Machine Learning Engineer Nanodegree

Capstone Project - Proteogenomics of breast cancer

Predicting breast cancer subtypes using clinical and protein expression data

Andra Tolbus
November 2019

I. Definition

Project overview

Breast Cancer is a the most common type of cancer among women worldwide and the second most common type of cancer overall, causing death in 672 000 women in 2018 alone [1][2]. Thus, early screening procedures as well as better understanding of the breast cancer mechanisms at a molecular level is crucial in order to increase the overall survival in these patients.

In the present project we aim to understand if clinical data on patients with breast cancer as well protein expression levels can be used to predict the cancer subtype.

There are 4 different subtypes of breast cancer based on the specific genes the cancer expresses : Luminal A, Luminal B, Basal-Like, and HER-2 enriched[3]. These subtypes are identified through immunohistochemistry and gene expression profiling and used together with clinical information in determining the right course of treatment for each individual patient. Relatively recently, it has been shown that proteogenomics connects somatic mutations to signalling of breast cancer [4].

Incorporating proteogenomics and additional molecular information together with the clinical data might yield in a better cancer subtypes classification , which will eventually lead to better treatments for those patients. Moreover, as personalised medicine progresses, discoveries in this area could potentially lead to a completely different classification of breast cancer subtypes.

In the present project I will utilise clinical and expression data originally generated for the above linked Nature publication[4] to determine the link between proteogenomics and somatic mutations in breast cancer. This data has been made available via Kaggle [5]. These datasets contain data on 105 patients, out of which only 77 genomically annotated

breast cancer provided high quality data. . A total of 12553 proteins per tumour were identified using mass spectrometry, which were used in determining the patient's molecular signatures. Moreover, a data set on the PAM50 genes and associated proteins which are known to play a role in breast cancer profiling [6] is provided. We will use this information to only include the relevant breast cancer related proteins in our analysis.

Problem statement

Can we predict the breast cancer subtypes including the patients clinical information (age, tumour size, stage of cancer, etc) and protein expression data?

In order to predict the cancer subtypes, I will approach the problem in two steps:

1) **Identify the patient's molecular signature** using the protein expression data and other available genetic information. This will be done using unsupervised machine learning (clustering) . Several algorithms (K-means, Gaussian Mixture Model, Affinity propagation and Agglomerative Clustering) will be used to identify these clusters and the best performing one will be selected. Prior to running the various clustering algorithms I will use PCA to reduce the number of features.

2) **Predicting cancer subtypes** by including relevant clinical data as well as the molecular signatures identified at step 1 and building a classification model to predict the cancers subtype . Several algorithms (Decision Tree, Ada Boost, Logistic Regression with a multi class setting, Random Forrest, Naive Bayes , XGBoost and SVM) will be used for the classification task and the best two will be selected for further tuning. The best performing model will be selected at the end the overall performance assessed.

These two steps will follow a thorough step of data preparation as suitable in both cases which will be detailed .

Metrics

1) **Clustering** - in order to evaluate the performance of the clustering models, I will look at the mean Silhouette score. This coefficient provides a good representation of how well each object has been classified and can take values between -1 and 1, with 1 being the best value and -1 being the worst value indicating incorrect assignment of an object to the clusters . Values very close to 0 indicate overlapping clusters [7].

Mathematically, the Silhouette score can be defined as [8]:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \textbf{if } a(i) < b(i) \\ 0, & \textbf{if } a(i) = b(i) \text{ and} \\ b(i)/a(i) - 1, & \textbf{if } a(i) > b(i) \end{cases}$$

Where $a(i)$ is the average distance of point (i) with regards to all the other points in the cluster and $b(i)$ the average distance of the point with regards to all the other points to its closest neighbouring cluster.

2) **Classification** - in order to assess the performance of the classification models, I will look at several metrics : accuracy on train and test, F1-score, F-beta score and an additional balanced accuracy score on the test set to account for imbalanced features on the test set.

Mathematically, these metrics can be defined using the terms: true positives (TP), true negatives(TN), false positives (FP) and false negatives (FN) , designed to help compare the classification results on the test set. Additionally, using these values we can also compute the values for **precision** (fraction of relevant instances among the retrieved instances) and **recall** (fraction of the total amount of instances that were actually retrieved)[9] as well as the above mentioned performance metrics:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$BalancedAccuracy = \frac{TPR + TNR}{2}$, TPR is the true positive rate, and TNR is the true negative rate.

F- Scores:

$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$, the harmonic mean between precision & recall , known as balanced F score.

$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$, weighted harmonic mean between precision & recall.

The F1 and F beta scores have the optimal values at 1, with the worst value being and at 0. We have a multi class classification at hand, so both F scores have been calculated in sklearn using the 'macro' parameter which estimates the metrics for each label and then finds their unweighted mean [10].

II. Analysis

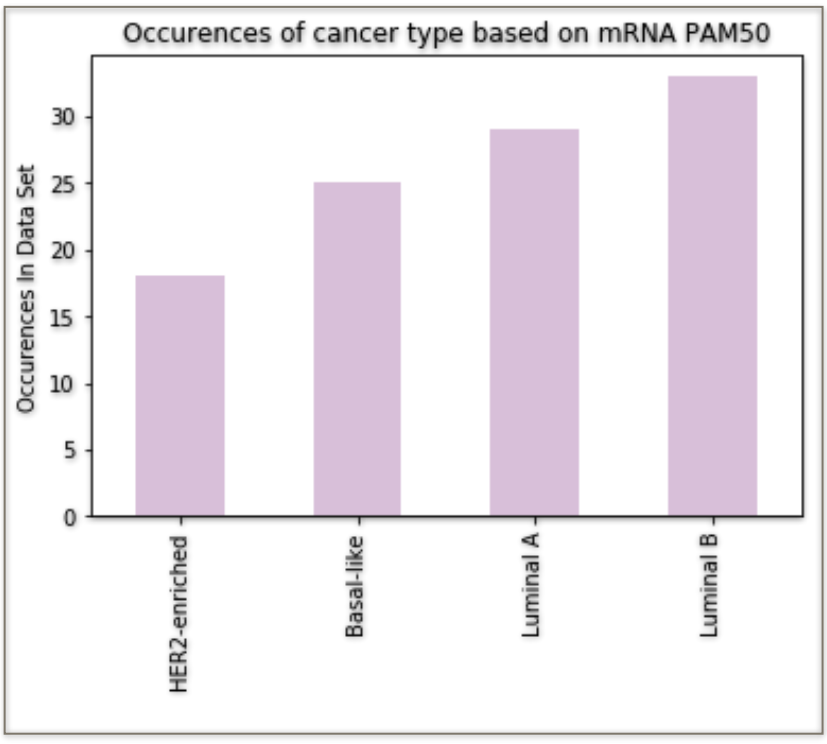
Data exploration & exploratory visualisation

