Advanced Cancer project

This project is used to demonstrate the power of a few of the machine learning techniques that can be used to significantly reduce the resources needed to reach a prediction with an accuracy that's almost identical to that achieved using significantly more resources.

Here are some of the problems that presented

- The data in the CSV file which had the information about the genes and type of the cancer was ordered/biased
  - This can cause serious problems in the classification process.
  - The normal train_test_split was going to work here because it wouldn't give a -good enough- shuffle the data would still be ordered.
  - So numpy.random.randint was used to ensure the data to be random.
- Pipelines were used to apply a series of transforms and fittings to then classify.
- Dimensionality Reduction techniques were essential to be used.
  - The data given was 150 samples and 54676 features which was going to require much more resources that the one implemented and maybe even hurt the classifier.
  - The Dimensionality Reduction techniques that were used are PCA and LDA
- Since the data had only a few samples 150 hence, LeaveOneOut was used because it's perfect for small data samples.
- Classification:
  - SVM was used for Classification.
  - PCA was compared to LDA to find which gives the best classification result across 6 Epochs –using train_test_split with shuffle- then taking the average of validation and test accuracies to see which performed better.

This project is used to demonstrate the power of a few of the machine learning techniques that can be used to significantly reduce the resources needed to reach a prediction with an accuracy that's almost identical to that achieved using significantly more resources.