CBioPortal Data Analysis

(60-499 Course Project)

By,
Johan Fernandes (104904941)
Julia Zheng (103795229)

1. Introduction:

CBioPortal houses various data sets extracted from studies conducted by various research centers. For this study, we have used the TCGA Provisional data set on Prostate Cancer. We have selected five datasets which consist of 5 clinical factors (one of which will be used to create classes) and variations of genetic data.

We wished to build a predictive model on one of the clinical factors and hence for this study we selected the *Disease Free Status* feature. The other four clinical factors which were active contributors in building this predictive model are:

1. Primary Gleason Score
2. Secondary Gleason Score
3. Age
4. *Overall Status*

To build our predictive model we had to gather two important factors

1. A large enough dataset for *cross validation*
2. At least 2 classifiers that could handle such a large dataset.

There are **499** cases (patient information) available in this dataset. However, the disease-free status feature was highly skewed (100 cases of reoccurrence and 399 cases of disease free). We extracted about 50 cases of reoccurrence and 50 cases of disease free patients to build a dataset of 100 cases with the 4 clinical features mentioned above.

Our first mission was to extract at least a few *genetic factors* per patient and combine it with the 4 clinical features and build a dataset of 100 cases. For this study we were able to extract and 3 such genetic factors.

1. Copy number of Alterations
2. DNA methylation
3. Gene Expression

For gene expression and copy number of alterations we have *linear* (actual) data as well as *normalized* data. Although normalized data is extracted from linear data we thought it would be best to process such data as well so that we can select the data set which provides the best estimator for out predictive model. Although we had 3 genetic factors we were able to build 5 datasets as gene expression and CNA had linear as well as normalized data.

Our next step is to begin cleanup and extraction of genetic information for each patient from the individual genetic factors that we have listed below.

2. Preprocessing Step:

The 5 datasets that were created using a combination of 5 clinical features of 100 patients and an n number of genetic information per gene depending on the genetic factor. The 3 genetic factors based on which these sets were created are as follows:

Note: *Normalized* genetic data can range from -2 to 2 based on the GISTIC scale and *Linear* data is the actual data pertaining to that gene.

Dataset	Genetic Information
1	Copy Number of Alterations (Normalized)
2	Copy Number of Alterations (Linear)
3	DNA Methylation (Linear)
4	Gene Expression (Normalized)
5	Gene Expression (Linear)

Table 1

Each of these 3 genetic factors had on an average about 20,000 and more genes. Some genes do have 0 values for each patient. They have been removed before the comparison with NCBI approved genes is done.

The *NCBI* is a governing body that assigns id's to each gene in the Human genome and decides which genes are essential enough to be studied and which aren't. On the official NCBI website one can find the gene and it's status, To get a comprehensive list of all genes that have an approved status from NCBI we had to reference the *HUGO* gene database.

On comparing the genes present in the 3 genetic factors (**Table 1**) and the HUGO approved gene list we were able to drop certain genes from each factor. Here is the result of that comparison:

Dataset	Total No. of Genes	No. of Genes after comparison
Copy Number of Alteration	24776	22235
DNA methylation	16481	14371
Gene Expression	20440	17511

Table 2

Now that we have eliminated the genes which are not approved by NCBI we can focus on building the 5 datasets we've been discussing for a while.

The five datasets are built using a combination of each genetic factor(normalized and linear are considered separate hence the number of datasets becomes 5) and clinical features. Another important aspect to note there is that the genetic information is mentioned on a per sample base. There could exist multiple samples per patient. But fortunately for us we found that each patient had only 1 sample in all of the genetic factors mentioned in **Table 2**.

Each patient's has a sample id which is used a reference to save the genetic features of that patient's sample. To simplify the extraction and combination of these genetic factors with clinical factors we have converted the sample id to patient id. Thus, merging clinical and genetic features no longer become an issue.

Dataset	Class	Clinical	Genetic Information	Feature Count
1	DFS Status	4 factors	Copy Number of Alterations (Normalized)	22239
2	DFS Status	4 factors	Copy Number of Alterations (Linear)	22239
3	DFS Status	4 factors	DNA Methylation (Linear)	14375
4	DFS Status	4 factors	Gene Expression (Normalized)	17515
5	DFS Status	4 factors	Gene Expression (Linear)	17515

Table 3

Note: DFS Status stands for Disease Free Status class.

Equipped with these 5 datasets as shown in **Table 3** we now begin our analysis phase.

3. Analysis

For the analyzing which dataset would be the most useful in predicting the disease-free status of patients we ran the Naïve Bayes, Support Vector machines and Random Forest Classifier on each dataset.