Overview by data set:

Clinical Data

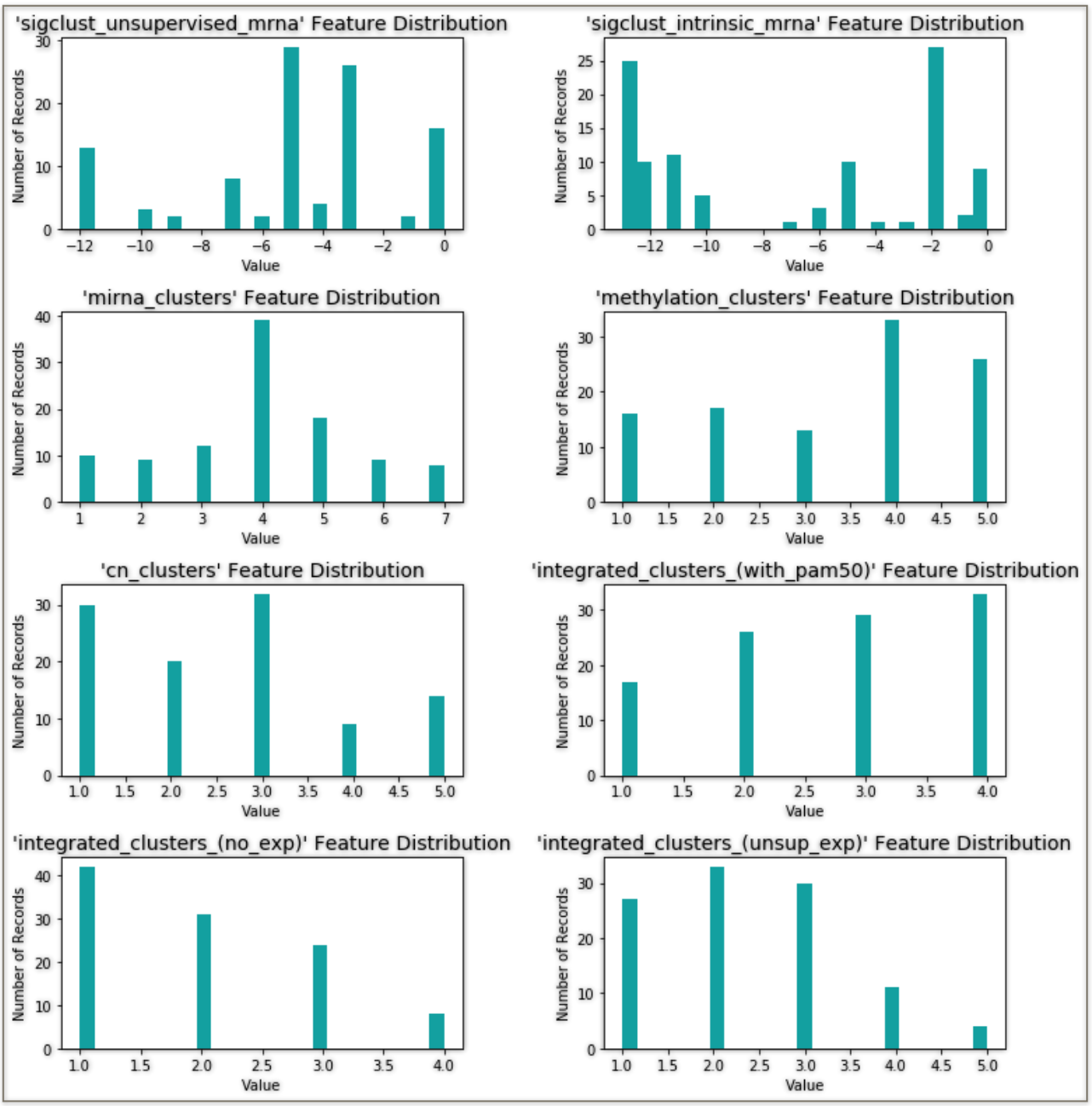
Patient characteristics and attributes related to the diagnosis (age, sex, metastasis, presence of lymph-nodes , tumour size , tumour shape , etc) . Total number of samples : 105. Reference file: *clinical_data_breast_cancer.csv .*

- **Target variable** : cancer subtype (column PAM50 mRNA) - overall balanced distribution of the 4 classes that we are trying to predict with A count of 25 Basal-Like, 18 HER-2 enriched, 29 Luminal A and 33 Luminal B .



- **Molecular data** available in the clinical data set .

The general cancer subtype classification is done following gene expression levels investigations and genome wide DNA methylation from tumour samples. Features such as the number of miRNA clusters, intrinsic mRNA clusters or methylation clusters are available for the patients included in the data set. The 8 features below will be analysed together with the expression data in an attempt to identify the clusters of molecular signatures.



