Project 3: Regression

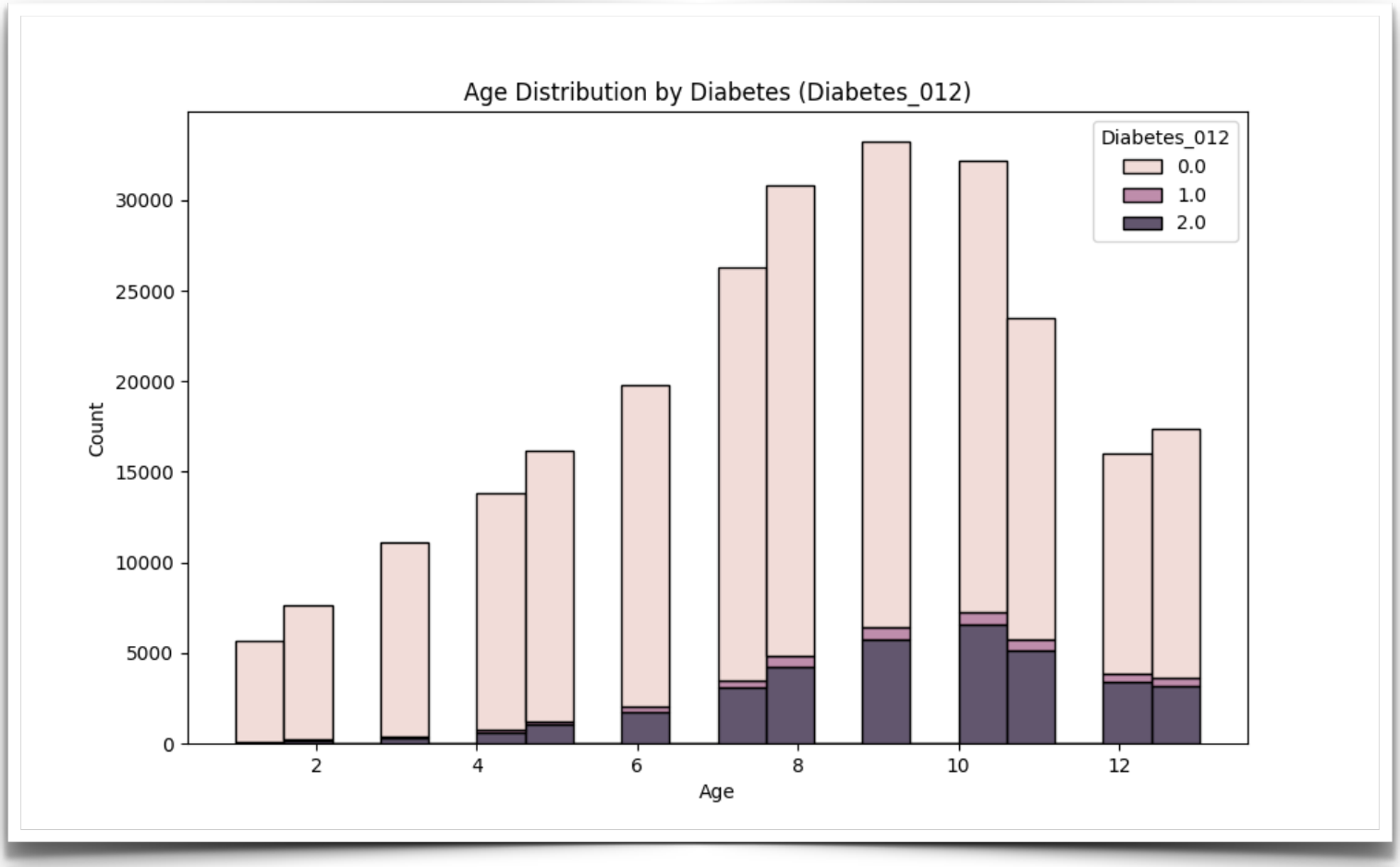
Ehu Shubham Shaw
Daniel Onyema
Roee Shraga
October 22, 2024

Task 1 – BYOD (Bring Your Own Data) [10 points]

