# Stakeholder report

Emma Pedersen, Melani Lærke Pedersen og Heidi Andersen

27 September, 2021

This report investigates the occurence of breast cancer in fine needle aspiration (FNA). Using unsupervised machine learning, we will analyze the data for the characteristics of the diagnoses to investigate whether the diagnoses are separated across clusters/groups according to their features. Furthermore, using supervised machine learning, we will create a prediction model to predict whether an FNA results in a malignant or benign diagnosis based on the features of the FNA.

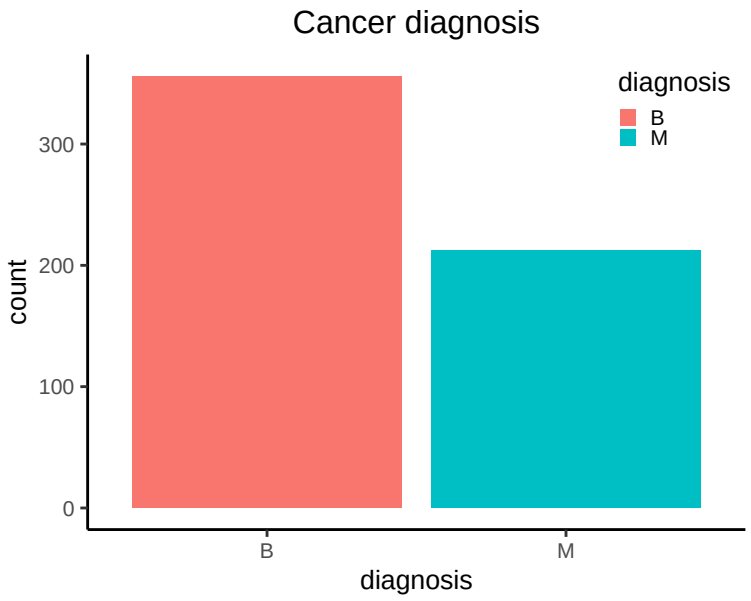
For the analysis, we have obtained the data set *Breast Cancer Wisconsin (Diagnostic)* from Kaggle, https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. This contains 568 observations of different FNA attributes. The attributes consist of the ID number of the FNA as well as the diagnosis. In addition, they consist of ten real-valued features are computed for each cell nucleus, all of which are divided between mean, standard error and worst/largest values. The attributes are summarized as follows:

- id: ID number
- diagnosis: Diagnosis of breast tissues (M = malignant, B = benign)
- radius: Distance from center to points on the perimeter
- texture: Standard deviation of gray-scale values
- perimeter: Size of the core tumor
- area
- smoothness: Local variation in radius lengths
- compactness: perimeter^2 / area - 1.0
- concavity: Severity of concave portions of the contour
- concave points: Number of concave portions of the contour
- symmetry
- fractal_dimension: “coastline approximation” - 1

All feature values are recoded with four significant digits. The attributes are

## Exploratory Data Analysis

Before the data set can be used for predicting cancer diagnoses, it requires some data preparation and a clearer overview of the cancer types. This overview of the number of possible cancer diagnoses, and how many actually get that diagnoses is summarized in the following bar plot.



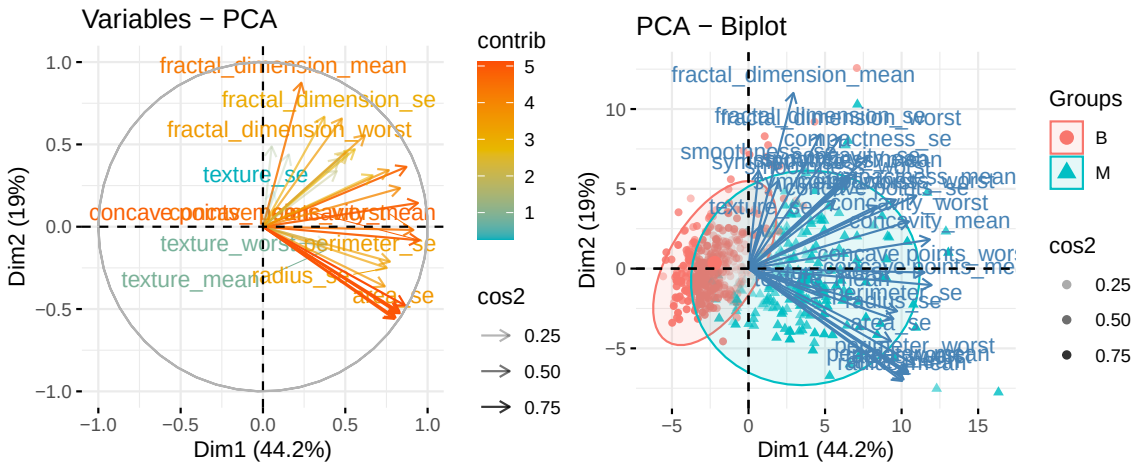

It turns out that 212 of the FNA tests are malignant, while 356 are benign. As can be seen from the bar plot, this corresponds to approcimately 1/3 of the diagnoses being malignant. However, it is not evident from the EDA what features affect the diagnosis. For further analysis of this question, we will use machine learning techniques,including dimensionality reduction and clustering, which is elaborated in the following chapter.

