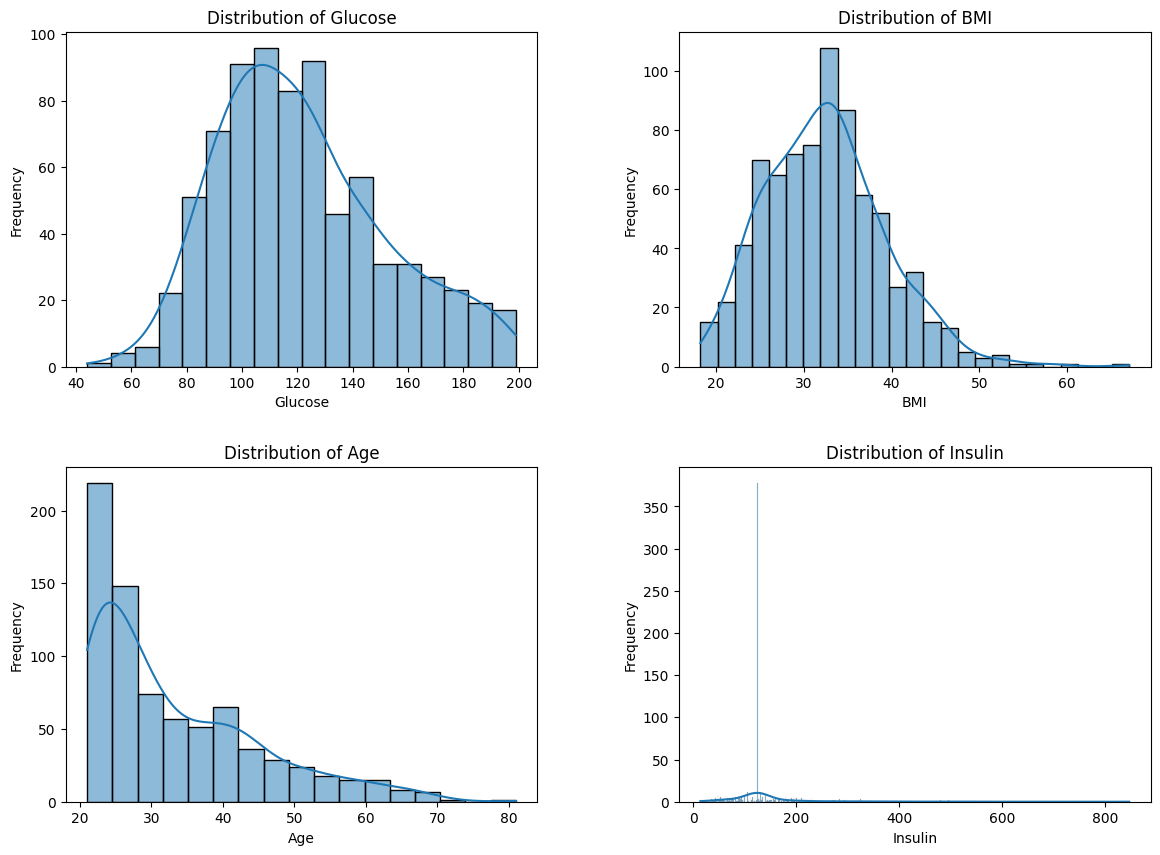
**Week 1   
Diabetes Prediction using Patient Data**