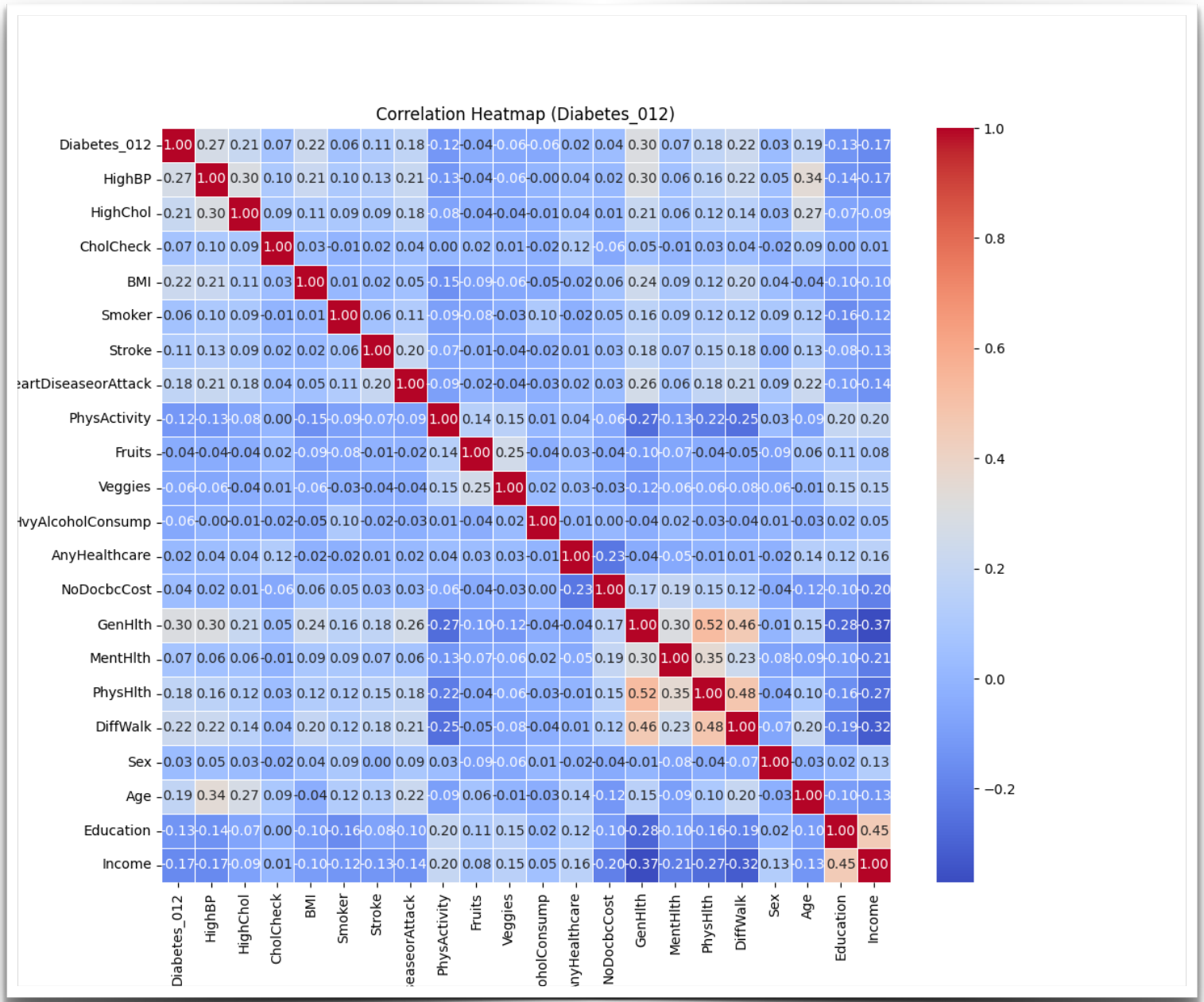
Diabetes is one of the most common chronic diseases in the United States, affecting millions of people each year and causing a huge financial burden on the economy. Diabetes is a dangerous, chronic condition in which people lose the ability to efficiently regulate glucose levels in their blood, resulting in a lower quality of life and a shorter lifespan. Various foods are broken down into sugars during digestion, which are then released into the bloodstream. This prompts the pancreas to release insulin. Insulin helps cells in the body use glucose in the bloodstream for energy. Diabetes is often characterized by either the body not producing enough insulin or being unable to use the insulin produced as well as needed. Secondly diabetes patients are more likely to develop complications such as heart disease, eyesight loss, lower-limb amputation, and kidney disease due to chronically high amounts of sugar in their system. While there is no cure for diabetes, efforts such as weight loss, healthy diet, physical activity, and medical therapy can help many patients avoid the negative effects of the illness. Early detection can lead to lifestyle adjustments and more successful treatment, making diabetes risk prediction models valuable tools for public and health officials.

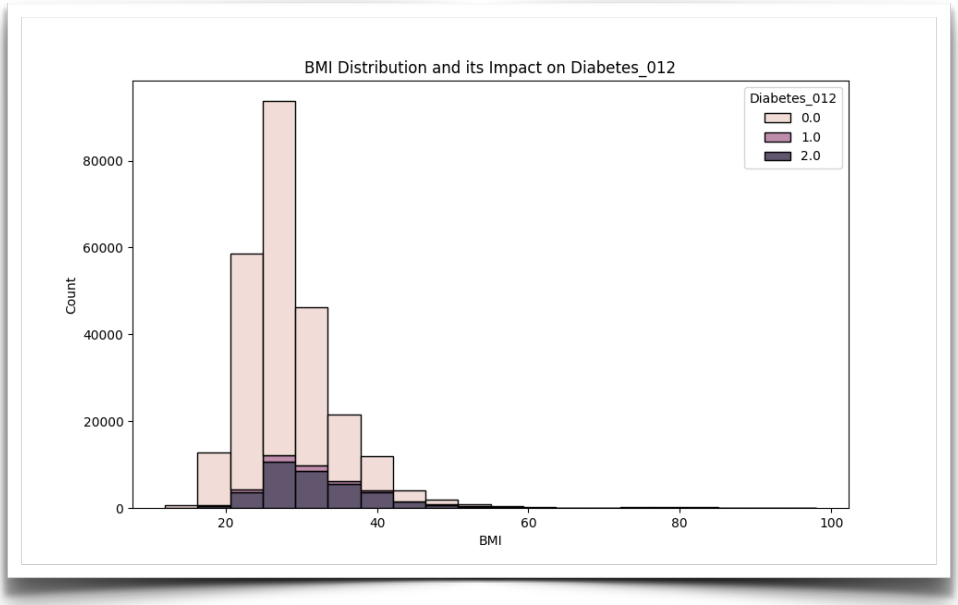
Task 2 – EDA [10 points]

1. Diabetes Distribution by Age: This graphic depicts the prevalence of diabetes among various age groups. It is useful to identify which age groups have a higher prevalence of diabetes, demonstrating that the risk of diabetes increases with age.



