Project Summary

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

In this project, we are focusing on using those machine learning techniques and steps that we learned from the class to implement this Iris dataset.

For the first step, we imported those necessary libraries that will be needed for data processing and visualization, including pandas, numpy, matplotlib, and seaborn. Then, the Iris flower dataset is loaded using the scikit-learn library. Then, we start working on data preprocessing, the dataset is transformed into a pandas data frame for ease of manipulation. Target variable encoding is performed to enhance interpretability by mapping integer labels to corresponding feature names. This preprocessing step must be done to ensure the usability of the dataset for the analysis later.

The next steps what we would do is data splitting and checking for missing values to ensure data integrity. Subsequently, the dataset is normalized using the Z-score method to identify potential outliers. Visualization is also very important in the exploratory process. In our visualization section, scatter plots and correlation matrix heatmap let us better understand the relationship between different features (correlation between each petal length and sepal length in this case).

Follow after the visualization, we do the 80/20 data splitting and also doubling the training data and adding Gaussian noise to the training data. Then next we begin the evaluation and validation process, we implement K-fold cross-validation with shuffling, setting the value of K as 5, which divides the dataset into 5 folds. After that, we follow the standard way to validate the performance by using these evaluation metrics, including accuracy, precision, recall, and F1 score, this allows comparison of SVM model with different kernels. Some kernels, like Rbf in this case, can have higher accuracy(97.5%) compared to other kernels. Later, at the end we perform visualization such as a confusion matrix, to further analyze the classification performance of the SVM model. The confusion matrix shows the correct classification of Iris flower species by the model and identifies potential misclassifications.

The dataset we used for this Iris project was actually considered as a relatively cleaned dataset, so less time was spent on data cleaning and pre-processing. In real-world scenarios, data cleaning is generally much more complex and time consuming. However, for the Fundamental Machine Learning course, this Iris project is considered a good example for implementation.