## Unsupervised Machine Learning

Data collected and used to solving *"real-world problems"*, like the prediction of the type of breast cancer, often has a large number of rows, and a great deal of columns. These columns represent "features" in a machine learning model, for example, the mean size of the core tumor called perimeter_mean in this dataset. Each row represents a specific FNA-test, noted by an ID number to anonymize the test.

While a larger dataset is necessary in terms of observations to perform an accurate prediction of future outcomes, this is not always the case in terms of variables. The problem that may arise here is that some of the variables are strongly correlated which implies that they are measuring the same phenomenon, and our model is therefore double counting the same thing. We have therefore run a *dimensionality reduction analysis* to construct new uncorrelated variables while maintaining a majority of the variation from the original dataset.

The result from the dimensionality reduction is an initial deduction from 30 features to 6 dimensions. However, from further analysis, it can be concluded that fewer dimensions are sufficient since some dimensions have a large degree of correlation. The two dimensions with the largest explained variation of the data are plotted in the following biplots, to investigate how the features are allocated according to the dimensions.

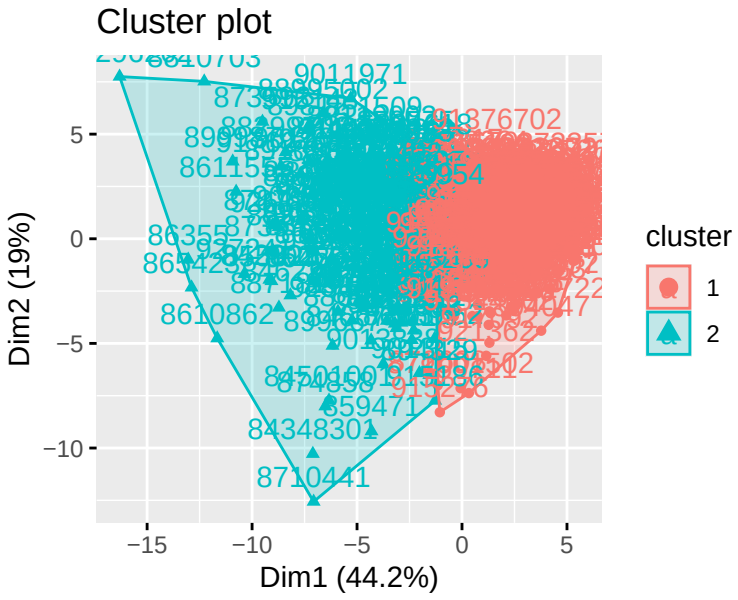



From the figure to the left it can be seen that the features move jointly according to the first dimension shown on the horizontal axis, meaning they move in the same direction. However, they split in the second dimension/vertical axis. It can furthermore be seen that radius, perimeter and area are correlated, while fractal dimension, smoothness and compactness are correlated. This could reflect that the second dimension is a measure of the complexity of the observations.

The figure to the right illustrates the biplot grouped by the diagnosis. Here, the benign cases are scattered across the red ellipsis while the malignant cases are scattered across the blue ellipsis. However, some of the observations lie outside their corresponding ellipsis. From the biplot it can be seen that the malignant cases are furthest to the right on horizontal axis, hence the features are correlated with a malignant diagnosis. In terms of the vertical allocation, there is no clear distinction between benign or malignant diagnoses.

Since the data set is now minimized to include the most essential, a clustering analysis can now be performed. This is done to classify data into structures that are more easily manipulated and understood.

The data is divide into two clusters reflecting the two possible diagnoses.