- **Numerical features**

	age_at_initial_pathologic_diagnosis	days_to_date_of_last_contact	days_to_date_of_death	os_event	os_time
count	105.000000	105.000000	11.000000	105.000000	105.000000
mean	58.685714	788.390476	1254.454545	0.104762	817.647619
std	13.066630	645.283040	678.050642	0.307715	672.026613
min	30.000000	0.000000	160.000000	0.000000	0.000000
25%	49.000000	240.000000	947.500000	0.000000	240.000000
50%	58.000000	643.000000	1364.000000	0.000000	665.000000
75%	67.000000	1288.000000	1627.500000	0.000000	1305.000000
max	88.000000	2850.000000	2483.000000	1.000000	2850.000000

Overall survival (OS):

- **OS event** - whether a treatment has been registered clinic.
- **OS time** - The length of time from either the date of diagnosis or the start of treatment for a disease, such as cancer, that patients diagnosed with the disease are still alive.

Most patients (90% of the records) do not have an os event registered (90%) so we will drop this column, while keeping the overall survival time. This makes sense to include only if we add the vital status (deceased/not).

Days to date of last contact/days to date of death - clinic log information, these are not predictors for the cancer subtype . We only want to use molecular/expression data/biomarkers and personal information about the patient (age, sex, etc). These columns will be dropped.

- **Categorical features**

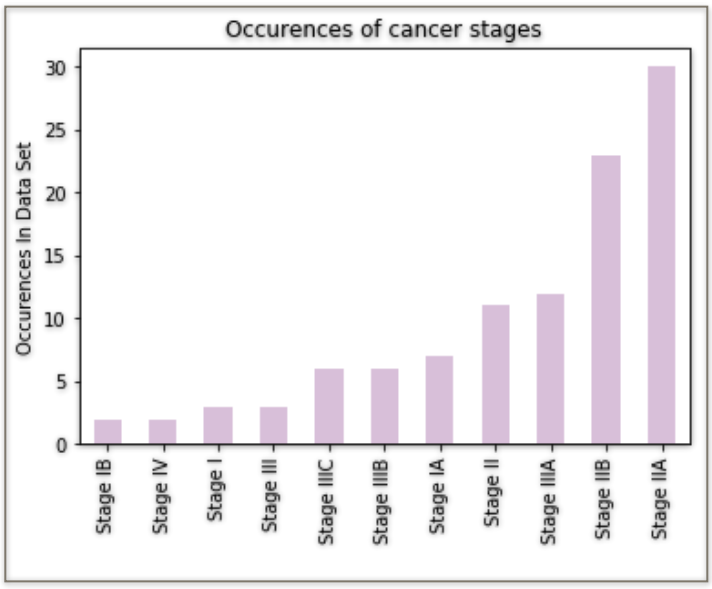
Gender - gender of the patient. Breast cancer is predominantly found in women and we only have 2 out of 105 records on male patients. This feature will be dropped.

PR status: breast cancer with progesterone receptors - 54 positive, 51 negative,
ER status: breast cancer with estrogen receptors - 68 Positive , 36 negative, 1 indeterminate
HER2 status: 77 Negative , 27 positive , 1 Equivocal; status of the human epidermal growth factor receptor 2 - known to play a role in breast cancer.
Node & Node coded - N0 (53), N1 (29), N2 (14), N3(9) - degree to which the cancer is extended to the lymph nodes. N0 - means that the cancer has not reached the lymph nodes. We will keep the coded feature here (positive/negative on whether the cancer has spread or not)

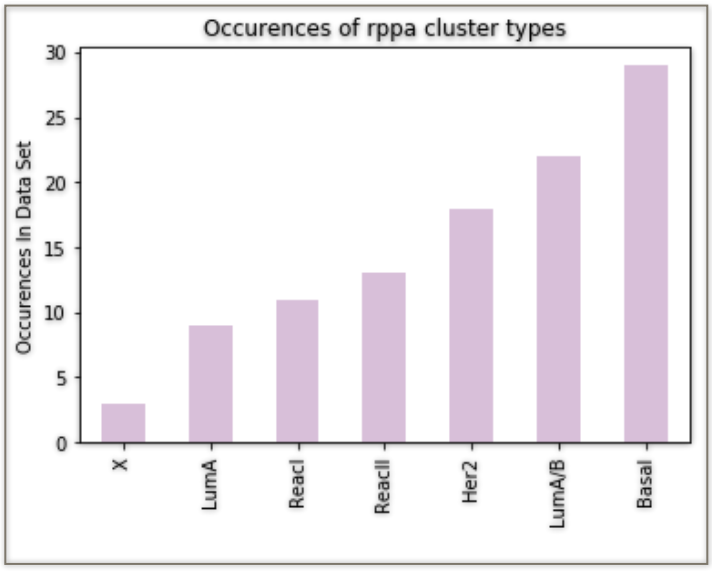
Metastasis Coded & Metastasis - we keep only one of the features as they are the same (Positive/Negative vs M1/M0). Most of the features in this data set are Negative (103 Negative vs 2 Positive). However this could be an important predictor should we train the models on larger data set.

Vital status : Deceased/Living - very important to include if we include the overall survival time as well. (11 out of 105 deceased) . This could be important in assessing the cancer subtype either in a clinical environment or during a post-mortem biopsy.
Tumour & Tumour T1_coded - they refer to the same metric which is the tumour size. In the data set T1 coded is a binary feature on wether it is T1 or otherwise (T0,T2,T3) leading to 90 samples marked as T_other. I will disregard the provided coded feature and focus on developing a new feature based on the tumour size distribution in the data set

Cancer stage (AJCC stage) - the cancer stage subgroups are very spread. These sub classifications can be condensed in larger groups to decrease the complexity.



RPPA - clusters (Reverse Phase Protein Array data) - the Nature paper [4] lists the initial use of the RPPA arrays followed by the alternative use of mass spectrometry to analyse the proteomes to counteract its limitations. Moreover the “mapping” of the clusters is very close to our predictor variable. Thus, this classification should not be considered in the analysis as it will bias the classifiers.



Protein expression data

We have expression levels across 12553 proteins (rows) for 83 patients with breast cancer, with the last 3 columns belonging to healthy individuals. This data set can be linked to the clinical data set through the TCGA id columns(included in the column names for each of the samples). Reference file: *77_cancer_proteomes_CPTAC_itraq.csv .*

This data has been normalised by the research group so there is no need to perform additional normalisation steps. Out of these proteins I will select the ones which can be retrieved in the PAM50 proteins data set (details below)

Special attention is needed during the data processing step on the rate of missing values and cases where handling missing data through imputation makes sense.