**Data Analysis**

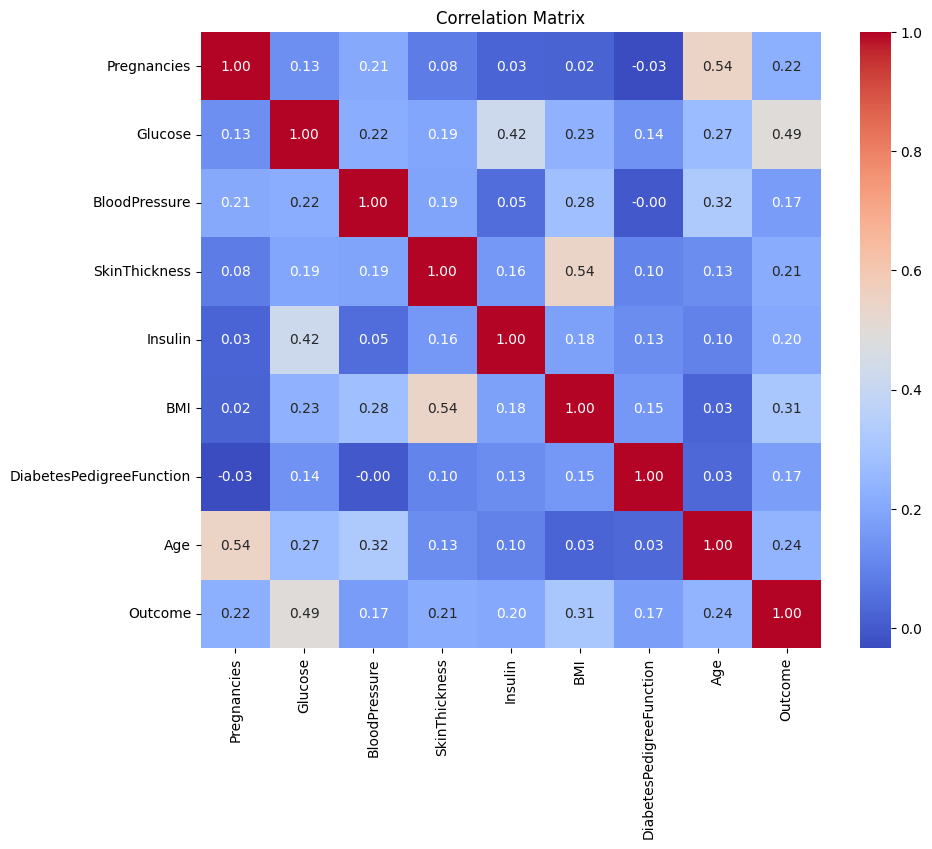
****

**Glucose:** The distribution appears to be slightly right skewed but relatively normal. Most values are clustered around the median, which is typical for a healthy population interspersed with diabetic individuals.

**BMI:** This shows a nearly normal distribution, slightly right skewed. The bulk of values center around the median, indicating a common range of body mass index among the participants.

**Age:** Age is right-skewed, which is expected in a general population dataset, showing more younger individuals than older.

**Insulin:** The distribution of Insulin is heavily right-skewed. Even after replacing zero values, it shows considerable variability, indicating different levels of insulin requirement or secretion among individuals.



**Glucose and Outcome**: There is a relatively strong positive correlation (0.47), suggesting that higher glucose levels are associated with a higher likelihood of diabetes.

**BMI and Outcome**: Also shows a moderate positive correlation (0.31), indicating that higher BMI may increase the risk of diabetes.

**Age and Outcome**: Moderate positive correlation (0.24), which could imply that the risk of diabetes increases with age.

**Pregnancies and Outcome**: Shows a correlation of 0.22, which might reflect the risk associated with gestational diabetes and its impact on overall diabetes risk later.

**Insulin and Skin Thickness**: These do not show very strong correlations with the outcome but are linked with each other and with other factors like BMI and glucose.

1. **Using Logistic Regression**

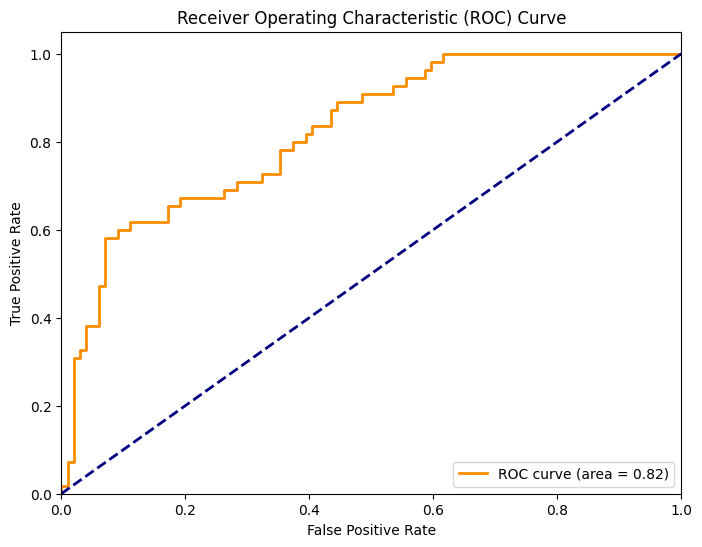
**Model Performance**

**Optimized Logistic Regression Model Results**

The best parameters from the grid search are:

