MAIS 202
Team #3
Anastasiia Nemyrovska, Shirley Ding, Michelle Sayeh, Lucie Shi

# MAIS 202 - PROJECT DELIVERABLE 2

**Problem Statement**

The goal of this project is to create a machine learning model that predicts the likelihood of heart disease in patients based on health indicators like cholesterol levels, age, and blood pressure. The dataset chosen provides comprehensive patient data with no missing values, making it ideal for accurate model predictions. By developing a predictive model, we aim to assist healthcare professionals in identifying high-risk patients efficiently.

**Data Preprocessing**

We are using the Heart Disease Prediction dataset, which includes labeled data indicating the presence or absence of heart disease. Since the data was clean with no missing values, we focused on preparing it for model input by:

- Label Conversion: Converting “Presence” and “Absence” in the "Heart Disease" column to binary labels (1 and 0).
- Feature Scaling: Applied standardization using StandardScaler to ensure that features like age and cholesterol are on the same scale.
- Class Balancing: Using SMOTE to handle class imbalance, ensuring the model doesn’t overemphasize the majority class.

The dataset contains X features (e.g., age, blood pressure) and a target variable y ("Heart Disease"), with a balanced sample count after SMOTE application.

**Machine Learning Model**

In Deliverable 1, we initially proposed using Logistic Regression due to its efficiency with smaller datasets and interpretability. We retained Logistic Regression after assessing that it aligned best with the dataset size and our need for transparency in understanding feature impacts.

- Framework and Tools: Implemented with scikit-learn and utilized tools like SMOTE for balancing.
- Model Architecture: The model is a simple logistic regression with one output layer, chosen to maintain simplicity and interpretability in predictions.
- Training/Validation/Test Splits: We split the data 80/20 for training and testing, respectively, which allowed us to maintain an adequate sample size for validation.
- Regularization and Hyperparameters: Logistic Regression's default settings were used, with the model fitted to scaled and balanced data.
- Validation and Challenges: The model was evaluated for overfitting/underfitting using test data and validated with metrics like precision, recall, and specificity.
- Challenges: Balancing classes was a key challenge due to the dataset's initial imbalance; however, SMOTE helped mitigate this.

**Preliminary Results**

Our evaluation metric is the confusion matrix, supported by precision, recall, specificity, and F1 score, which align with medical diagnostic needs.

- Confusion Matrix: Shows correct vs. incorrect classifications.
- Precision: Indicated a high proportion of accurate heart disease predictions.
- Recall and Specificity: High recall ensures minimal missed diagnoses, and high specificity avoids over-diagnosing.
- Graphs: A series of graphs (e.g., precision-recall curve, F1 score distribution) illustrate the model's performance, confirming that Logistic Regression is performing as expected for this data.
    - Confusion Matrix: We visualized the confusion matrix as a heatmap to highlight the true positives, true negatives, false positives, and false negatives. This will make it easy to see where the model is performing well and where it might be making errors. The result indicated that the model is performed well, with a high number of correct predictions and few false positives/negatives.
    - Precision-Recall Curve: Since this is a binary classification problem, a precision-recall curve helped us assess the trade-off between precision and recall

at various threshold levels. Our curve indicates a good balance between precision and recall across most thresholds.

- ROC Curve (Receiver Operating Characteristic Curve): An ROC curve plots the true positive rate (sensitivity) against the false positive rate, providing a visual summary of the model's performance across different threshold values. The area under the ROC curve quantifies overall performance, with values closer to 1 indicating a strong classifier. A steep initial curve suggests that our model has a strong ability to distinguish between positive and negative cases.