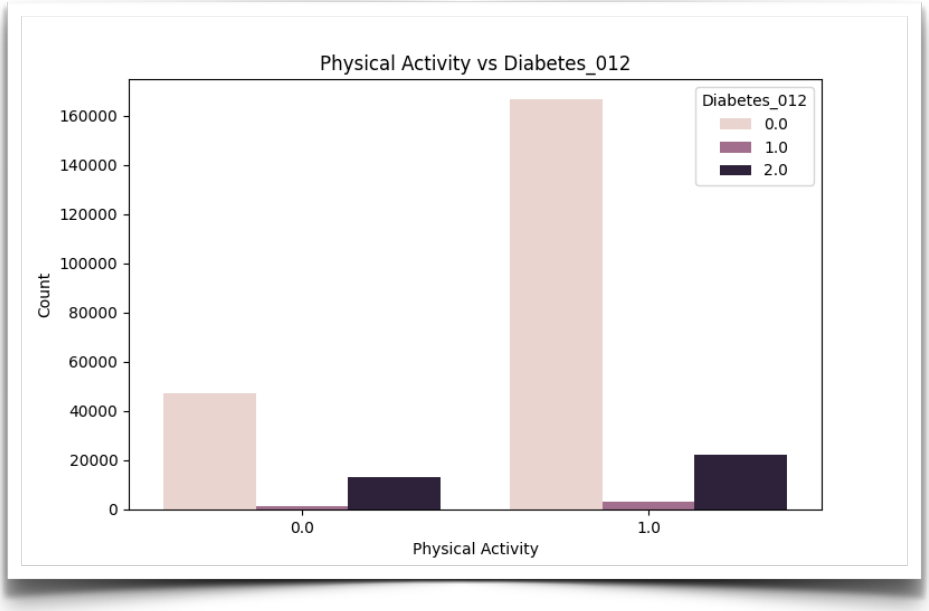
2. Correlation Heatmap: Defined as a visual representation of the correlation between dataset characteristics. Strong associations with diabetes (such as BMI or physical activity) are highlighted, providing insight into which health factors are most closely related to diabetes.

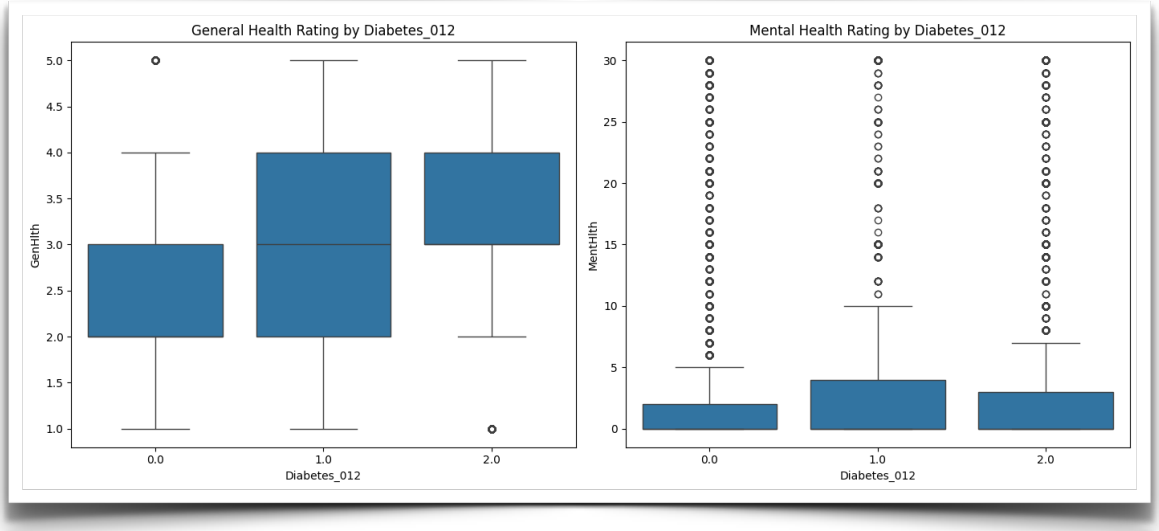




3. BMI Distribution and Diabetes Impact: This plot compares the distribution of BMI between those with and without diabetes. The distribution frequently shows that people with higher BMIs are more likely to have diabetes, showing a substantial link between weight and diabetes risk.

4. Physical exercise vs. Diabetes: This bar plot contrasts the prevalence of diabetes among those who engage in physical exercise vs those who do not. The study aims to determine whether regular physical exercise is associated with a lower incidence of diabetes.





5. Box-plots compare general and mental health evaluations between individuals with and without diabetes. Individuals with diabetes report poorer general health and emotional well-being, emphasizing the disease's broader health implications.

Regression

Task 3 – Problem Definition [10 points]:

Diabetes has evolved as one of the most common chronic diseases in the United States, affecting millions and providing considerable problems to public health. It is defined by the body's inability to adequately manage blood glucose levels, which can result in serious health issues such as cardiovascular disease, kidney failure, and vision loss. Diabetes has a significant financial impact on the economy, thus understanding its risk factors and promoting early detection measures are critical.

