Building a classifier with Weka for the Breast Cancer Wisconsin (Original) Data Set

Naomi Hindriks

10/25/2021

Abbreviations

AUC	Area under the ROC curve
FNA	Fine needle aspiration
PCA	Principal component analysis
TNR	True negative rate (specificity)
TPR	True positive rate (sensitivity)
UCI	University of California Irvine
Weka	Waikato Environment for Knowledge Analysis

Contents

1 Introduction

Breast cancer is a major health threat for women. In 2020 it was the most commonly diagnosed cancer with 11.7% of all newly diagnosed cancer cases being breast cancer. Furthermore it is the leading cause of cancer death in females with 685,000 deaths (15.5% of all female cancer deaths) in 2020 [1]. Early detection of breast cancer is a crucial factor in the prognosis and survival rate of the patients [2].

Fine needle aspiration (FNA) is a type of biopsy used to collect a sample of cells from a lump or mass. This sample can be viewed under a microscope and different cytological characteristics can be observed. FNA is a cost-effective, fast and complication-free technique to investigate a lump or mass [3]. But even though some of the cytological characteristics observed with FNA show a statistical significant difference between benign and malignant samples, not one single characteristic can be used to accurately separate the benign from the malignant samples [4]. If breast FNA samples are used to triage possible breast cancer patients it is of utmost importance to have a high level of certainty in determining which of the samples are malignant. It is very important not to classify a malignant sample as benign, as those patients will not go for further examination and treatment. The other way around, classifying a benign sample as malignant, is less disastrous, as the further examination of the patients will identify those samples as benign.

To distinguish the malignant from the benign samples the practice of data mining might be able to help. Data mining is a modern technique used to find patterns in large batches of data. Between January of 1989 and November 1991 Dr. William H. Wolberg from the University of Wisconsin Hospitals has collected 699 breast FNA samples. Of these samples nine cytological characteristics were scored on a scale from 1 to 10 with 1 being the closest to benign and 10 being the most to anaplastic [5]. These nine characteristics are all considered to differ significantly between benign and malignant samples [4]. In the *Breast Cancer Wisconsin (Original) Data Set* that was assembled from this data, Dr. Wolberg added the correct classification to each sample: benign or malignant [6]. This report will revolve around determining whether this data set is well suited for the purpose of data mining and cleaning it up where needed, and trying to make a suitable classifier based on this data.

2 Materials & Methods

For analyzing and cleaning the data as well as for building the machine learning classifier an assortment of materials and methods was use.

2.1 Materials

The data that was being used for writing this report is the *Breast Cancer Wisconsin (Original) Data Set* [6]. The data was collected by Dr. Wolberg between 1989 and 1991. It was later published on the *University of California Irvine (UCI) Machine Learning Repository Center for Machine Learning and Intelligent Systems*, where it was downloaded from to be used in this paper. The data set reports 9 attribute values, an ID code and a class label for the 699 instances it encompasses. Each attribute is a cytological characteristic scored on a scale from 1 to 10. The structure of the data is described in table 1.

To analyze the data the programming language R (version 4.0.4) [7] has been used. R is a programming language widely used for statistical analyses, data mining and data visualization. In table 2 the R packages that were used are listed.'

For the development of a classifier, running the classifier on the data set, and evaluating the performance of the learning algorithms the Waikato Environment for Knowledge Analysis (Weka), version 3.8.4 [8] was used.

The Java programming language (Java SE 11) [9] has been used to develop a command line interface application with which the classifier can be used by others.