PAM 50 gene-protein mapping data

PAM50 protein data - This is official list of genes and proteins according to the PAM50 classification system. This data set will only be used to select the breast cancer relevant proteins from the expression data set using the RefSeq id . Header:

	GeneSymbol	RefSeqProteinID	Species	Gene Name
0	MIA	NP_006524	Homo sapiens	melanoma inhibitory activity

Algorithms and Techniques

General reference for algorithms description[11] and own Udacity MLND assignments [12] and [13].

Feature Engineering

Several transformations have been done to the data prior to feeding it to the classification model:

- Subcategories of categorical variable have been grouped based on the “parent” category in order to decrease the complexity.
- Missing data (valid for the expression data set) - features belonging to the expression data set that had over 20% of entries missing have been removed; for the rest of the features I have used imputation by the median value. The 20% threshold is commonly accepted when dealing with biological data. If there is a high number of missing values, the imputation step will alter the results since there is a strong relationship between these features biologically.

- Numerical features with skewed distributions have been log transformed ; other numerical features have been normalised (min-max scaling).
- The target values have been coded to numerical values
- One hot encoding has been performed on the categorical data before applying the machine learning algorithms

Dimensionality reduction and clustering techniques

The data included for identifying the molecular signatures contains multiple features many of them correlated .Thus we need to decrease the dimensionality to reduce the overlapping information contained by these features.

PCA

In order to do that, I used principal component analysis (PCA) . PCA is one of the most commonly used dimensionality reduction algorithms by defining a new coordinate system in which the first axis follows the highest variance in the data. The next axis will be orthogonal on the first one, following the direction of the next highest variance in the data. The following axes are all orthogonal on each other, following the same principle.

In sklearn, we able to define the threshold of variance that the principal components should explain - in this case I selected 80% to account for the imputed values and potential experimental noise.

K-means - is an unsupervised learning algorithm which given a specific number of clusters (k) will attempt to partition the data by assigning k centroids to a feature space and computing the Euclidean distance (or a similar metric) between each sample to each centroid . Based on this, the algorithm assigns the closest centroid to each sample. Then we calculate the average feature vector of the samples that have been labelled with it which become the new locations of the centroids. We repeat the process until the assignments do not change anymore. The model will then output a list of labels assigned to the samples. K-means is a hard clustering technique.

Gaussian Mixture model is soft clustering technique where all points belong to all clusters but they have a different membership scores. This is actually one of the advantages in using this algorithm, where probabilities for belonging to a certain cluster are assigned to points, allowing for mixed memberships.[12].

Affinity propagation - is a clustering algorithm commonly used on biological and healthcare related applications such as detecting genes from microarray data. In this case we do not define the desired number of clusters like we do when running k-means or GMM. Instead, the algorithm takes into consideration measures of similarity between pairs of data points [13].

Agglomerative clustering - is a hierarchical type of clustering where groups of samples are assigned to clusters by computing their similarity metrics and linkage functions . The linkage

function takes into account the distance information . This process is repeated until all samples are included and linked in a hierarchical tree.

Classification

The task at hand is predicting cancer subtypes, which is a labelled target class for each example in our data. The suitable approach in this case is classification and there are many algorithms which can be used in this case, with various strengths and weaknesses which will be outlined below:

Naive Bayes Classifier - is one of the simplest classifiers, based on Bayes' Theorem and has a strong assumption of independence between the features. This makes it suitable for small datasets such as the one we are analysing as it is not taking into account the relationships between the features. This might however be a disadvantage here as all the features characterise the breast cancer patient are highly overlapping.

Logistic regression - is a classification learning algorithm generally used for binary classification, but can be adapted to multi class classification as well.It has a low number of parameters, it's quick and outputs probability estimates for the outcomes.

Decision Tree - is a rule based classifier. Basically the data is split under certain conditions in order to classify the outcome. Other definitions describe it as a an acyclic graph used for decision making. The advantage here is that it is highly explainable and it might work well on smaller datasets with not too many features. However, considering the features vs samples ratios in our case, this method will likely overfit .

Support Vector Machine (SVM) - is a classifier which divides the data using hyperplanes or higher order polynomials functions making it very powerful in high dimensional cases. These algorithms perform very well in certain cases such as text classification - but they are not very explainable and it can be very difficult to select and tune the right hyper parameters.

Ensemble Learning (Adaboost, Random Forrest, XGboost) - is an approach where instead of training one classifier at a time, we train a large number of weak learners with low accuracy levels and then combine them in a "meta-model" with high accuracy through either bagging (Random Forrest) or boosting (XGboost and Adaboost) . These methods counteract overfitting in cases where simple methods such as Decision trees fail. With a high number of features and not many parameters to tune, we will probably have a high chance in getting higher accuracy levels with one or more of the selected algorithms in this category.

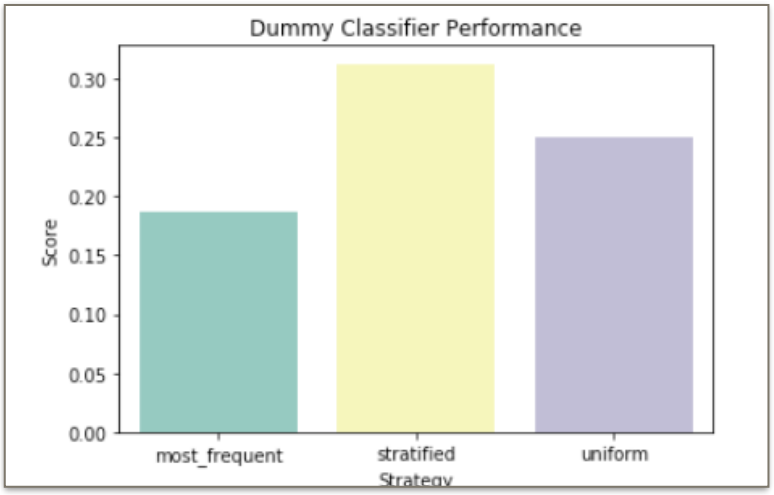
Cross Validation using Grid Search

Considering the low sample size, I cannot split our data in training, validation and testing sets as it would be ideal. Instead I will only split it in training and testing sets.

