Weekly progress report 2:

This week, we focused on implementing Linear Regression and Logistic Regression models to analyze and predict heart disease using machine learning techniques. We began by preprocessing the dataset, handling missing values, and splitting it into training and testing sets. Linear Regression was applied for continuous predictions, while Logistic Regression was used for classification. We evaluated model performance using various metrics, optimized hyperparameters, and addressed challenges like data imbalance. Our goal is to enhance model accuracy and explore advanced techniques for real-world application.

Implementation of the Linear Regression Model

For predicting continuous outcomes, we implemented a Linear Regression model following these key steps:

- Trained the model using Scikit-learn's LinearRegression class.
- Making predictions on the test set and evaluating the model's performance using:
    - Accuracy Score
    - Precision, Recall, and F1-score
- Plotted regression lines and residual errors to check for patterns or inconsistencies.

The model achieved an accuracy of 90.5%, with high precision and recall values, indicating strong predictive performance.

We observed that while Linear Regression captured some relationships, it struggled with categorical variables, indicating the need for alternative approaches for classification tasks.

Implementation of the Logistic Regression Model

Since our target variable was binary (presence or absence of disease), we implemented Logistic Regression, which is more suitable for classification tasks. The workflow included:

- Training a Logistic Regression model using Scikit-learn's LogisticRegression class.
- Making predictions on the test set and evaluating the model's performance using:
    - Accuracy Score
    - Precision, Recall, and F1-score
- Plotted an ROC curve and calculated the AUC score to assess classification performance.

The model achieved an accuracy of 96.5%, with high precision and recall values, indicating strong predictive performance.

Some key challenges included handling imbalanced data and improving model generalization. Moving forward, we plan to:

- Experiment with other machine learning models, such as Decision Trees and Random Forests.
- Conduct deeper feature selection analysis to improve model interpretability.

Challenges and Proposed Solutions:

Challenge: Fractional Predictions for Binary Targets

- Linear regression outputs continuous values, whereas binary classification requires discrete 0 or 1 values. Using a fixed threshold (e.g., 0.6 < y_pred < 1.2) for classification can be arbitrary and inconsistent.

Solution:

- Use Logistic Regression instead of Linear Regression, as it maps predictions to probabilities using the sigmoid function.
- Alternatively, apply a threshold-based classification (e.g., y_pred_binary = (y_pred > 0.5).astype(int)) for better consistency.

Challenge: Accuracy Fluctuations with Threshold Adjustments

- Changing the prediction threshold affects accuracy unpredictably.

Solution:

- Use techniques like ROC Curve and Precision-Recall analysis to determine an optimal threshold dynamically instead of manually selecting one.