In this study, we hope to address the critical need for accurate diabetes risk prediction models. We will use a dataset containing crucial health characteristics such as Diabetes_012, HighBP, BMI, PhysActivity, and demographic factors such as Age and Income to investigate the correlations between these variables and diabetes prevalence.

Our exploratory data analysis (EDA) revealed numerous remarkable findings. For starters, higher Body Mass Index (BMI) levels have a substantial link with diabetes prevalence, emphasizing the need of weight management in diabetes prevention. Second, our findings show that people who engage in regular physical activity had lower incidence of diabetes, implying that lifestyle changes can play an important role in lowering risk. Furthermore, the findings show a disturbing trend in which older age groups are more likely to be impacted by diabetes, emphasizing the importance of targeted healthcare measures for aging populations.

The goal of this research is to create a reliable diabetes risk prediction model that can drive public health policy and empower people to make informed lifestyle decisions. We hope that by better understanding the interactions between various health and demographic factors, we may contribute to ongoing efforts to combat diabetes and enhance the quality of life for individuals affected.

Task 4 – Preprocessing [10 points]

1. *Data Cleaning & Data Transformation & Normalization:*

We verified data consistency across datasets by checking for null values using the pandas null function. There were no outliers or data inconsistencies. I split my dataset into training and testing sets using the train_test_split function from scikit-learn, which divided the data into two parts: one for training the model and another for testing it. After the data was cleaned, Min-Max normalization was used to scale numerical features (such as BMI and age) across all three datasets to a range of 0 to 1. This may aid by normalizing numerical features, which improves model convergence during training. As a result, this normalized data will be used for future model training and testing.

2. *Specific Data Type Processing*

Each dataset's categorical variables (for example, Sex, Smoker, and PhysActivity) were encoded using one-hot encoding to turn them into a numerical format suited for machine learning techniques. Many machine learning algorithms require numerical input, and one-hot encoding effectively represents categorical data without introducing ordinal relationships. Target Variable Identification svariable (Diabetes_012) was identified for supervised learning purposes, ensuring clear demarcation between features and labels in each dataset.

Task 5 – Model Selection, Training, and Optimization [20 points]

So we have used : 'Linear Regression','Random Forest','Gradient Boosting','KNN Regressor','Decision Tree'. We have apply GridSearchCV for hyper parameter tuning, cross-validation, and feature engineering techniques to improve performance. And we found that

Cross-Validation Results: MSE scores - FOR MODELS

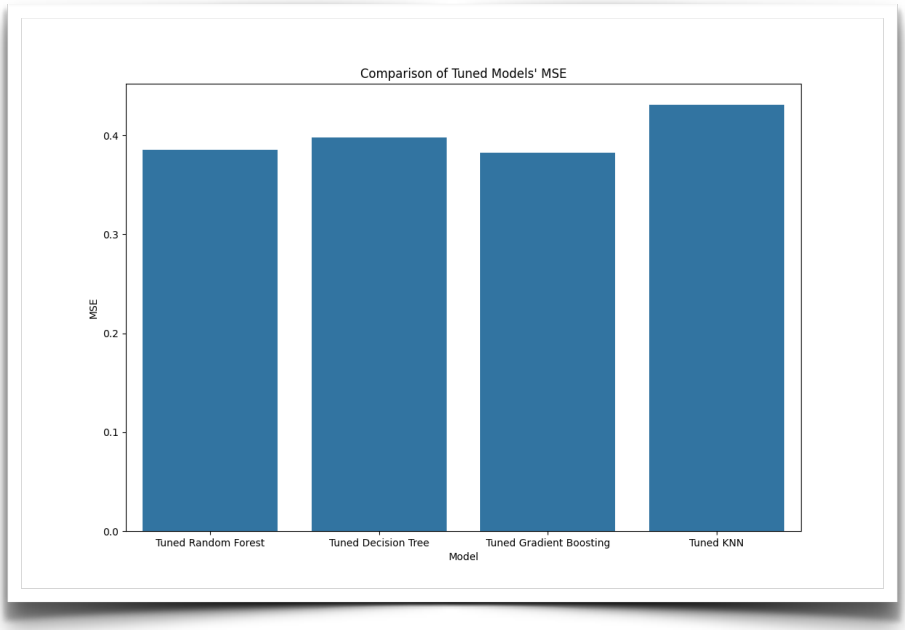
RandomForest: 0.43048062567289824,
DecisionTree: 0.8204766306645943,
GradientBoosting: 0.39021610820300145,
LinearRegression: 0.40426308944622064,
KNN: 0.46919640901308607

Five regression models were originally examined using cross-validation, with Mean Squared Error (MSE) serving as the major evaluation statistic. The findings are summarized below.• Random Forest MSE = 0.4305.• Decision Tree: MSE = 0.8205.• Gradient Boosting: MSE = 0.3902.• Linear regression with MSE = 0.4043.• KNN Regressor has an MSE of 0.4692.
These findings show that Gradient Boosting performed best during cross-validation, with the lowest MSE of 0.3902, followed by Linear Regression and Random Forest. On the other hand, Decision Tree performed the poorest, with the highest MSE of 0.8205, suggesting significant prediction error.

Hyper-parameter Optimization and GridSearchCV:

GridSearchCV was used to tune each model's hyper-parameters to boost its performance even more. Following adjustment, the models were tested again with MSE. The results of optimization are detailed below.