In order to get an accurate idea of what level or rate at which each classifier was able to predict the disease-free status in patients we have used three performance measurements:

1. Accuracy
2. Specificity
3. Sensitivity

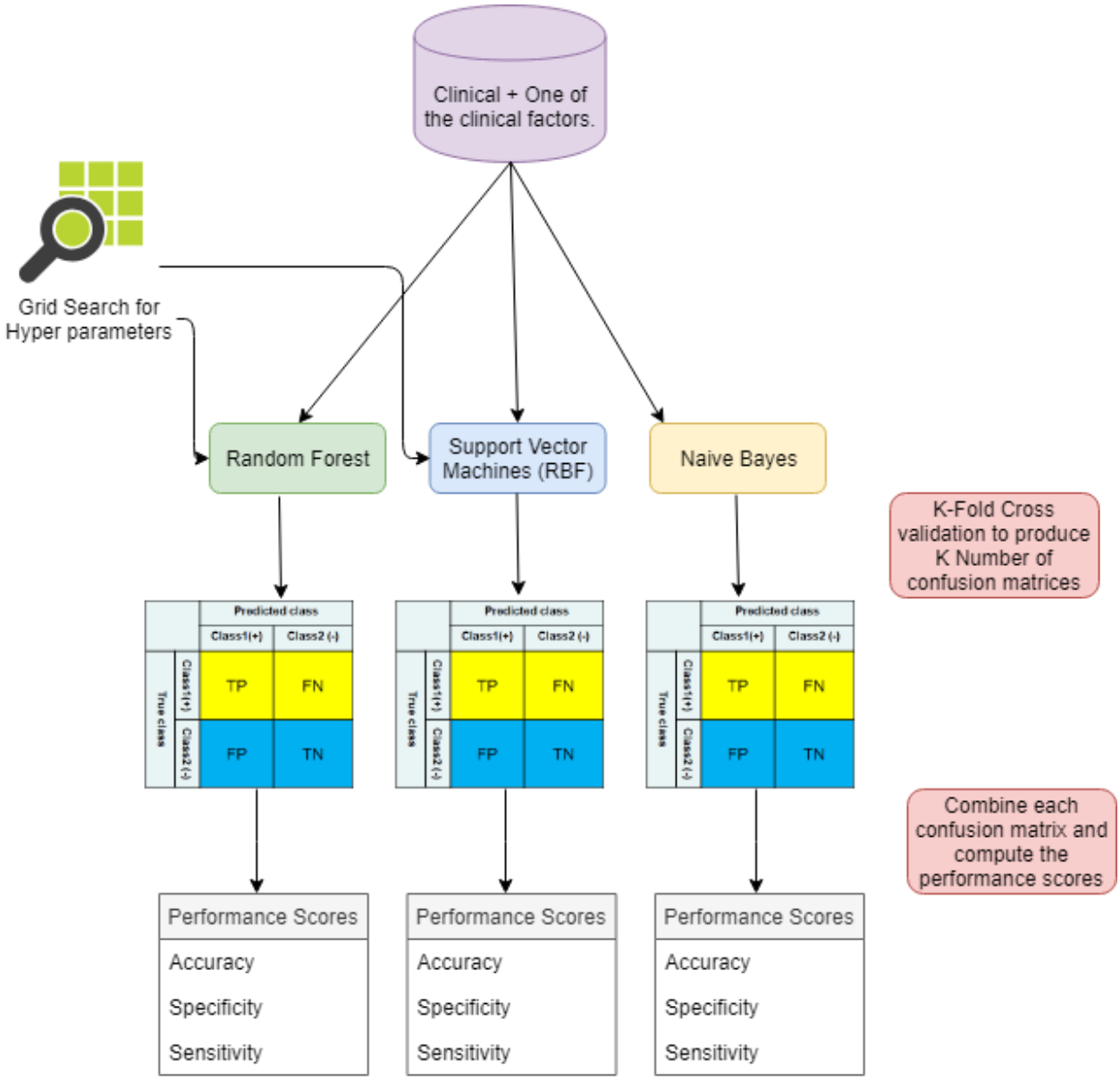
Accuracy: The rate at which positive values were correctly predicted as positive and negative values were correctly predicted as negative.

Specificity: The rate at which negative values are correctly identified as negative values.

Sensitivity: The rate at which positive values are correctly identified as positive values.

In order to get the most accurate values of performance measurements we performed a five-fold cross validation on each dataset. Each iteration provided a confusion matrix which contains values such as *True Positive, True Negative, False Positive and False Negative*. By taking an average of all these values we were able to generate the above-mentioned performance scores.

