COMP3212 Major Assignment 2020

# Introduction

The aim of this project is to explore the causes of different types of cancers, in order to show how based on these causes we can classify types of cancer.

I will be exploring and performing classification on two different datasets. The first one gathers the stem loop expression of Prostate cancer. The second dataset gathers the stem loop expression of Stomach cancer. An analysis of the importance of the different features is performed on the Model section of this report.

Three different machine learning classifiers are implemented, Decision trees, Random forest, and Support vector machines. They will help us analyze the importance of the different features, as well as allow us to classify the data. Python will be used alongside the sklearn library to implement the classification algorithms, and Pandas and NumPy are used for preprocessing the data.

# Data

As introduced above, two datasets are used in this project. Both are coming from the National cancer institute, collected in the Genomic Data Commons Data Portal. This institution provides quality data to cancer researchers and bioinformaticians.

The Ids of the datasets selected are the following:

* TCGA-PRAD/Xena\_Matrices/TCGA-PRAD.mirna.tsv
* TCGA-STAD/Xena\_Matrices/TCGA-STAD.mirna.tsv

The first one is a dataset contains 551 samples and 1882 identifiers; it is the Prostate cancer dataset. The second one is the stomach cancer dataset containing 477 samples and the same number of identifiers 1882.

The fact that both datasets have the same identifiers was what influenced me to most towards choosing them, as later on when classification algorithms are performed is important that the data shares the same or similar features.

Once the datasets are selected it is time to move on the preprocessing of the data. Pandas and NumPy are used at this step. Once each dataset is in a pandas dataframe I transpose it in order to have the samples as rows and the identifiers or features as columns.

Once the above key step is performed, I proceed to add a new column into each dataset, a field called “Cancer Type”. This has the values “Prostate” for the prostate dataset and “Stomach” for the Stomach dataset and is a label that will be used for the classification process. Now that we have each dataframe preprocessed as needed we can merge them together.

Finally, we just need to split the data before implementing the classifiers. The data is randomly split 10 times in order to produced results that are non-dependent in the way the data was split. The split is 80% used for training and 20% used for testing.

# Model

Three machine learning classifiers are implemented, Decision trees, Random forest, and Support vector machines. The implementation is done using the sklearn library.

The first classifier used is Decision trees, the data is fitted into the model and the following tree was produced.



*Figure 1: Decision tree*

As we can see in the tree the most relevant features used for the classification are the ones higher up in the tree, being the most important feature has-mir-215. This is due those features being used to make the key decisions. The classification accuracy obtained using this method was of a 99.03%.

The second model that was used is Support vector machines, this classification algorithm uses linear hyperplanes in order to split the data. Is important the hyperparameters we use for the SVM in order to produce the better results. A linear kernel is used, together with a gamma value of 1 and a C value of 100. These two parameters mean:

* Gamma - the higher this value is the more exact the hyperplane is fitted on the training data.
* C - the penalty parameter, that controls the trade-off between a correct classification of the points and a smooth hyperplane.

We can look at the top features used by the classifier by looking at the coefficients used to draw the hyperplane.



*Figure 2: Feature importance SVM*

As we can see in the table above, the importance of each feature individually is low, the importance adds up to 1 so as an example has-mir-215 has an importance of above 3.5%, and it is the most important feature.

The classification accuracy obtained using this method was of a 100%, manly thanks to a correct tuning of the hyperparameters of the classifier.

The third method used is Random forest classifier, after the data is fitted to the model, we can extract the feature importance as it is a built-in function of this classifier.



*Figure 3: Feature importance Random forest*

As we can see in figure 3, we have a similar case to the representation of the SVM feature importance shown in figure 2. The individual feature importance is low, having in this case has-mir-194-1 as the most important features with an importance of only above 2%.

The total accuracy for this classification method was of a 99.5%.

# Conclusion

In the top 10 features extracted from the SVM classifier with the top 10 features extracted from the Random forest. My guess would have been that some of the features match but surprisingly they don’t. The top 10 most important features of SVM and Random forest are completely different. This may be due how each classifier uses the data differently.

When comparing the Decision tree top features with SVM top features we find that there are two of them that match. Hsa-mir-125 and hsa-mir-194-2. Showing how both models identify these features as important.

In terms of the classification accuracy score of the classifier, the one that was able to classify the data better was the SVM classifier with a 100% accuracy, followed by random forest with a 99.5% accuracy. Finally, in terms of their classification accuracy score we have Decision tree classifier with an accuracy of 99.03%.