1. Random Forest: Initial MSE (Cross-validation) = 0.4305. Tuned parameters: n_estimators = 100 and max_depth = None. Optimized MSE to 0.4229. Tuning Results: The tweaked Random Forest model improved slightly in performance, lowering the MSE to 0.4229. This improvement shows that adjusting the hyper-parameters, such as the number of estimators and tree depth, enabled the model to better capture the data's patterns.



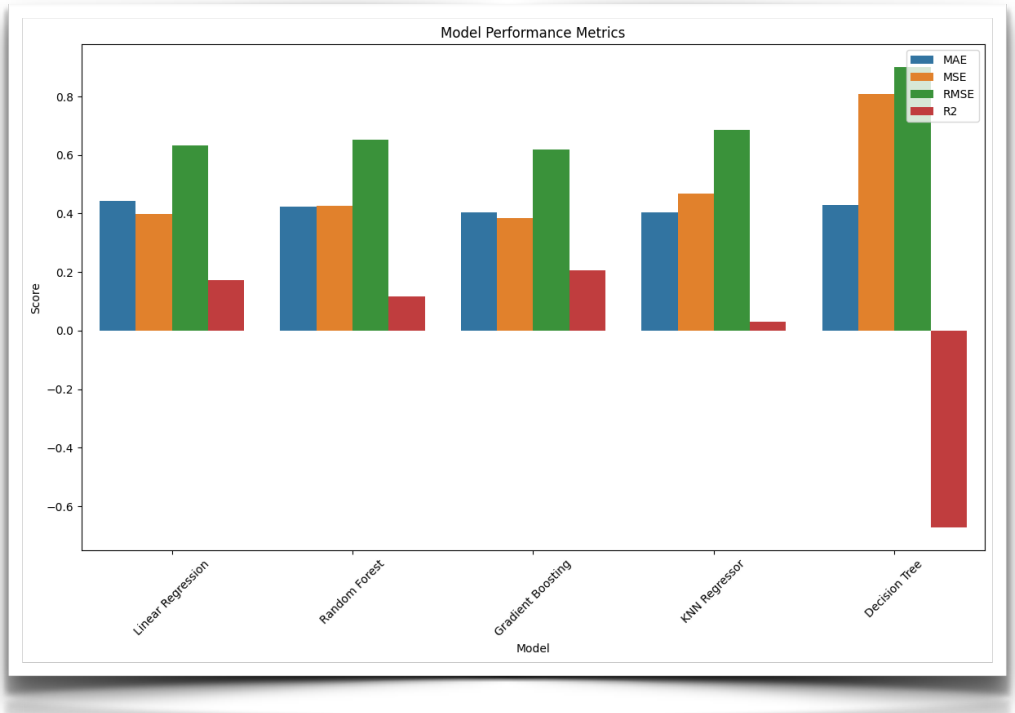
2. Decision tree: Initial MSE (Cross-validation) = 0.8205. Tuned parameters: max_depth = 10. Optimized MSE = 0.8113. tweaking Results: After tweaking, the Decision Tree model improved just slightly, with an MSE of 0.8113. The model still scored poorly, suggesting that a simple decision tree structure may not be adequate for this dataset.

3. Gradient-Boosting:• Initial MSE (Cross-validation) = 0.3902.• Tuned parameters include n_estimators = 100, learning_rate = 0.1, max_depth = 7, and subsample = 0.8.• Optimized MSE = 0.3856.
Tuning Results: The Gradient Boosting model, which was already the best performer, improved even more following hyperparameter tuning, lowering its MSE to 0.3856. This suggests that carefully adjusting parameters such as the learning rate, number of trees, and tree depth enabled the model to produce more accurate predictions.

4. KNN Regressor.• Initial MSE (Cross-validation) = 0.4692.• Tuned parameters: n_neighbors = 5, weights = distance, and p = 2.• Optimized MSE = 0.4628.
Tuning Results: The KNN Regressor improved slightly after tuning, with a lower MSE of 0.4628. However, the gain was minor, implying that KNN's performance is fundamentally limited in this context when compared to models such as Gradient Boosting.

Task 6 – Model Evaluation [20 points]

	Model	MAE	MSE	RMSE	R2
0	Linear Regression	0.442539	0.399754	0.632261	0.173310
1	Random Forest	0.424074	0.426963	0.653424	0.117040
2	Gradient Boosting	0.404781	0.384177	0.619820	0.205522
3	KNN Regressor	0.403433	0.469012	0.684845	0.030083
4	Decision Tree	0.425495	0.802983	0.896093	-0.660567



- Mean Absolute Error (MAE) is the average difference between projected and actual values. Lower values indicate improved performance.
- • The Mean Squared Error (MSE) calculates the average difference between anticipated and actual data. A lower MSE suggests higher accuracy.
- • Root Mean Squared Error (RMSE) is the square root of MSE and provides error in the same units as the target variable. Lower RMSE indicates improved performance.
- • R^2 (R-Squared): The amount of variance in the dependent variable that is predictable from the independent variables. Higher numbers imply a better fit for the model.

Cross-Validation Results:

Model	Cross-Validation MSE
Linear Regression	0.40
Random Forest	0.43
Gradient Boosting	0.39
KNN Regressor	0.47
Decision Tree	0.82

3. Observation and Analysis:

- Gradient Boosting was the best performing model, with the lowest MSE (0.38), RMSE (0.62), and greatest R^2 (0.21), explaining the most variance in the target variable. This demonstrates its ability to handle the data well after preprocessing.

