learning models equipped with interpretability tools like SHAP values can significantly influence public health strategies and policy making:

- **Public Health Monitoring:** Predictive analytics can identify potential outbreaks and health crises before they become widespread, allowing for preemptive action.
- **Resource Allocation:** By predicting disease trends and outcomes, policymakers can better allocate healthcare resources to areas of greatest need, improving overall healthcare efficiency.

References

[1] Marcilio, W. E., and D. M. Eler. "From Explanations to Feature Selection: Assessing SHAP Values as Feature Selection Mechanism." 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2020, doi: 10.1109/sibgrapi51738.2020.00053.

[2] Pore, N. "Healthcare Diabetes Dataset." 2023, www.kaggle.com/datasets/nanditapore/healthcare-diabetes/data. Accessed 9 June 2024.

[3] Ahsan MM, Luna SA, Siddique Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare (Basel). 2022;10(3):541. Published 2022 Mar 15. doi:10.3390/healthcare10030541

[4] Lundberg, S.M., and Lee, S.-I. (2017). ”A Unified Approach to Interpreting Model Predictions.” *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA.

[5] Lundberg, S.M., Nair, B.G., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T.L., Liston, D., Low, D.K., Newman, S., Kim, J.H., & Lee, S. (2017). Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. bioRxiv.