At this step, I will select the best performing classifier (aside from the simple Decision Tree) and will use cross validation through grid search for tuning a predefined set of hyper-parameters with the goal of increasing the model's F score. Once we have identified the best configuration, I will test the performance of the model on the test data set .

Benchmark

For **clustering**, I will use the results provided in the peer reviewed Nature paper [4] where they identified 3 distinct molecular signatures using k-means, while keeping in mind that the exact variables and data processing steps are not easily followed in the publication. They have also focused on the full protein expression data set, not just on the breast cancer relevant proteins. Moreover, the research group has very specific domain knowledge not available for the current analysis. The aim here is to get similar results , with acceptable values between 2 and 5 clusters.



For **classification** , the F score for a dummy benchmark model will be used. These models provide predictions using very simple rules. Sklearn provides an implementation for such a classifier with different strategies. The best scoring dummy classifier is the one with the *stratified strategy* with an **F-score of 0.3125.**

III. Methodology

Data preprocessing

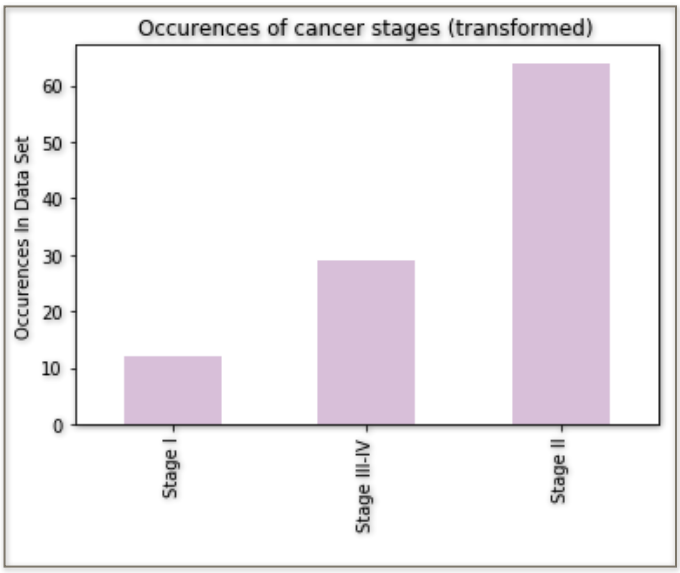
Clinical data

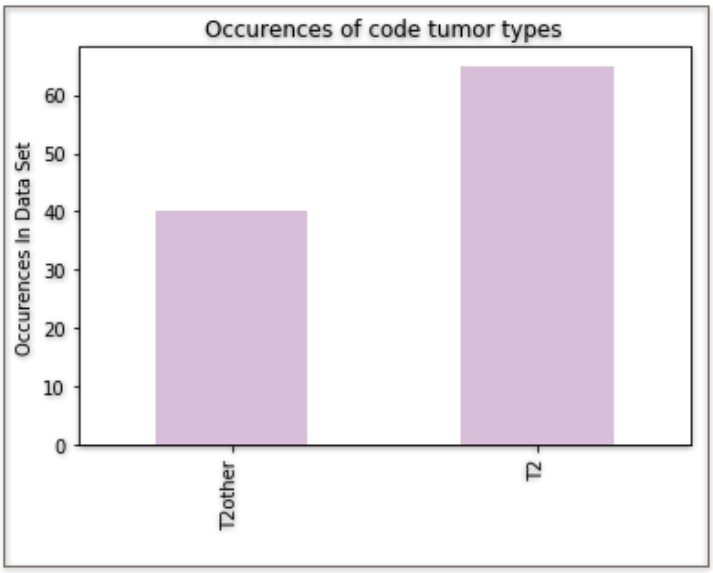
Several features that are not relevant for the analysis or are encoded versions of other features have been dropped (as described in the data exploration step).

Other features have gone through a preparation step prior to one-hot-encoding.

The cancer stage feature has been mapped by condensing the cancer stage subcategories in the main categories. An additional grouping has been done for the advanced stages of cancer (3-4) to obtain a more balanced distribution.

Similarly, the tumour codes have been encoded to T2 and T2 other (T1,T3,T4)
Post processing, the cancer stages and tumour types distributions look as follows:





The numerical features selected from the clinical data are **age** which has been normalised (min-max scaling) and **os_time** which has been log transformed as the data was skewed to right.

The molecular data from this data set has been extracted to be combined with the expression data for clustering.
The final features to be used in modelling from the clinical data set:

- **Age** (scaled), **ER** status, **PR** status, **HER2** status, **tumour** coded, **node** coded, **metastasis** coded, **cancer stage** transformed, **vital status** and **overall survival time** (log transformed)

Additionally we keep the unique patient ID (TCGA id) as index for further merging with the expression data.

Expression Data

The expression data set was delivered in a long format, with patient ids as columns and the expression levels for the 12553 proteins as rows.
This table was converted to a wide format with the patient id as index to prepare for the clustering task and merge it later on with the clinical data.
The protein expression levels have been normalised by the research group so no further transformation is necessary in this case.
Additional preprocessing:

- Removing duplicate patient id entries
- Removing proteins with a high ratio of missing values (>20%)- 2491 columns removed
- Impute by column median the rest of the missing values

From the clean data set, only the PAM50 breast cancer proteins were select using the provided PAM50 mapping file.
The resulting table was merged by patient id with the molecular features extracted from the clinical data set : 43 features in total for 77 patients. The final number of patients is the same as the one used in the reference Nature publication.
This is the final table that will be used for identifying molecular signatures.