- Effective Preprocessing and Optimization: Using StandardScaler to scale features increased model performance, particularly for KNN algorithms that rely on feature scaling. Optimization by hyperparameter adjustment (for example, GridSearchCV) boosted performance much further. This optimization helped models such as Random Forest and Gradient Boosting.

- The Decision Tree model underperformed, with an MSE of 0.81 and a negative R^2 (-0.67), indicating weak generalization and severe overfitting

Key benefits of using :

1. Mean Absolute Error (MAE):

MAE is useful for evaluating typical prediction error and is less sensitive to outliers than MSE or RMSE, making it an excellent general-purpose metric for direct comparison.

2. Mean Squared Error (MSE):

MSE is useful in cases where big prediction errors are particularly undesirable, as it increases the cost for such huge deviations. This sensitivity is important for detecting models that may occasionally cause substantial errors.

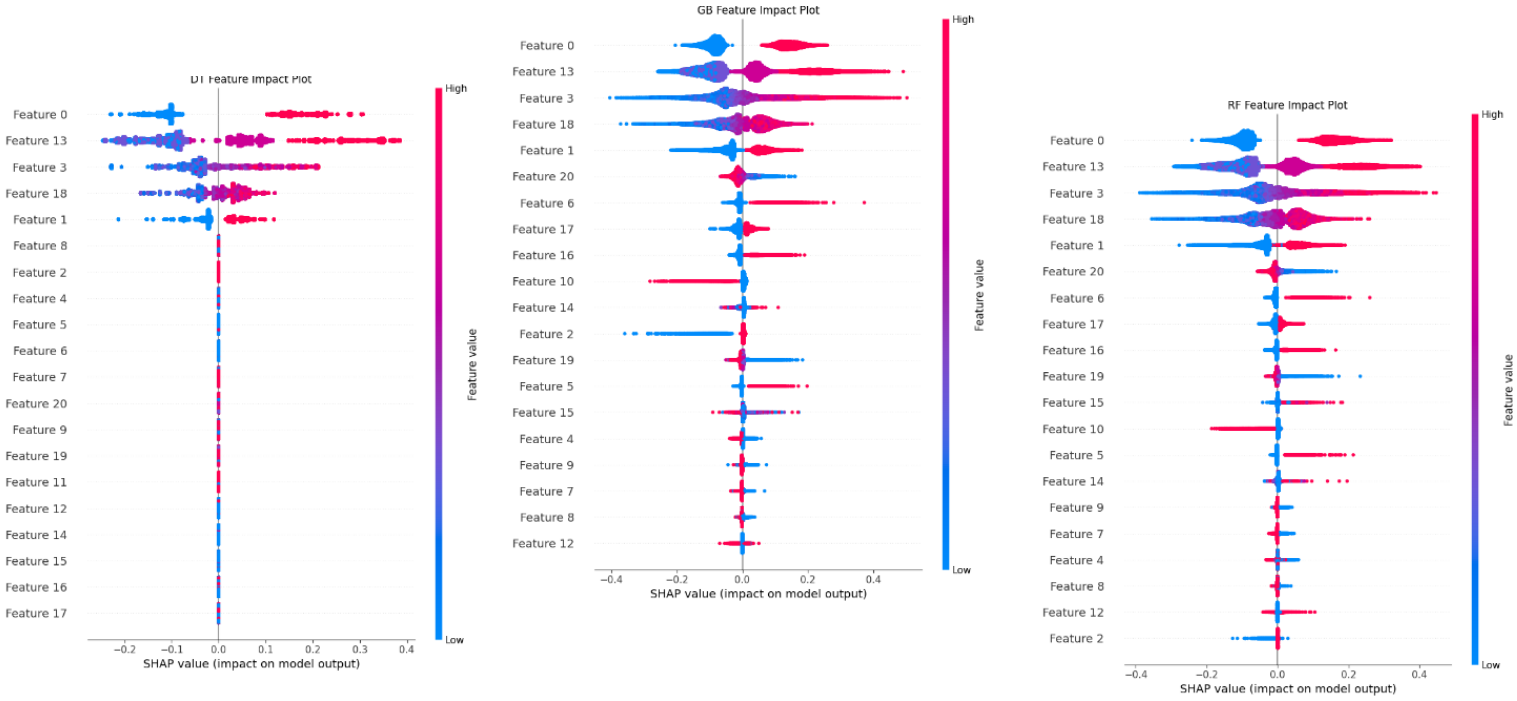
3. Root Mean Squared Error (RMSE):

RMSE is especially useful when you need an interpretable measure of the error magnitude in the same units as the output variable, and when significant errors must be punished more severely than minor ones.

4. R-Squared (R^2):

R^2 is beneficial for determining how well the model captures overall trends in the data and provides an intuitive view of the model's fit compared to a basic mean-based model.

