**Q1: Which Model was used as a solution to our problem?**

A1: We chose logistic regression for our binary classification task because it is well-suited for predicting a binary outcome—in this case, the presence or absence of heart disease. Logistic regression is effective when the dependent variable is binary, as it estimates the probability of a given outcome using a logistic function, mapping the results between 0 and 1. This approach allows us to interpret probabilities directly, which is crucial for understanding the likelihood of heart disease occurrence. Logistic regression also provides clear insights into the relationship between independent variables (such as age, smoking status, and physical health metrics) and the dependent variable (heart disease status). By assigning weights to each independent variable, logistic regression highlights which factors are most influential, making it an interpretable and reliable choice for this health-focused application.

To address the imbalance, we used undersampling on certain datasets to equalize the classes. The encoded datasets with 10, 20, and 37 features were also re-balanced by undersampling to ensure the model didn’t favor the majority class.

**Q2: Which evaluation techniques worked best, and what insights did you gain from them?**

A2: In our heart disease prediction project, we utilized several evaluation techniques to assess the performance of our logistic regression model. The primary metrics employed included accuracy, precision, recall, F1-score, and area under the curve (AUC). Each metric served a distinct purpose in evaluating the model's effectiveness, particularly in the context of class imbalance that is common in medical datasets.

Accuracy: While accuracy is a straightforward metric that indicates the overall proportion of correct predictions, it can be misleading in imbalanced datasets. In our case, the accuracy remained consistently high (around 0.91) across the three encoded datasets. However, this high accuracy was somewhat deceptive, as it masked poor performance in identifying heart disease cases. The model could predict the majority class (no heart disease) correctly, leading to inflated accuracy while failing to capture the minority class effectively.

Precision: Precision measures the proportion of true positive predictions among all positive predictions. It provides insight into the model’s reliability when it predicts heart disease. For instance, in the 10-feature encoded dataset, the precision was 0.52, indicating that when the model predicted heart disease, only about 52% of those predictions were correct. This metric highlighted the model's tendency to produce false positives, suggesting a need for improvement in distinguishing between the two classes.

Recall: Recall, also known as sensitivity, measures the proportion of actual positive cases that were correctly identified by the model. In the context of predicting heart disease, recall is critical because missing a diagnosis could have serious health consequences. The recall values for the encoded datasets were notably low (0.09 for the 10-feature dataset, increasing slightly for the 20-feature and 37-feature datasets). This indicated that while the model may have high accuracy, it was failing to identify a significant number of patients who actually had heart disease.

F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives. The F1-scores for the encoded datasets were relatively low (0.15 for the 10-feature dataset), reflecting the challenges posed by class imbalance. However, this score illustrated that there was substantial room for improvement in model performance, particularly in correctly identifying positive cases.

Area Under the Curve (AUC): AUC is a robust metric that assesses the model's ability to discriminate between classes across various probability thresholds. The AUC values increased from 0.80 for the 10-feature dataset to 0.84 for the 37-feature dataset, suggesting that adding more features improved the model’s classification ability. This increase in AUC indicated that the model was better at distinguishing between individuals with and without heart disease as more informative features were included. A higher AUC indicates that the model is less likely to misclassify instances, which is particularly valuable in medical diagnoses where accurate predictions are crucial.

Confusion Matrices: The confusion matrices provided detailed insights into the classification results for each dataset, breaking down the counts of true positives, false positives, true negatives, and false negatives. For example, the confusion matrix for the 10-feature encoded dataset showed 54440 true negatives, but only 491 true positives, highlighting the model's struggle to correctly identify heart disease cases amid a majority of non-cases. This visual representation underscored the class imbalance issue, prompting further investigation into data preprocessing methods like undersampling

From the undersampled datasets, the evaluation metrics demonstrated significant improvement:

Precision, Recall, and F1-Scores: After applying undersampling techniques, the performance metrics for the 10-feature undersampled dataset improved markedly, with an F1-score of 0.70. This indicates that the model became more reliable in its predictions, correctly identifying a higher proportion of actual heart disease cases. The precision increased to 0.74 and recall to 0.67, showing that the model was now able to balance identifying true positives while minimizing false positives effectively.

In summary, these evaluation techniques provided a comprehensive view of model performance, revealing the strengths and weaknesses of the logistic regression model in predicting heart disease. The insights gained highlighted the importance of choosing appropriate metrics, especially in scenarios with class imbalance, and underscored the necessity of employing data preprocessing strategies to enhance predictive accuracy.

**Q3: Did you learn something new about your dataset after expanding your experiments to include more evaluation metrics/techniques?**

A3: Yes, we learned important insights

Data Preprocessing: Undersampling: Balancing the dataset improved model performance. It highlighted that relying solely on accuracy can be misleading in imbalanced datasets, as the model made more accurate predictions after reducing the majority class.

Feature Selection: Using different feature sets (10, 20, and 37 features) showed that more features typically improved AUC scores. However, not all features contribute equally, indicating the need for careful selection to enhance predictive power without adding unnecessary complexity.

**Q4: Which evaluation metrics are more important/reliable for your application? Which are less reliable?**

A4: For our application, the following evaluation metrics are deemed more important:

Recall is particularly crucial in the context of health predictions, as it reflects the model's ability to identify true positive cases of heart disease. In scenarios where missing a diagnosis could have serious health implications, a high recall is prioritized over precision.

F1-Score balances precision and recall, making it a reliable metric for assessing model performance when dealing with class imbalances.

Conversely, metrics like accuracy were found to be less reliable, especially in imbalanced datasets. For instance, in the encoded datasets, the accuracy remained high (around 0.91) while precision and recall were considerably low, indicating that the model was biased toward the majority class (no heart disease).

**Q5: How do the evaluation metrics indicate the performance related to different classes?**

A5: The evaluation metrics provide insights into the model's performance for different classes as follows:

Accuracy gives an overall performance measure but can be misleading in imbalanced scenarios. High accuracy may conceal poor performance in the minority class.

Precision focuses on the positive class (heart disease), showing how many predicted positive cases are actually positive. A high precision indicates that when the model predicts heart disease, it is likely correct.

Recall measures the model's ability to capture all actual positive cases. A high recall indicates that the model effectively identifies most of the heart disease cases, which is critical for patient care.

F1-Score combines precision and recall, providing a single metric to evaluate the model's effectiveness, particularly in situations with class imbalance.

AUC provides a broader perspective on model performance by analyzing true positive rates against false positive rates across various thresholds. A higher AUC indicates better overall classification ability.