* Regularization Type (Penalty): L1 (Lasso)
* Regularization Strength (C): 1
* Solver: liblinear

**Accuracy:** 0.77

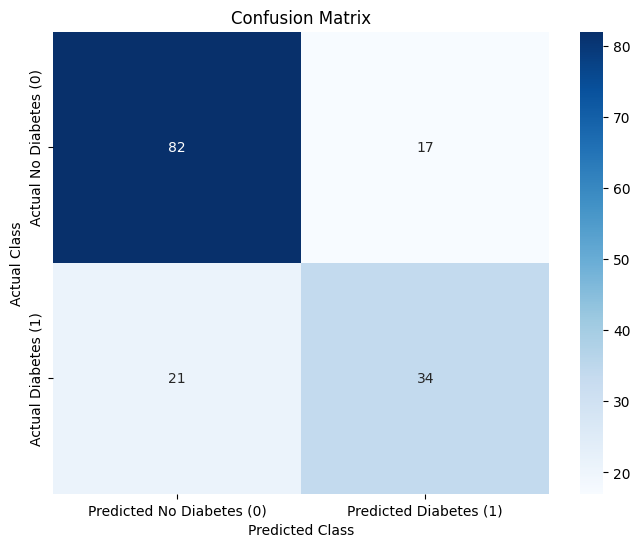
****

The ROC curve with an area of 0.82 indicates a higher performance. An AUC of 0.82 suggests that the model has a very good ability to distinguish between the classes effectively, which is a significant improvement.

The ROC curve, plotted above, shows the trade-off between the true positive rate (TPR) and false positive rate (FPR) at various threshold settings.

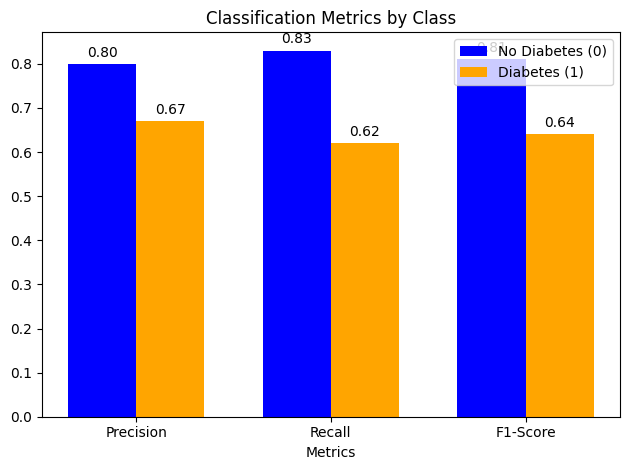
An AUC of 0.82 indicates a good discriminatory ability of the model to differentiate between the diabetic and non-diabetic cases. It's not perfect, but it shows reasonable predictive power.

The curve is above the diagonal line (which represents a random guess), showing that the model provides a better prediction than random guessing.



* **True Negatives (TN)**: 82 - The number of correct predictions that an individual does not have diabetes.
* **False Positives (FP)**: 17 - The number of incorrect predictions that an individual has diabetes when they do not.
* **False Negatives (FN)**: 21 - The number of incorrect predictions that an individual does not have diabetes when they do.
* **True Positives (TP)**: 34 - The number of correct predictions that an individual has diabetes.

This matrix provides insights into the types of errors the model is making and its accuracy in predicting each class. It also highlights the model's balance between sensitivity (ability to detect positive cases) and specificity (ability to detect negative cases).