Task 7 – Explainability [20 points]



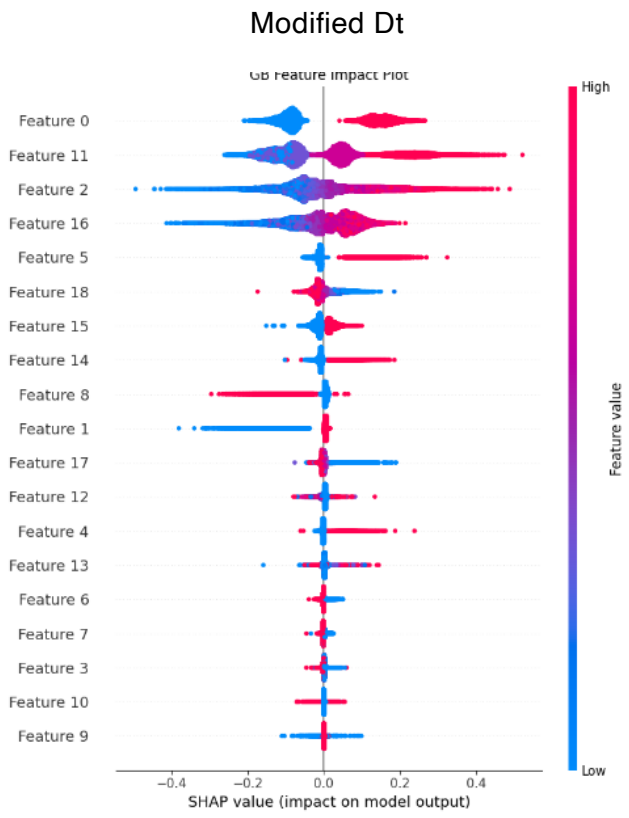
The **SHAP** summary graphs provide useful information on how specific features affected a tweaked model's ability to predict diabetes diagnosis. These plots show whether entries in the training set with higher or lower values for specific attributes influenced the model's predictions positively, negatively, or had no impact.

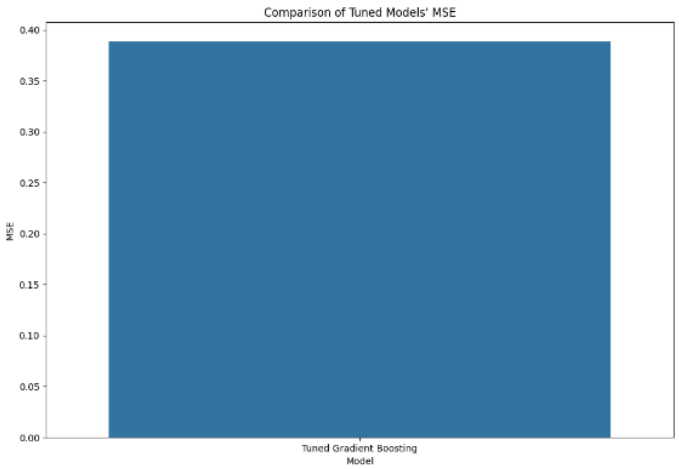
For example, assessing the best estimate for the Gradient Boosting (GB) model reveals that feature 13 is the most positively influential predictor. This feature is binary, with a value of 1 indicating that the patient was unable to see a doctor in the last 12 months owing to expense restrictions. In the SHAP value plot, all red entries to the right of the 0.0 line show occurrences when patients were unable to obtain medical care (i.e., feature 13 is "1"), and these events considerably improved the model's predictions. Conversely, the blue entries to the left of the 0.0 line correspond to occurrences where feature 13 is "0" (patients who could see a doctor), which had a negative impact on the model's predictions.

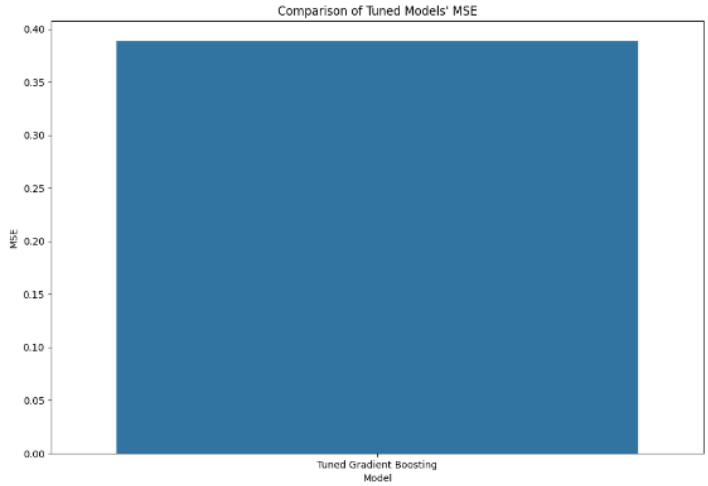
The divergence in SHAP values indicates a substantial positive link between patients with a "1" in feature 13 and the model's predicted performance. In contrast, entries with a "0" for this characteristic show a negative correlation with the model's ability to predict properly, indicating that those patients may have other health indicators that are less connected with diabetes risk.

Interestingly, not all traits have a clear correlation. For example, in the tweaked Random Forest (RF) model, feature 10—a binary indication of daily vegetable consumption—produces a distinct dynamic. Entries rated "0" for this attribute, suggesting that patients do not consume vegetables on a daily basis, had no significant effect on the RF model's predictions. This implies that a lack of vegetable consumption is not highly related to the model's predictive power. In fact, whereas entries with a "0" had no meaningful influence on the predictions, those with a "1" (representing daily vegetable consumption) showed a possible negative link with the model's ability to forecast accurately.