Table 1: Attribute Information from the Breast Cancer Wisconsin (Original) Data Set

Column	Attribute	Unit	Description
1	Sample code number	id number	Unique number given to each sample
2	Clump Thickness	1-10	Assesses if cells are mono or multi-layered
3	Uniformity of Cell Size	1-10	Evaluate the consistency in size of the cells in the sample
4	Uniformity of Cell Shape	1-10	Evaluate the consistency in shape of the cells in the sample
5	Marginal Adhesion	1-10	Quantifies proportion of cells that stick together
6	Single Epithelial Cell Size	1-10	Measures the enlargement of epithelial cells size
7	Bare Nuclei	1-10	Proportion of nuclei surrounded by cytoplasm versus those that are not
8	Bland Chromatin	1-10	Rates the uniform texture of the nucleus in a range from fine to coarse
9	Normal Nucleoli	1-10	Determines whether the nucleoli are small and barely visible or larger, more visible, and more plentiful
10	Mitoses	1-10	Describes the level of mitotic activity
11	Class	2 or 4	Classification: 2 for benign and 4 for malignant

The cytological characteristics of breast FNAs (seen in rows 2-10) get a score from 1 to 10 by an examining physician with 1 being the closest to benign and 10 the most anaplastic.

2.2 Existing methods

Different existing methods have been used to analyze the data set, build a machine learning classifier and measure the performance of different classifiers.

To test if the difference in class distribution is significant for the attributes the Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon, Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test) [10]. This is a non-parametric test that can be used for ordinal data to test that a randomly selected instance from one set of data has an equal probability to be greater than an instance from the second set of data.

To assess the extent correlation between the different attributes the Kendall rank correlation coefficient [11] [12] (commonly referred to as Kendall's τ coefficient) has been used. This choice was made since this test is appropriate for ordinal (it is a non-parametric test), and capable of handling ties.

Another existing method that has been used is Principal component analysis (PCA) [13], PCA is a method to rotate the data so that the largest variance is captured on one axis, it can be used to reduce the dimensions of the data and show multidimensional data in a two dimensional plot.

Different existing learning algorithms have been tested to evaluate their effectiveness on the Breast Cancer Wisconsin (Original) Data Set. The algorithms that were used are included in Weka (version 3.8.4): ZeroR, OneR [14], Naive Bayes classifier [15], simple logistic [16] [17], Support Vector Machines (SMO in Weka) [18] [19] [20], K-nearest neighbours (IBk in Weka) [21], J48 decision tree [22]. Some meta classifiers were also used: Random Forest [23], AdaBoostM1 [24], bagging [25], voting [26] [27] and cost-sensitive classification and cost-sensitive learning.

Different measures have been used to asses the performance of the classifiers: accuracy, true positive rate (TPR), true negative rate (TNR), area under the ROC curve (AUC) and the F_2 score. Paired two sample t-tests [28] were used to test the difference between the performance measure between the different classifiers.

2.3 Developed methods

A combination of different learning algorithms (cost-sensitive, voting, Naive Bayes, IBk and Random Forest) were combined into a classifier that was trained with the Breast Cancer Wisconsin (Original) Data Set to classify new instances. The process in which this classifier was build can be found in Thema09: Building a classifier with Weka repository [29]. After the making of this classifier a command line application was build with Java to use the classifier with an easy command line interface. The application can be found in the Breast Cancer Classifier repository [30].

Table 2: R packages used for analyzing, manipulating and visualizing data

Package name	Version	Usage description
tidyr	1.1.4	Reshaping the data (e.g. from wide to long format)
kableExtra	1.3.4	Formatting tables to present data
dplyr	1.0.7	Manipulating the data (e.g. grouping the data and calculating new values)
ggplot2	3.3.5	Making a visualization of the data in plots
ggpubr	0.4.0	Custumazition of ggplots
scales	1.1.1	Used for giving color to ggplots
ggalt	0.4.0	Additional options for ggplot (e.g. encircling data in plot)
ggcorrplot	0.1.3	Visualization of correlation matricx for ggplot
forcats	0.5.1	Used for working with categorical data (ordering of factors)
patchwork	1.1.1	Making plot compositions of ggplot plots
foreign	0.8.82	Reading and writing ARFF files (data from Weka)
ggthemes	4.2.4	Used for getting colors for ggplot
latex2exp	0.5.0	Parsing of LaTeX math formulas to R's plotmath expressions, to be used as titles/labels in ggplots.
ggbiplot	0.55	Used for making principal component ggplot

