**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](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html) was going to work here because it wouldn’t give a -good enough- shuffle the data would still be ordered.
  + So [numpy.random.randint](https://numpy.org/doc/stable/reference/random/generated/numpy.random.randint.html) was used to ensure the data to be random.
* [Pipelines](https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html) 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](https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html) and [LDA](https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html)
* Since the data had only a few samples 150 hence, [LeaveOneOut](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.LeaveOneOut.html) was used because it’s perfect for small data samples.
* Classification:
  + [SVM](https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html) 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.