Case Study: Principle Component Analysis (PCA)

Problem Statement:

You work as a data scientist in a flower research company. The company has a sample dataset of pre-labeled data on iris dataset with features like 'sepal-length', 'sepal-width', 'petal-length', 'petal-width' and 'Class'. They plan to extend this dataset and train a RandomForestClassifier on it. But they expect the dataset to grow quite large i.e. millions of rows and are worried that a million rows and 4 features is going to be too big for them to be able to train their classifier. They wish to reduce the number of features or dimensions without a sharp decrease in accuracy of the classifier.

You have been asked to:

1. Read the sample dataset given to you.
2. Use PCA to figure out the number of most important principle features.
3. Reduce the number of features using PCA
4. Train and test the RandomForestClassifier algorithm to check if reducing the number of dimensions is causing the model to perform poorly.
5. Figure out the most optimal number of components that produce good quality results i.e. they do not cause a sharp decrease in prediction accuracy.
6. Do this for all possible number of principle components and find out the smallest number of components that our dataset can be reduced to with good prediction accuracy.