3 Results

3.1 Duplicated data

While inspecting the instances of the data set it became apparent that there were a lot of duplicated sample code numbers, even thought these are supposed to be unique. There are 100 instances that share their sample code number with at least 1 other instance and there are 46 sample code numbers that are found at least twice in the data set. In table 3 and table 4 all the instances with duplicated sample code numbers are displayed. In some cases not only the sample code number is duplicated, but every attribute of the instance is the exact same as another instance. In tables 3 and 4 the rows with the instances with sample code numbers that have an exact copy are colored red.

When inspecting these tables it can be seen that the duplicated data is sometimes in consecutive rows, but not always. It can also be seen that most of the instances with duplicated sample code numbers have the same class label, but not always. Most duplicates come in pairs, but they also come in bigger groups, up to 6 instances with the same sample code number. Since no reason can be found as to why these double sample code numbers and instances exist, it can not be verified that these samples are not from the same origin. Therefore the choice has been made to remove all but one of every duplicate sample code number to guarantee the uniqueness of every sample.

3.2 Missing data

In table 5 all the instances that have at least one missing attribute are shown. It can be seen that there are 16 instances that do not have a complete record, all of them missing the *Bare Nuclei* attribute. Since this concerns only a fraction of the total number of instances (less than $\frac{1}{40}$) and since it is undesirable that missing attributes will influence the data mining the choice has been made to delete these instances from the data set.

3.3 The order of removing data

The missing data will be removed from the data set before the instances with duplicated sample code numbers are removed. This is done so that when one of the instances with a duplicated sample code number has missing data the other instance of this sample code number can be kept in the data. If it were to be done the other way around it could happen that an instance with a duplicated sample code number with a full record would be removed and an instance with missing data kept in the data, only for the instance with the missing data to be removed in the next processing step. After these filtering steps there are 630 instances left in the data set.

Table 3: Instances with duplicate sample code number (the rows with the instances with sample code numbers that have an exact copy are colored red)

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
268	320675	3	3	5	2	3	10	7	1	1	Malignant
273	320675	3	3	5	2	3	10	7	1	1	Malignant
270	385103	1	1	1	1	2	1	3	1	1	Benign
576	385103	5	1	2	1	2	1	3	1	1	Benign
272	411453	5	1	1	1	2	1	3	1	1	Benign
608	411453	1	1	1	1	2	1	1	1	1	Benign
684	466906	1	1	1	1	2	1	1	1	1	Benign
685	466906	1	1	1	1	2	1	1	1	1	Benign
372	493452	1	1	3	1	2	1	1	1	1	Benign
373	493452	4	1	2	1	2	1	2	1	1	Benign
291	560680	1	1	1	1	2	1	1	1	1	Benign
375	560680	3	1	2	1	2	1	2	1	1	Benign
690	654546	1	1	1	1	2	1	1	1	8	Benign
691	654546	1	1	1	3	2	1	1	1	1	Benign
578	695091	1	1	1	1	2	1	2	1	1	Benign
692	695091	5	10	10	5	4	5	4	4	1	Malignant
315	704097	1	1	1	1	1	1	2	1	1	Benign
339	704097	1	1	1	1	1	1	2	1	1	Benign
322	733639	3	1	1	1	2	NA	3	1	1	Benign
323	733639	3	1	1	1	2	1	3	1	1	Benign
443	734111	1	1	1	3	2	3	1	1	1	Benign
444	734111	1	1	1	1	2	2	1	1	1	Benign
526	769612	3	1	1	2	2	1	1	1	1	Benign
527	769612	4	1	1	1	2	1	1	1	1	Benign
338	798429	1	1	1	1	2	1	3	1	1	Benign
528	798429	4	1	1	1	2	1	3	1	1	Benign
345	822829	7	6	4	8	10	10	9	5	3	Malignant
613	822829	8	10	10	10	6	10	10	10	10	Malignant
698	897471	4	8	6	4	3	4	10	6	1	Malignant
699	897471	4	8	8	5	4	5	10	4	1	Malignant
5	1017023	4	1	1	3	2	1	3	1	1	Benign
253	1017023	6	3	3	5	3	10	3	5	3	Benign
9	1033078	2	1	1	1	2	1	1	1	5	Benign
10	1033078	4	2	1	1	2	1	2	1	1	Benign
536	1061990	1	1	3	2	2	1	3	1	1	Benign
619	1061990	4	1	1	1	2	1	2	1	1	Benign
30	1070935	1	1	3	1	2	1	1	1	1	Benign
31	1070935	3	1	1	1	1	1	2	1	1	Benign
43	1100524	6	10	10	2	8	10	7	3	3	Malignant
254	1100524	6	10	10	2	8	10	7	3	3	Malignant
48	1105524	1	1	1	1	2	1	2	1	1	Benign
469	1105524	4	1	1	1	2	1	1	1	1	Benign
388	1114570	5	3	3	2	3	1	3	1	1	Benign
389	1114570	2	1	1	1	2	1	2	2	1	Benign
62	1115293	1	1	1	1	2	2	2	1	1	Benign
491	1115293	1	1	1	1	2	1	1	1	1	Benign
63	1116116	9	10	10	1	10	8	3	3	1	Malignant
255	1116116	9	10	10	1	10	8	3	3	1	Malignant
65	1116192	1	1	1	1	2	1	2	1	1	Benign
538	1116192	5	1	2	1	2	1	3	1	1	Benign