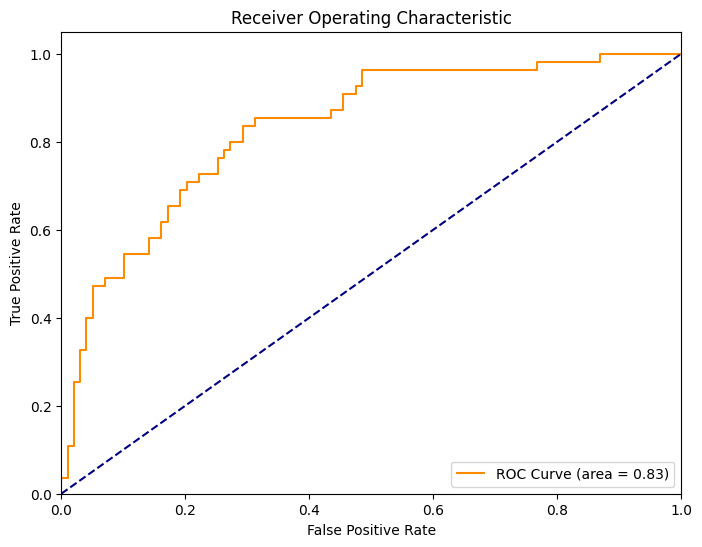
1. **Using Random Forest**

**Model Performance**

Grid Search Best Optimizers

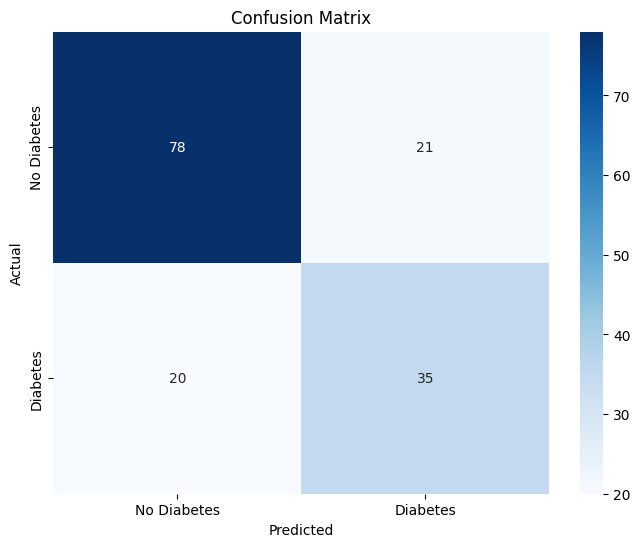
* max\_depth: None
* max\_features: log2
* min\_samples\_leaf: 1
* min\_samples\_split: 10
* n\_estimators: 300

**Best Model’s Accuracy:** 0.75



**High AUC**: An AUC of 0.83 suggests that the model has a high degree of separability between the positive and negative classes. This is a good sign, especially for medical diagnostic tests where the cost of false negatives can be high.

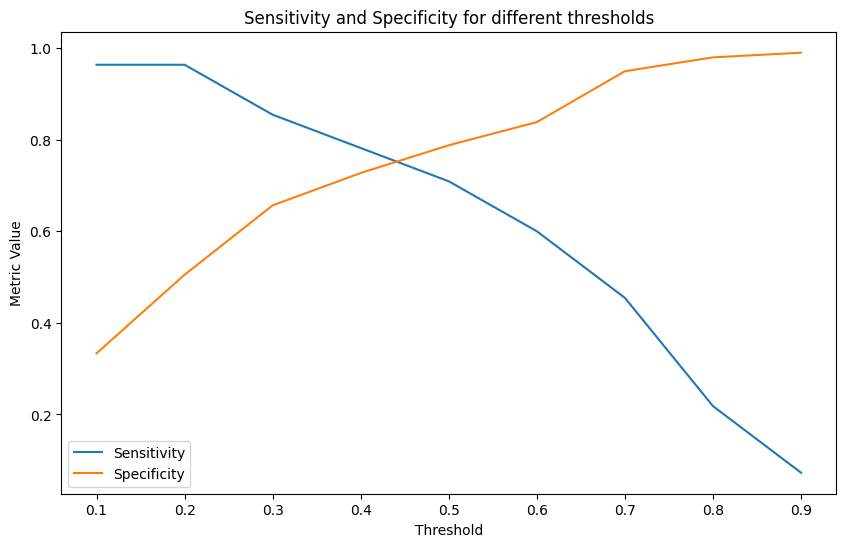
**Curve Position**: The ROC curve is significantly above the line of no-discrimination (the diagonal dashed line), which illustrates that the model predictions are not random and have a substantial predictive capability.