Here is a brief overview of the whole process:



The process of model evaluation and confusion matrix image are the intellectual property of Dr. Luis Rueda at the University of Windsor's School of Computer Science. All components in the image, except for the confusion matrix are developed using draw.io

Image 1

As shown in **Image 1** we calculated the performance scores for each classifier on each dataset using the confusion matrices obtained from the five-fold cross validation. An additional point here is the fact that we tuned the hyper parameters for SVM (Cost function and Gamma) and Random Forest (Number of estimators) in order to improve the accuracy of both classifiers.

Mentioned below are the performance measurements of each classifier on each dataset. The image below describes the performance scores of accuracy, specificity and sensitivity for each of the classifiers.

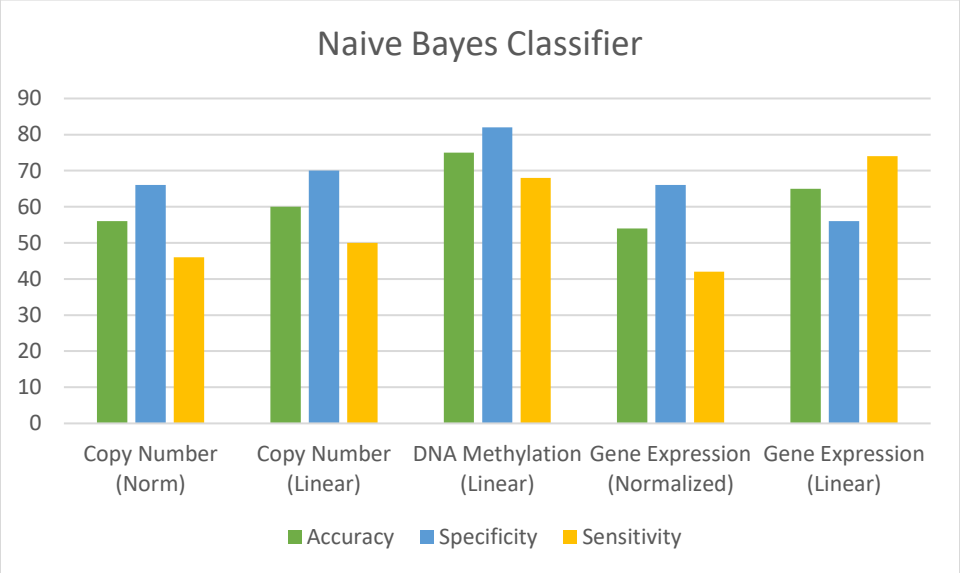


Image 2

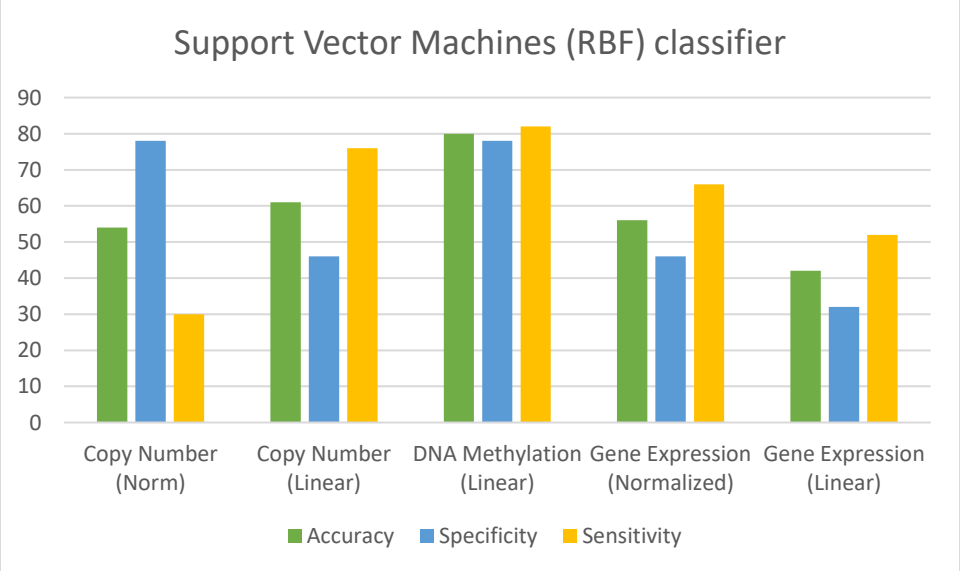


Image 3

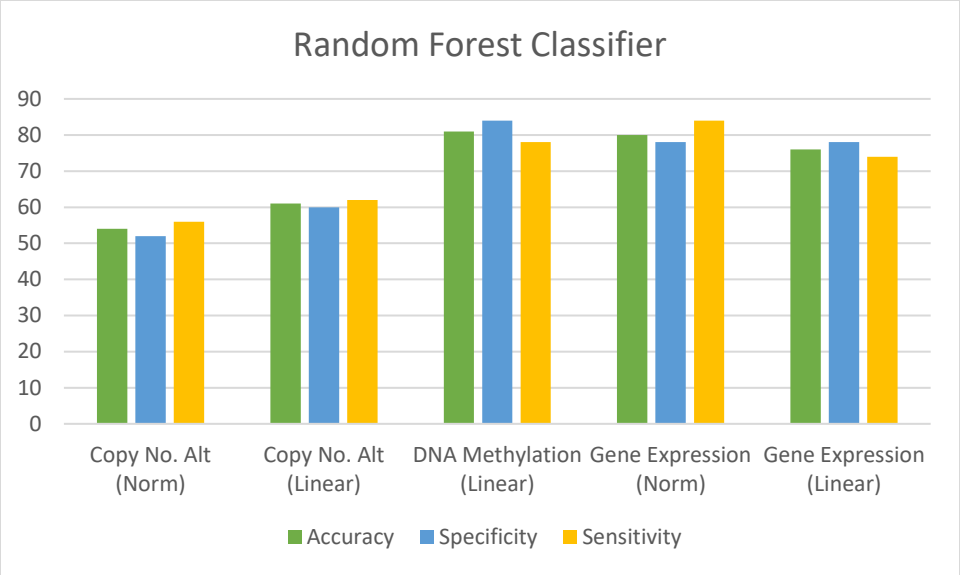


Image 4

Comparing **Image 2 – 4** it is quite evident that the Random Forest classifier proved to be the most efficient in predicting the disease-free status of cancer patients using not only the DNA methylation but also the Gene Expression (Linear) and Gene Expression (Normalized).

Result of classification based on order of accuracy from highest to lowest.

Classifier	Genetic Information	Accuracy	Specificity	Sensitivity
RF	DNA Methylation (Linear)	81	84	78
SVM	DNA Methylation (Linear)	80	78	82
RF	Gene Expression (Norm)	80	78	84
RF	Gene Expression (Linear)	76	78	74
Naïve	DNA Methylation (Linear)	75	82	68
Naïve	Gene Expression (Linear)	65	56	74
SVM	Copy Number (Linear)	61	46	76
RF	Copy No. Alt (Linear)	61	60	62
Naïve	Copy Number (Linear)	60	70	50
Naïve	Copy Number (Norm)	56	66	46
SVM	Gene Expression (Norm)	56	46	66
Naïve	Gene Expression (Norm)	54	66	42
SVM	Copy Number (Norm)	54	78	30
RF	Copy No. Alt (Norm)	54	52	56
SVM	Gene Expression (Linear)	42	32	52

Table 4

Conclusion:

After building these datasets and constantly discussing with Dr. Rueda and his team members we realized the possibilities of using machine learning in cancer research. For this study, we had accurately proved with not one but multiple classifiers and genetic factors that we could predict a clinical attribute of a patient if we have the genetic and certain clinical attributes of the patient.

