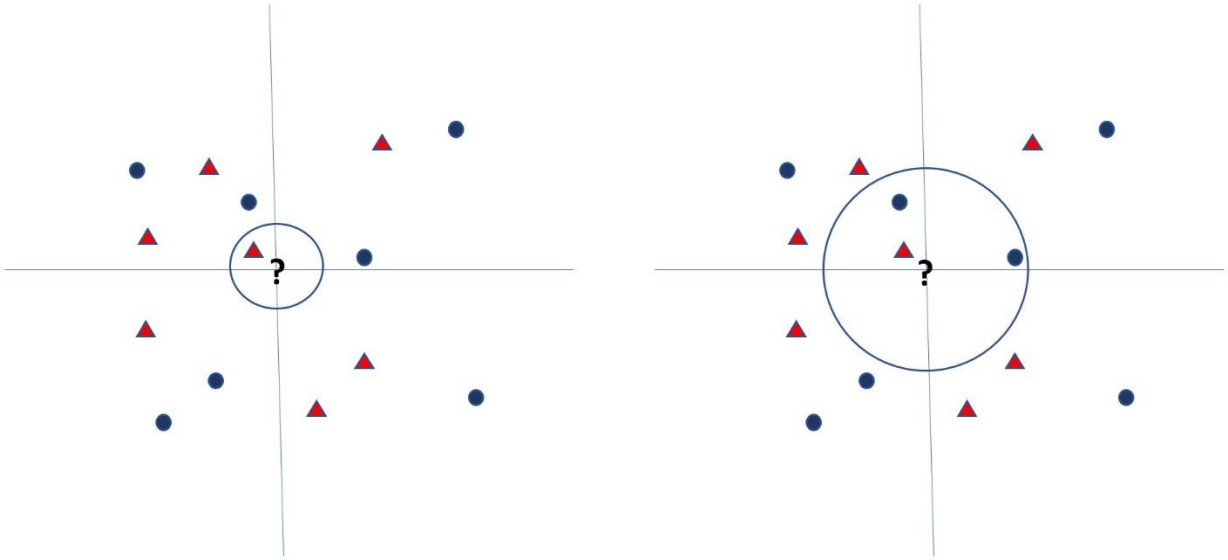
In this project, the in-built data set for iris is used that is split into train and test data for supervised machine learning. The randomly selected data will be used to train and evaluate a **k-nearest-neighbor(KNN) machine learning model** to be able to predict the type or class of the label of new data. This is an example of classification model.

Understanding KNN classification:

Given a known set of cases, a new case is classified by majority vote of the K (k= 1,2,3..) points nearest to the values of the new case; that is the nearest neighbors of the new case. In the example below are two features, the values are shown in the X and Y axis. There are two cases as represented by blue and red color, and each case has a value for the two features represented by the axes. KNN classifies cases with unknown label.

In the figure below, for k=1, the nearest neighbor to the case **'?'** is a red triangle thus, classified as a red triangle. But for k=3, the test case has two blue dots closest to it as compared to just one red triangle. Thus, with a different 'K' value, here k=3, the test case **'?'** is classified as blue triangle.



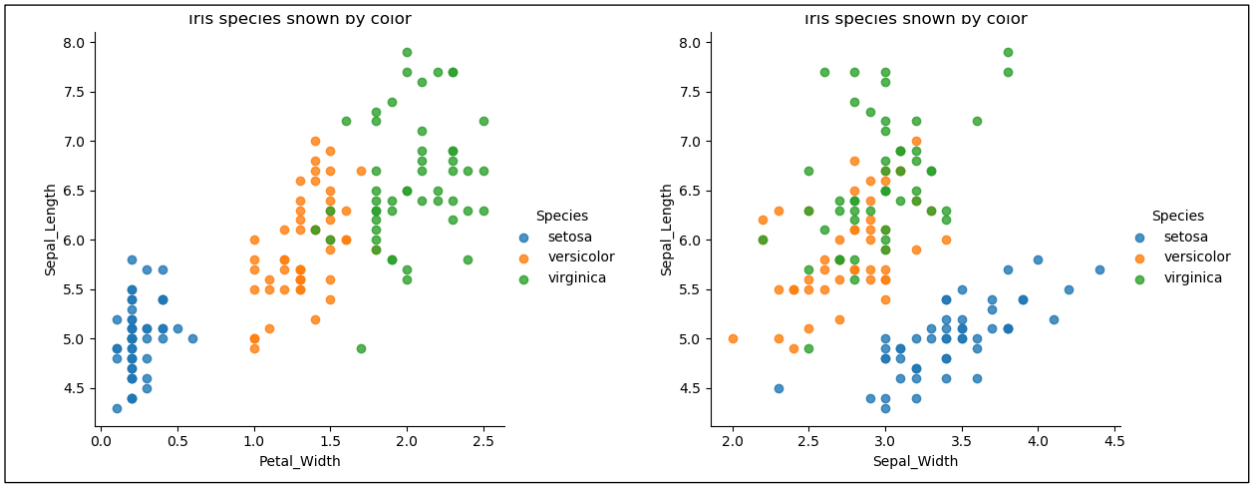
The data set of choice is 'Iris' which is the built-in data from sklearn. This data set is a list of dictionary like, 1. 'data' : containing sepal and petal dimensions

2. 'target' : containing the label assignment 0,1,2 for setosa, versicolor and virginica.