The sensitivity indicates that the model can identify 63.6% of all actual diabetic cases correctly.

The specificity shows that the model correctly identifies 78.8% of all non-diabetic cases.

Adjusting the decision threshold can help you fine-tune your model's performance, particularly balancing between sensitivity (recall) and specificity. By modifying the threshold used to classify predictions as positive (diabetes) or negative (no diabetes), you can potentially increase the model's ability to correctly identify more positive cases, albeit possibly at the cost of increasing false positives.



**Sensitivity (Recall)** starts high and decreases as the threshold increases. This indicates that lowering the threshold helps capture more true positives but at the risk of increasing false positives.

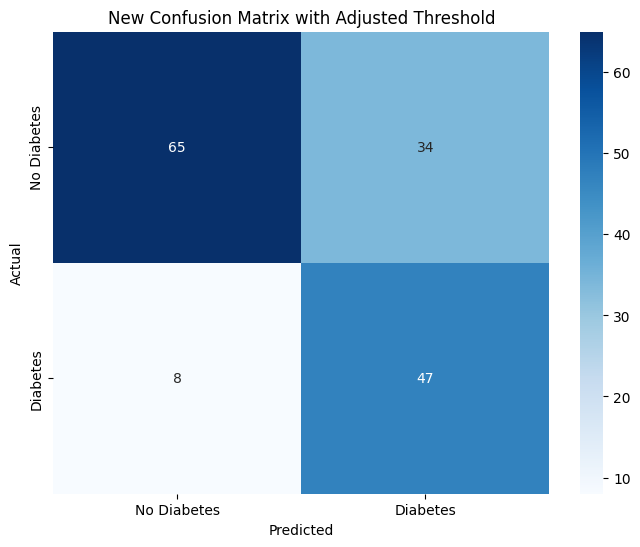
**Specificity** increases with the threshold, meaning higher thresholds reduce false positives but also miss more true positives.

**Selecting the Optimal Threshold**

The optimal threshold depends on the relative importance of sensitivity versus specificity, which can vary based on the application:

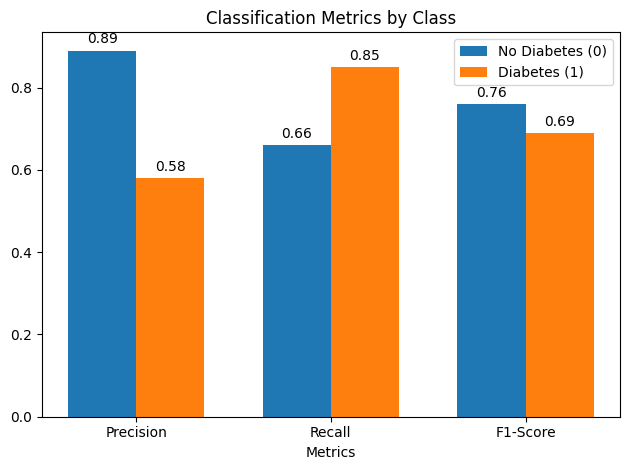
* For medical applications like diabetes prediction, a higher sensitivity might be preferred to ensure that most diabetic cases are captured, even if it results in more false positives (lower specificity).
* The crossover points around the threshold of 0.5 seems to be a balance, but if missing a diagnosis has serious consequences, you might consider a lower threshold.

**Threshold around 0.3 to 0.4**: This range might be a good starting point if increasing sensitivity is a priority. It maintains specificity above 60% while keeping sensitivity relatively high.



**Increased Sensitivity (Recall):** The model now correctly identifies more true positives (diabetic cases), with a decrease in false negatives. This suggests that the model is now better at catching diabetic cases, which is crucial for clinical applications.

**Decreased Specificity:** The increase in false positives indicates a loss in specificity. This shift means more non-diabetic patients are incorrectly identified as diabetic.



