Project Summary

Team Member: Alan Xing, Chien-I Chao, Nicholas Rasmussen

In our machine learning project, we applied the techniques learned in class to the widely recognized and frequently used Iris dataset. We began by importing essential data processing and visualization libraries, such as pandas, numpy, matplotlib, and seaborn. The Iris dataset, a staple in the machine learning community, was loaded via sci-kit-learn and transformed into a pandas data frame, with target variable encoding for better interpretability. This also allowed us to visualize the dataset as a table with ease, providing insight into the structure of the data.

We then meticulously ensured the integrity of the data by checking for missing values. This was followed by a thorough normalization process using the Z-score method, which is instrumental in outlier detection. Visualization tools, such as scatter plots and heatmaps, were employed to comprehend the relationships between features and visually confirm the presence of any outliers in the dataset.

The preprocessing phase included splitting the data into an 80/20 ratio, increasing the training data fourfold, and introducing Gaussian noise. While techniques like PCA and LDA were not utilized in this project, they could be employed for additional data transformation before feeding the data into the SVM. The best kernel was evaluated using k-fold cross-validation on the training data, a rigorous process that involved testing multiple kernels and comparing their performance. The RBF kernel emerged as the most accurate, boasting a 97.5% accuracy rate across the folds. Other metrics were also gathered and displayed for all kernels, including precision, recall, and F1 score. This process gave us a clear picture of the best kernel for this data type.

In the final stage, the selected kernel was tested against the test data, and we achieved a perfect score, showing the effectiveness of our approach. Visualizing the confusion matrix provided a clear picture of the SVM model's classification performance, confirming the accurate identification of Iris species and revealing no misclassifications.

This study is a practical demonstration of applying fundamental machine learning concepts using a relatively clean dataset. It is noteworthy that real-world data often necessitates more exhaustive cleaning and preprocessing.