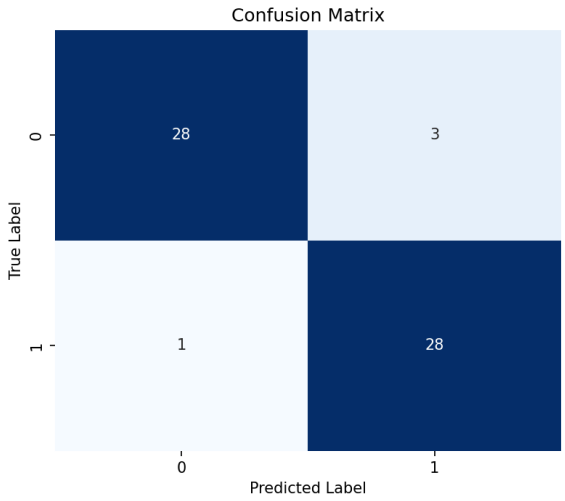


**Figure 1.** Confusion matrix visualization shows that 1) the model correctly predicted 28 cases where there was no heart disease; 2) the model incorrectly predicted 3 cases as having heart disease when there was none; 3) the model missed 1 case of heart disease; 4) the model correctly identified 28 cases with heart disease.

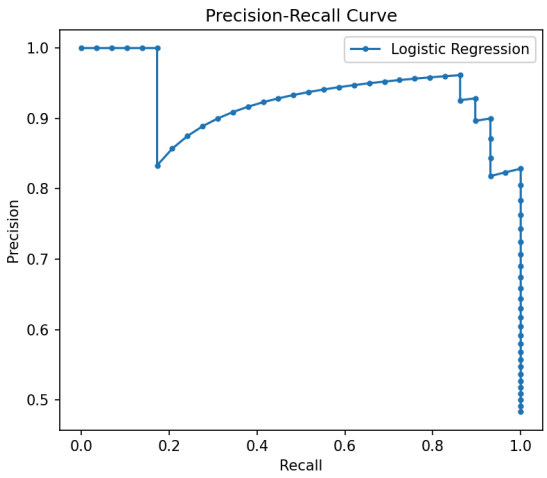


**Figure 2.** Precision-Recall Curve graph shows how precision and recall vary with different threshold values. In medical applications, a high recall is essential to avoid missing cases of heart disease, even if it slightly reduces precision.

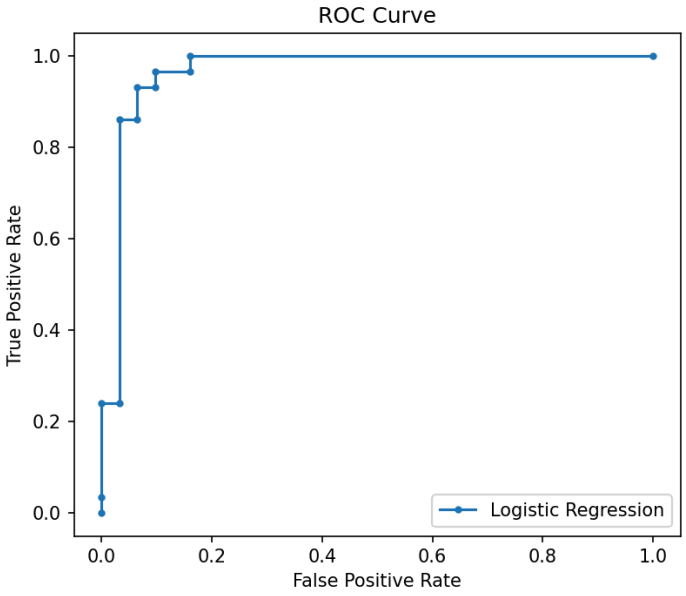


**Figure 3.** The ROC Curve shows the trade-off between the true positive rate (sensitivity) and the false positive rate at different thresholds.

- Feasibility: These results are promising for developing a practical prediction tool, though we may consider model refinement.

**Next Steps**

- Model Improvement: We may explore fine-tuning logistic regression's regularization parameter or trying alternative models like Random Forest if Logistic Regression underperforms in final testing.
- Pros and Cons: While Logistic Regression offers simplicity and interpretability, it may lack the predictive power of more complex models. Balancing this with the need for clear, interpretable predictions is a consideration.
- Future Work: Fine-tuning hyperparameters and exploring other resampling methods could enhance the model's generalizability.