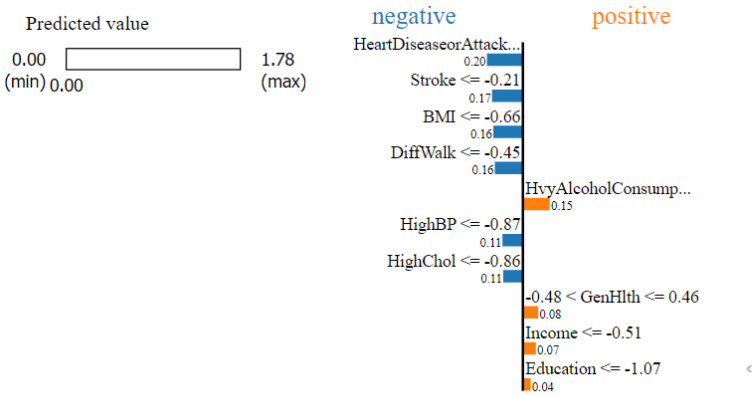
Overall, these findings illustrate the complexities of feature interactions in diabetes prediction models, demonstrating that the impact of key traits varies dramatically between models. Such findings not only improve our understanding of the model's decision-making process, but also highlight the significance of feature selection in designing successful diabetes prediction systems.



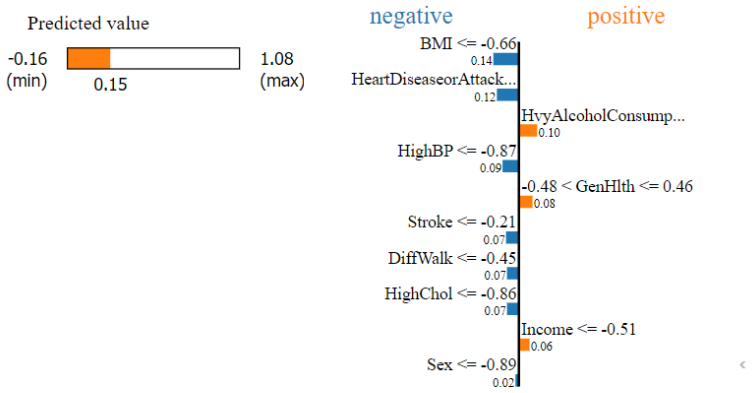




The **LIME** output for the Linear Regression model shows how each feature affected the model's forecast for this particular instance. The following is an interpretation of this output:
Overall Prediction:
The model's expected value for this instance is 0.15, as indicated by the orange bar in the upper left.
The prediction ranges from -0.16 (lowest) to 1.08 (highest).
Positive and negative contributions:
Positive (Orange): Orange features increase the prediction's value.
Negative (Blue): The blue features reduce the prediction's accuracy.



Feature	Value
HeartDiseaseorAttack	-0.32
Stroke	-0.21
BMI	-1.12
DiffWalk	-0.45
HvyAlcoholConsump	-0.24
HighBP	-0.87
HighChol	-0.86
GenHlth	0.46
Income	-1.96
Education	-1.07



Feature	Value
BMI	-1.12
HeartDiseaseorAttack	-0.32
HvyAlcoholConsump	-0.24
HighBP	-0.87
GenHlth	0.46
Stroke	-0.21
DiffWalk	-0.45
HighChol	-0.86
Income	-1.96
Sex	-0.89

Contributions for the Features
Each feature in the LIME output consists of:
Feature Name: The name of a feature in the dataset.
Value: The specific value of that feature for this instance (as given in the table to the right).
Impact on Prediction: The bar's length and associated number reflect how much this feature influences the forecast.
A detailed analysis of each feature's contribution.

Negative Contributions: (Left/Blue)
BMI (-1.12): BMI makes a significant negative contribution (-0.14) to the prediction, decreasing the expected outcome. A lower BMI may be related with a lower risk of the projected outcome.
HeartDiseaseorAttack (-0.32): This feature has a negative impact of 0.12, indicating that having a low or no history of heart disease reduces the model's prediction.

Other Health-Related Features: HighBP (-0.87), Stroke (-0.21), DiffWalk (-0.45), and HighChol (-0.86) all had moderate unfavorable influences, reducing the prediction by 0.07-0.09. The model reads lower scores as a decrease in the likelihood of the expected event.
Sex (-0.89): This feature contributes marginally negatively, implying that demographic factors play a minor effect in the model's predictions.

Positive Contributions (right/orange)
HvyAlcoholConsump (-0.24): Despite its low score, this characteristic adds a tiny positive push (0.10) to the prediction, showing that alcohol consumption is associated with a higher expected outcome.
GenHlth (0.46): With a moderate positive effect (0.08), the general health score indicates that specific health issues are positively correlated with the result in this case.
Income: (-1.96) Interestingly, even with a negative score, income has a somewhat positive impact (0.06), implying that lesser income may be associated with a higher likelihood of the expected outcome.

Main factors: The largest negative factors are BMI and HeartDiseaseorAttack, which reduce the anticipated score due to their low levels. In contrast, HvyAlcoholConsump and GenHlth are the most significant positive factors, boosting the prediction higher.
Overall Prediction: The final prediction of 0.15 is the outcome of a balance of positive and negative influences, with negative impacts being somewhat offset by positive factors, resulting in a low overall prediction value.
This LIME output shows the features the model ranks as most influential for this specific instance, providing transparency into the model's logic and assisting in interpreting its prediction in a human-readable manner.