Reflection

In this project, the primary design decision was to structure the model development in three stages: baseline, intermediate, and final. I began with **K-Nearest Neighbors (KNN)** as a baseline due to its simplicity and intuitive nature. Although its accuracy was low (48%), it provided an essential reference point against which more complex models could be compared. This decision ensured that improvements in subsequent models could be quantitatively justified.

Next, I selected a **Support Vector Machine (SVM)** as the middle model. The rationale was its strong performance in high-dimensional spaces and its ability to maximize class separation through margin optimization. This resulted in a significant performance jump to 85%. While SVM required more computational resources and careful feature scaling, it demonstrated the effectiveness of margin-based classifiers for this dataset.

Finally, I chose **Logistic Regression (LR)** as the top model. Despite its simplicity, LR offered interpretability, robustness, and surprisingly strong accuracy (96%). This decision emphasized the principle that complex models are not always superior; with proper preprocessing, a well-regularized LR can outperform more computationally intensive methods. Additionally, LR provided clearer insights into feature importance, which aligned with the project's goal of building not only an accurate but also an interpretable model.

Another important design choice was the emphasis on **accuracy as the primary metric**, given the relatively balanced dataset. However, I also considered failure modes such as class overlap and minority class misclassifications. This guided my decision to recommend LR for deployment while suggesting future exploration of ensemble methods or class imbalance handling.

Overall, the staged progression from KNN to SVM to LR allowed me to balance **simplicity, interpretability, and performance**, resulting in a well-justified and effective final model.