It is seen that the observations are split with a clean cut between the two clusters. Therefore, it seems that this method is working well on dividing the data between groups. However, the clustering does not show how the diagnoses are separated across the clusters; does the clustering distinguish between malignant and

benign cases, or are they spread ambiguously? The following table shows the distribution of diagnoses across clusters.

```
    1   2
B 342  14
M  37 175
```

It seems that the clustering does a good job in separating the diagnoses across clusters. Hence, 3.93258% of benign cases and 17.45283% of malignant cases seem to be clustered incorrectly given a clustering dividing the diagnoses completely. These percentages are low enough for us to assume representative clusters for the diagnoses. Therefore, it is possible to investigate the features of the model according to the clusters.

| cluster | radius_mean | texture_mean | perimeter_mean | area_mean | smoothness_mean |
|---|---|---|---|---|---|
| 1 | 12.434 | 18.244 | 79.898 | 487.20 | 0.09221 |
| 2 | 17.557 | 21.359 | 116.408 | 993.67 | 0.10492 |

It is seen that all features take on a higher value for the cluster containing the malignant diagnosis, indicating that on average, a higher value of all features respectively indicate a higher possibility of the diagnosis being malignant.

## Supervised Machine Learning

After having gained the clusters and seen that clustering separate the diagnoses neatly across the two clusters, we are now proceding to perform supervised machine learning. This is done in order to predict whether a patient who underwent a FNA analysis will be diagnosed with benign or malignant cancer, based on the features of the model.
To create a model that can predict the results of FNA tests, it is required that the data set is divided into a training and test set. The training dataset can then be used to construct the model, while the testing data set is used to test the accuracy of the models. The training set is thus used to create the recipe for the model whereafter the models are created and the data is fit onto the models.

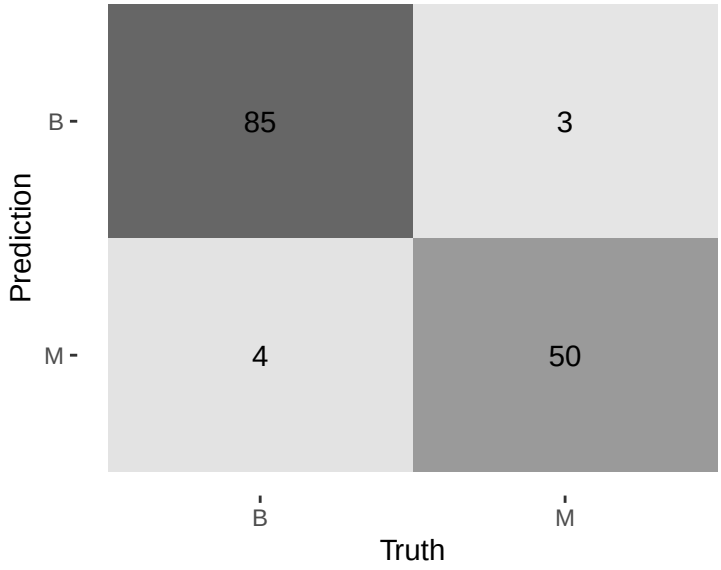
```
[10:08:22] WARNING: amalgamation/../src/learner.cc:1095: Starting in XGBoost 1.3.0, the default evaluati
```

We have chosen to set up a logistic regression model, xgboost model and a random forest model. The accuracy of these models is set out in the table below. This is done in order to choose which model does the best job in predicting the data.

```
# A tibble: 3 x 2
  model .estimate
  <chr>     <dbl>
1 glm       1
2 rf        0.991
3 xg        0.953
```

It is seen that the model with the best accuracy is the logistic model with an accuracy of 100%. However, an accuracy of 100% seems unreliable. Furthermore, even though the other models are not as accurate as the logistic one, they still have a high accuracy of 99.5% and 97.4%, respectively.

Since the logistic model has the best fit, it is used for the final predictions. To see whether the model predicts the true values correctly, a confusion matrix is drawn. Here it is shown how the predicted values fit the true values.




The model fit is shown to predict the model quite accurately. It is seen that out of 142 observations, 129 are predicted accurately. The accuracy is furthermore examined below together with sensitivity and specificity.

```
# A tibble: 3 x 3
  .metric  .estimator .estimate
  <chr>    <chr>          <dbl>
1 accuracy binary         0.951
2 sens     binary         0.955
3 spec     binary         0.943
```

From the table above, it can be seen that the model has managed to accurately predict 95.07% of the test observations from the FNA correctly. This is seen by series *accuracy*, where the number of true predictions is divided by all observations. In addition, it can be seen from the table that the model sensitivity is 95.51%, i.e. out of all predictions of benign cancer, only 4.49% are predicted incorrectly. On the other hand, the specificity computes the share of truly predicted outcomes of malignant with a specificity of 94.34%, i.e. out of all predictions of malignant cancer, only 5.66% are predicted incorrectly. Due to the high values of the metrics, it is concluded that the model is a good predictor for the results, benign or malignant cancer, of a fine needle aspirate (FNA) of a breast mass in Wisconsin.