We could pick any of the top 4 classifiers and datasets mentioned in **Table 4** and build a predict model to predict the disease-free status of a cancer patient, provided we know certain genetic information about that patient. The possibilities of finding different results and of building different kinds of models is infinite at this point of time.

We hope our small project attracts more people to this field of cancer research where we solve problems using the latest machine learning methodologies.

Credit: We must credit the amazing support and knowledge provided by our colleagues:
Dr. Luis Rueda and his team members:

1. Dr. Abed Alkhateeb
2. Osama Hamzeh
3. Nazia Fatima.

Reference:

1. Dr. Luis Rueda's lecture notes from his 60-473: Advance AI Machine Learning Course.
2. Pattern Classification by David G. Stork, Peter E. Hart, and Richard O. Duda.
3. Sklearn documentation available on official sklearn website
4. Python Data Science Handbook by Jake VanderPlas
5. Machine Learning Mastery website by Dr. Jason Brownie
6. CbioPortal : http://www.cbioportal.org/
7. NCBI gene cards: https://www.ncbi.nlm.nih.gov/gene
8. HUGO Gene database: https://www.genenames.org/

Index:

Disease Free Status: Cancer patients who have been diagnosed with prostate cancer and have received treatment may be at risk, due to certain genetic factors, of having a reoccurrence of prostate cancer.

Cross Validation: Process of dividing the data into test and training set iteratively such that after ever iteration each data point would have the chance to be in the training set and in the test set at least once.

Genetic factors: DNA Methylation, Gene Expression and CNA are all genetic information about various genes and are used for different purposes.

NCBI: The National Centre for Biotechnology Information houses datasets, journals and papers on different types of cancers that are currently being researched around the world

HUGO: Genetic database that houses the data related to each gene such as it's id, approval status etc.

True Positive: Positive classified as positive

True Negative: Negative classified as negative

False Positive: Negative classified as positive

False Negative: Positive classified as negative