Lab 4:

In this lab, I would like to select a dataset from python's Seaborn package. First, I checked the list of the dataset in the Seaborn package and looked for the information of most datasets in the list. Then I chose the dataset of penguins' information, including penguin species, regions, weights, genders, etc. I think this is an interesting dataset because we can analyze and use the model to predict penguin species through different factors. The penguin species in this dataset include Adelie, Chinstrap, and Gentoo. I will use some algorithms of data mining to make a decision tree model to predict penguin species and generate a classification report for the model.

First of all, I determined the correlation relation between different features and targets to analyze the deeper relationship between them from the results of data visualizations. According to the results, the penguin species of Adelie has the most data in the dataset, the penguin species of Gentoo has the highest mass weight and longest flipper length. Through the heatmap, I found that there was a small amount of missing data in this dataset, so I dropped all the rows with missing data and converted the category data of the species and islands column into indicator variables.

Next, I would like to split data and fit the decision tree model. In this dataset, features (X) are all the factors that may affect penguin species (target, y). I set all columns except the three species columns to features and set species columns to targets. After the dataset splits into test data and train data, the python helped fit the training data into the decision tree model. And then, we can obtain the parameters of the decision tree to better understand the model. The decision tree model analyzes whether the penguin belongs to a particular species through various factors. Then, we can use the test data features to make the prediction of the species and compare it with the actual species and generate a classification report. Through this classification report, I will understand the accuracy of this model in predicting penguin species.

According to the classification report, the accuracy of the model to predict penguin specie of Adelie, Chinstrap, and Gentoo are 98%, 92%, and 100%, respectively. In the results of recall, the model's ability to find all positive instances of Adelie, Chinstrap, and Gentoo was 96%, 100%, and 97%, respectively. The recall score of Chinstrap is 1, this means that the number of penguin species of Chinstrap predicted by the model is the same proportion as the actual number of Chinstrap. Also, the f-score shows that the accuracy of the three kinds of optimistic prediction is close to 1. The result of support means that the times of occurrence of the three kinds of predictions in the dataset are relatively balanced. It can be seen from these results that the prediction accuracy of the decision tree model is high. In addition, I also generate a classification report for the random forest model. The prediction accuracy of the random forest model may be higher than that of the decision tree model because the precision, recall, and f1-score of the model are closer to 1.

Moreover, I generated a classification report of the logistic regression model at the end of this lab. In the logistic regression model, I reset the target to specie "Adelie", so the model will analyze whether this penguin belongs to Adelie through different factors. The logistic regression model also got scores very close to 1 in the classification report. The results of the three classification reports show that different factors in the dataset are related to penguin species.