1. **Using SVM**

**Model Performance**

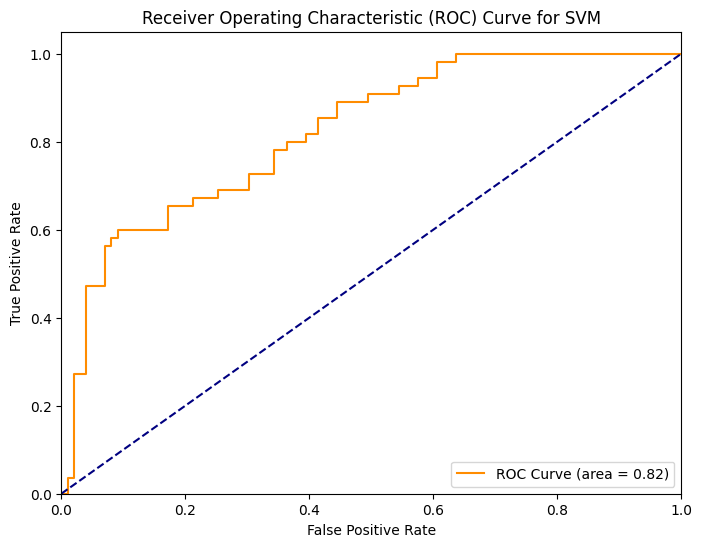
Using Grid Search, best parameters with Classification Report are:

**Best SVM Model Parameters:** {'C': 1000, 'gamma': 0.01, 'kernel': 'rbf'}

**Best Model Accuracy:** 0.7597402597402597

**Classification Report:**

|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| --- | --- | --- | --- | --- |
| **Class 0 (No Diabetes)** | 0.82 | 0.81 | 0.81 | 99 |
| **Class 1 (Diabetes)** | 0.66 | 0.67 | 0.67 | 55 |
| **Macro avg** | 0.74 | 0.74 | 0.74 | 154 |
| **Weighted avg** | 0.76 | 0.76 | 0.76 | 154 |

****

**Observations from the ROC Curve:**

1. **AUC Score:** The ROC curve shows an AUC (Area Under Curve) of 0.82, which indicates that the model has a good ability to discriminate between the two classes (diabetic and non-diabetic). The closer the AUC score is to 1, the better the model's performance. An AUC of 0.82 suggests that the model performs well but there is still some room for improvement.
2. **Curve Shape:** The curve is significantly above the diagonal dashed line, which represents a random classifier. This is a positive sign, showing that the model's predictions are far better than random guessing. As the true positive rate (TPR) increases, the false positive rate (FPR) also increases, indicating that the model is correctly identifying more positive cases (diabetes) as we reduce the threshold.
3. **Classifier Performance:** The model does a relatively good job of distinguishing between the diabetic and non-diabetic classes, as evidenced by the curve rising quickly towards the top-left corner of the plot.
4. **Impact of Threshold:** The curve illustrates how varying the decision threshold impacts the balance between True Positives (correctly identified diabetic cases) and False Positives (healthy patients misclassified as diabetic). A higher threshold will increase specificity but may lower sensitivity, while a lower threshold might capture more diabetic cases at the cost of false positives.

**Key Takeaways:**

* **True Positives (TP):** 37 diabetic patients were correctly predicted as diabetic, which is a good outcome, but could be improved further.
* **False Positives (FP):** The model made 19 false positive predictions, meaning it incorrectly identified some healthy patients as diabetic. This could be reduced by adjusting the threshold.
* **False Negatives (FN):** The model missed 18 diabetic patients, which is concerning. A false negative means the model failed to identify some diabetics, which could have serious implications in a real-world scenario.
* **True Negatives (TN):** 80 healthy patients were correctly predicted as non-diabetic, which indicates the model is doing well at recognizing non-diabetic cases.

**Best Model Performance:**

Looking at the results it can be seen that using Logistic Regression (0.77) our accuracy is far better than SVM (0.75) and Random Forest (0.75). Though these results can be improved but using full potential from this dataset and just for now it can be seen Logistic Regression is better.