3. 'target_names': the names of the corresponding flowers.
4. ‘DESCR’ and more.

A data frame created with the four features, containing the dimensions of parts of the iris flower structures. The label column is the Species of the flower and the goal is to build a KNN model to classify the flowers correctly as much as possible.

Creating plots with the features is useful in determining which pair of features are more helpful in better classification. In an **ideal case**, the label cases are perfectly separated by one or more of the feature pairs, but such cases are very rare in real-world situation. Two of the plots for the iris data are:



As observed, **Petal_Width** and **Sepal_Length** offer a very good classification.

Preparing the data set:

Two important steps to prepare the data:

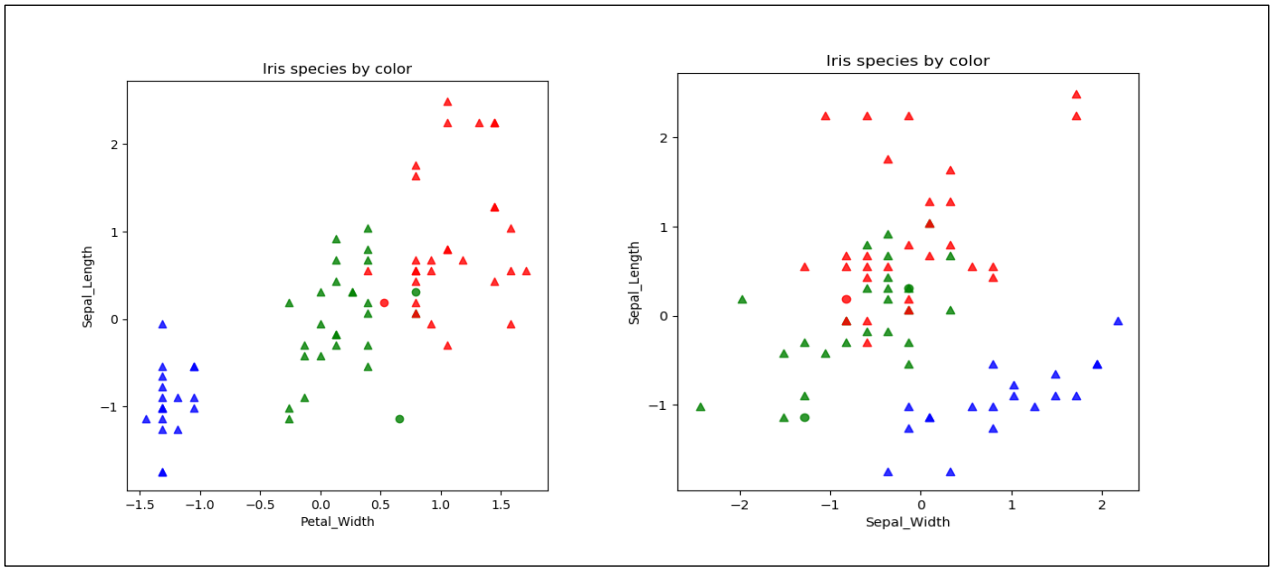
1. Scaling the numeric values of the features is important. This is done to have similar range of values so that the features with larger value may not dominate the remaining features in the model training, even though the feature with smaller values may have more important information. Thus, **Zscore normalization** is used. This normalization process scales each feature so that the mean is 0 and the variance is 1.0.
2. Split the dataset into randomly sampled training and test data set. Using **Bernoulli Sampling** to randomly select the cases prevents any leakage of information between the test and train set and allows for a better modeling.

Train and evaluate the KNN model:

Following the preparation of the data, next step is to train and evaluate a KNN model for k=3 as we have three different species of flowers.

The model was applied on the test data set. The predicted labels were compared with the real labels. The accuracy of the model = 94.67 % which is a good performance.

Creating plots for the prediction of classes for the pairs of features as earlier, we observe:



In the plots above, the **colors** represent the predicted class of the flowers. The **triangles** indicate correctly classified cases while the **circles** represent misclassified cases. The prediction accuracy is quite impressive. I mentioned earlier, that the pair of features “Sepal_Length vs Petal_Width” offer almost ideal case of classification. However, I now realize that the pair does not help us confirm or measure the precision of the machine learning model. Instead, “Sepal_Length vs Sepal_Width” turned to depict the model’s performance very well.

Conclusion:

KNN is an effective machine learning model. It can correctly classify cases even if the separation of cases is not much clear.

With the completion of this simple project, it has helped me get an understanding of the complete end-to-end machine learning process including data exploration, data preparation, modeling and model evaluation along with a basic understanding of the underlying principle and associated terminology.