Implementation

Identifying molecular signatures

Considering the high number of features vs the examples in the data set, as well as a degree of correlation between the expression levels of various proteins, dimensionality reduction is crucial in this step.

PCA was performed to identify the top most important components that will explain up to 80% of the variance. In this case 9 components explain 80% of the variance in the data and will be used in the clustering algorithm.

K-means and **GMM** were ran for a range of clusters between 2 and 19 and the Silhouette scores were reported .

Affinity propagation was ran for a various range of values (0.5 to 0.9, sep 0.1) for the damping ratio parameter or the extent to which the current value is maintained relative to incoming values.
The Silhouette scores were reported for the identified clusters

Agglomerative clustering was ran using the euclidean affinity and ward linkage and a graph of the hierarchy was reported . This algorithm was added to visualise the hierarchy of clusters to aid in drawing a conclusion on the results obtained using the other methods.

The labels from the clustering model results with the highest Silhouette score are the patient's molecular signature and were added to the clean clinical data set.

Predicting Cancer Subtypes

The clinical data set enhanced with the molecular signatures from the previous step has been one hot encoded and the target categorical variable converted to numerical values .

The data set has been split between train and test (80% vs 20%) . Due to the low size of the samples we are not able to add a validation set.

A simple scoring function has been implemented to take as parameters the predictions on train and test sets, as well as the train and test sets (features + target)
This function will report the performance metrics (Accuracy, Balanced Accuracy, F scores). The training scores are reported on a subset of the training set.

The classification models described in the Algorithms section have been trained and tested on the current data set split with default settings and their performance scores reported. The two best performing algorithms have been selected for future tuning and improvements .

If one of the simpler models (Naive Bayes, Decision Tree or Logistic regression) performs similarly to the more complicated ones, it will be selected for future improvements together with one of the more complex ones. The goal here is to see if we can achieve improvements on the category of models that have a higher degree of explainability.

Considering the size of the data set, the speed of the algorithms has not been taken into account in this case as all of them will run very quickly.

For model tuning, cross validation through grid search has been implemented with the goal to find the best set of parameters for the model in order to maximise the F score.

Additionally, a check on feature importance has been implemented and the performance metrics were reported for the 5 most important features.

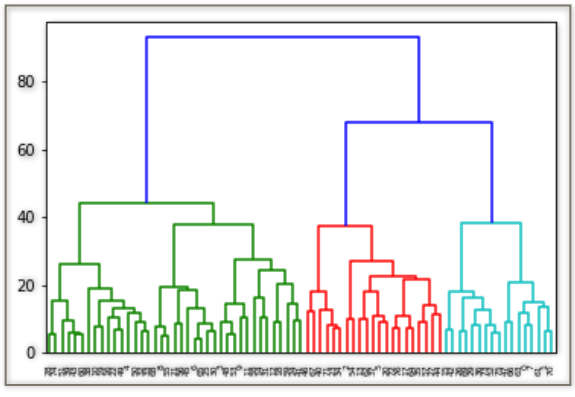
Refinement

Identifying molecular signatures

The Silhouette scores for K-means ,GMM and Affinity propagation are reported in the table below (top 7 results)

Number of clusters	K-means Silhouette Score	GMM Silhouette score	Affinity Propagation Silhouette score
2	**0.269**	**0.264**	
3	0.228	0.182	
4	0.2	0.164	
5	0.193	0.192	
6	0.192	0.153	
7	**0.194**	**0.182**	**0.179**
8	0.186	0.196	0.167

Agglomerative clustering delivered the following cluster structure :



Both K-means and GMM returned the highest Silhouette scores for n_clusters=2. On the other hand, Affinity propagation returned the highest score at n_clusters=7 (but significantly lower than for the other two models).

If we look at the hierarchy delivered by at Agglomerative clustering, we can clearing see the two main clusters with several 7-8 sub clusters in total , which is similar to what Affinity Propagation identified. We can also see an increase in score around n_clusters=7 for both k-means and GMM.

This indicates that different molecular signatures can be identified using the expression data which should be further discussed with domain experts.

However, considering the size of the data set, estimated Silhouette scores and the original benchmark (n_clusters expected between 2-5), the selected clusters indicating the molecular signatures are n_clusters=2.

The scores for n_clusters=2 are very similar for GMM and K-means (0.264 vs 0.269). I will select the labels generated by GMM in this case due to fact that it allows mixed memberships and it is thus more suitable for our problem.

The molecular signatures (cluster labels) will be used as features in the classification task.

Predicting cancer subtypes

The selected classification algorithms have been trained on the data and the performance metrics reported on the test set.
The performance overview per model using default parameters can be visualised in the table below:

	Decision Tree	Ada Boost	Logistic regression	Random Forrest	Naive Bayes	XGBoost	SVM
acc_test	0.688000	0.750000	0.562000	0.688000	0.562000	0.625000	0.438000
acc_train	1.000000	0.700000	0.800000	0.900000	0.800000	1.000000	0.800000
balanced_accuracy_test	0.684524	0.625000	0.565476	0.636905	0.517857	0.601190	0.419643
f_1_test	0.694000	0.623000	0.597000	0.667000	0.560000	0.633000	0.412000
f_1_train	1.000000	0.593000	0.875000	0.937000	0.806000	1.000000	0.875000
f_beta_test	0.736111	0.644504	0.643519	0.702694	0.601190	0.673971	0.415675
f_beta_train	1.000000	0.585000	0.875000	0.943000	0.806000	1.000000	0.875000

Naive Bayes, Logistic regression (multi class , lbfgs solver) and SVM had the lowest F and accuracy scores, with SVM performing the worst in this case. They will not be further evaluated or optimised.
I will further compare: Decision Tree, Adaboost, XGBoost and Random Forrest.

The Decision Tree, Random Forrest and XGboost are severely overfitting to the point of (nearly) perfectly describing the data set (scores of 90% and 100% on train) , while failing to predict on the test set.
The Decision Tree algorithm had the highest F score on the test set among the 3 models and given its high explainability level, it will be further selected for improvements.

Ada Boost is the only algorithm which has a more balanced bias-variance score, although it looks like it is slightly underfitting. Adaboost also had the highest accuracy on the test set. It will be further selected for improvements.

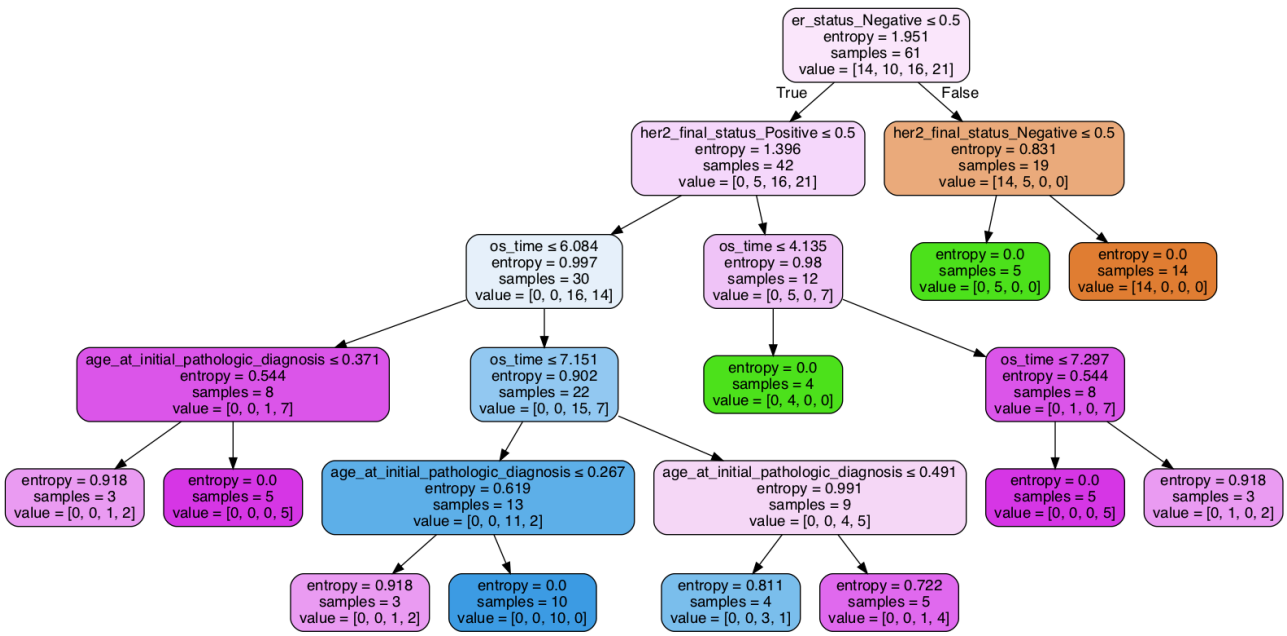
Winner models from initial model screening: Decision Tree and Adaboost.

Decision Tree Improvements

The model is highly overfitting and has an accuracy level of 69%, F1-score of 69%.
I will try to limit the depth of the tree to 5 (from 9) and add a threshold of minimum samples per leaf of 3. This setting has slightly decreased the degree of of overfitting, but without significant improvements on the scores:

```
'acc_train': 0.9,
'acc_test': 0.688,
'balanced_accuracy_test': 0.6845238095238094,
'f_beta_train': 0.943,
'f_beta_test': 0.736111,
'f_1_train': 0.937,
'f_1_test': 0.694}
```

Final Decision Tree:



In the simplified version of the tree, the main predictors are the HER2 status, os time and age , which is an oversimplification of the problem, especially since the os_time validity is dependent on the vital status.
Other parameters settings have been tried with even worse performance than the one included above.
The Decision Tree does not have a desirable performance in this case due to overfitting and oversimplification of the problem.

AdaBoost Improvements

The size of the train data set is 61 samples, whereas the size of the test set is 16.
Ideally we should try to assess the model performance on different subsets of the data so that we can get a better overview of the model's robustness .
However, due to the size of the data set the results are likely to be inconclusive.

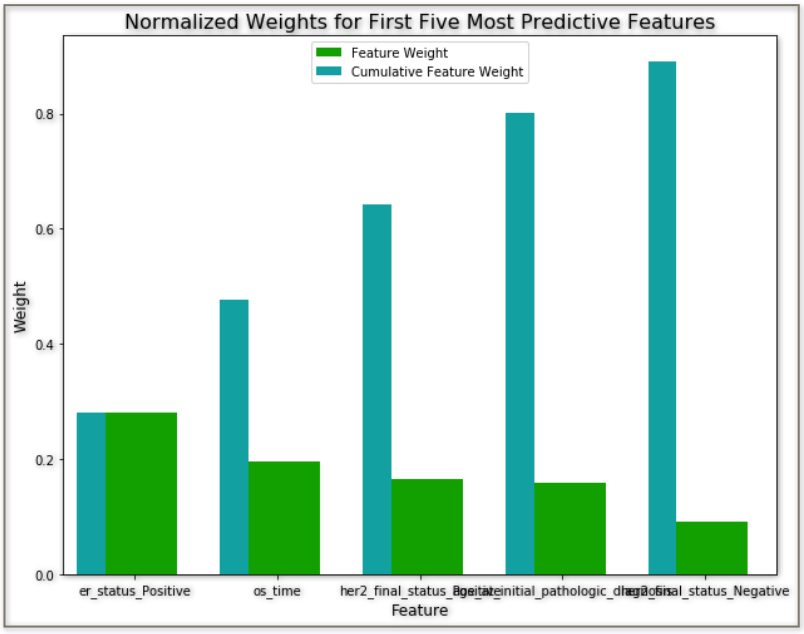
Instead, I will run cross validation through grid search on a set of predefined estimators and values for the learning rate . The learning rate shrinks the contribution of each weak learner by the defined rate value.
The best model selected through cross validation showed an improvement in F scores and Balanced Accuracy scores, with the overall accuracy staying the same:

```
Final accuracy score on the testing data: 0.7500
Final F-score on the testing data: 0.7806
Balanced Accuracy Score on test data: 0.8095
```

The best model configuration uses the Decision Tree Classifier as a weak learner with min_samples per leaf=1, learning rate =0.01 and n_estimators=100, algorithm = 'SAMME.R':

```
method BaseEstimator.get_params of AdaBoostClassifier(algorithm='SAMME.R',
   base_estimator=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
     max_features=None, max_leaf_nodes=None,
     min_impurity_decrease=0.0, min_impurity_split=None,
     min_samples_leaf=1, min_samples_split=2,
     min_weight_fraction_leaf=0.0, presort=False, random_state=None,
     splitter='best'),
   learning_rate=0.01, n_estimators=100, random_state=None)>
```

When looking at the features importance from the best model, unsurprisingly we have HER2 status(positive and negative) , ER status (negative), age and overall survival. If we train the model on a reduced dataset containing these features alone, the performance is however significantly worse:




```
Final Model trained on full data
------
Accuracy on testing data: 0.7500
F-score on testing data: 0.7806
Balanced Accuracy on testing data: 0.8095

Final Model trained on reduced data
------
Accuracy on testing data: 0.6250
F-score on testing data: 0.6588
Balanced_accuracy on testing data: 0.6905
```

IV. Results

Model evaluation and Justification

Using cross validation we identified the best scoring model for this problem, **Adaboost**, with the parameters described in the previous section. The performance of the final best model was assessed on the test set.

I would say that given our very imbalanced dataset (with high imbalances across the predictors of high importance), the final model is reasonable. However, we do not know how well it would generalise on unseen data. Ideally we would have a much higher sample size, that will likely give more balanced distribution across the important features.

With this configuration we run into the risk of misclassifying new data that had very few examples in the original training data set. Considering that the problem involves prediction of cancer subtypes and potentially a treatment course, I would not say that we can trust the model at this stage since any mis-classification would have serious consequences in the chances of overall survival for those particular patients.

However, with an **F-score** of **0.78**, the final performance results are significantly better than the dummy benchmark model, which had an F score of **0.31.**

V. Conclusion

Free Form Visualisation

Overview of the best classifier performance at different stage vs Dummy Classifier

Score on test set	Adaboost simple	Adaboost optimised	Adaboost optimised - reduced data	Dummy Classifier
F-score	0.64	**0.7806**	0.6588	0.3125
Balanced Accuracy Score	0.625	**0.8095**	0.6905	

Reflection

In the present analysis I tried to answer the question: *"Can we predict the breast cancer subtypes including the patients clinical information and protein expression data?",* using the datasets used by a research group who published a paper in the peer-reviewed Nature journal [4] on linking protegenomics to signalling in breast cancer.

I used a step wise approach: identifying molecular signatures for the patients from which the biopsies were taken through clustering, and the information was later used to predict the cancer subtype (Basal-Like, HER2, Luminal A and Luminal B). Two major signatures were identified.

For the prediction task, several classifiers have been tested, with two selected for further improvements (Decision Tree and Ada Boost). Although I managed to slightly decrease the level of overfitting in the Decision Tree, the performance did not significantly improve.

The winning classifier was Ada Boost which was further optimised through cross validation. The F score was increased by 14%through this process, whereas the balanced accuracy score was increased by 18% to a final core of 0.81. The balanced accuracy score in this case is highly relevant since overall, the data set is highly imbalanced.

Although the identified molecular signatures through clustering were not among the most important features in the classification model, they had a contribution in increasing the performance.

I think it was highly interesting to answer these questions using a data set that was used to discover important findings within breast cancer cell signalling research and try to replicate, for a part of the project, some of their findings (molecular signatures identification) . However, the mechanisms are highly complex and I do not have the necessary domain background to draw the right conclusions in my findings.
For example, I only selected the PAM50 relevant proteins to ensure relevance for breast cancer, but the whole data set had many more entries which could be manipulated in a way that makes sense biologically to look for specific signatures. Thus, I strongly believe that the a deep domain knowledge is crucial here.

This data was not originally designed for a prediction problem, but rather clustering across a high number of expressed proteins and deep investigations around it, the focus here being to have balanced sample of cancer subtypes for the analysis. The other clinical information which was used for classification in the analysis is however unbalanced. Thus, with a small unbalanced sample size, the final model is not robust enough to be used in a clinical setting.

To sum up, I would say that the model is not production ready, it can be seen as a proof of concept, with further investigations need and additional data acquired in order to increase the robustness, trustworthiness and performance scores.

Improvement

I believe that significant improvements can be achieved by acquiring clinical data and expression data on a significantly higher number of patients .

Steps that could be taken:

- Pair up with a group of researchers to get advice on handling the expression and available molecular data for identifying the signatures. This step could lead to more relevant clusters which can be used as an input to the prediction step .

- More data likely means a more balanced set across many of the features . This will allow for a thorough model testing and evaluation as well as testing the performance of unseen data.

- More data also means that relevant features such as the tumor size and node can used raw as they likely have specific molecular signatures which we could not capture with the encoding used.

- More ensemble methods should be optimised for the task as in the present analysis they had just a slightly lower performance on the test set so we might actually get even better performance with other algorithms.

References

[1] https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

[2] https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5715810/

[4]https://www.nature.com/articles/nature18003

[5] https://www.kaggle.com/piotrgrabo/breastcancerproteomes

[6] https://ww5.komen.org/BreastCancer/PAM50.html

[7] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

[8] https://en.wikipedia.org/wiki/Silhouette_(clustering)

[9] https://en.wikipedia.org/wiki/Precision_and_recall

[10] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[11] “The hundred page machine learning book” - Andryi Burkov

[12] https://github.com/andratolbus/creating_customer_segments/blob/master/customer_segments.ipynb

[13] https://github.com/andratolbus/ml_classification_finding_donors

[14] Brendan J. Frey and Delbert Dueck, “Clustering by Passing Messages Between Data Points”, Science Feb. 2007