Table 4: Instances with duplicate sample code number (the rows with the instances with sample code numbers that have an exact copy are colored red) continued

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
82	1143978	4	1	1	2	2	1	2	1	1	Benign
83	1143978	5	2	1	1	2	1	3	1	1	Benign
94	1158247	1	1	1	1	2	1	2	1	1	Benign
394	1158247	1	1	1	1	1	1	1	1	1	Benign
105	1168736	10	10	10	10	10	1	8	8	8	Malignant
256	1168736	5	6	6	2	4	10	3	6	1	Malignant
109	1171710	1	1	1	1	2	1	2	3	1	Benign
110	1171710	6	5	4	4	3	9	7	8	3	Malignant
116	1173347	1	1	1	1	2	5	1	1	1	Benign
117	1173347	8	3	3	1	2	2	3	2	1	Benign
121	1174057	1	1	2	2	2	1	3	1	1	Benign
122	1174057	4	2	1	1	2	2	3	1	1	Benign
137	1182404	4	1	1	1	2	1	2	1	1	Benign
257	1182404	3	1	1	1	2	1	1	1	1	Benign
258	1182404	3	1	1	1	2	1	2	1	1	Benign
266	1182404	5	1	4	1	2	1	3	2	1	Benign
449	1182404	1	1	1	1	1	1	1	1	1	Benign
498	1182404	4	2	1	1	2	1	1	1	1	Benign
169	1198641	3	1	1	1	2	1	3	1	1	Benign
259	1198641	3	1	1	1	2	1	3	1	1	Benign
267	1198641	10	10	6	3	3	10	4	3	2	Malignant
195	1212422	3	1	1	1	2	1	3	1	1	Benign
196	1212422	4	1	1	1	2	1	3	1	1	Benign
208	1218860	1	1	1	1	1	1	3	1	1	Benign
209	1218860	1	1	1	1	1	1	3	1	1	Benign
472	1238777	6	1	1	3	2	1	1	1	1	Benign
633	1238777	1	1	1	1	2	1	1	1	1	Benign
548	1240603	2	1	1	1	1	1	1	1	1	Benign
549	1240603	3	1	1	1	1	1	1	1	1	Benign
242	1276091	3	1	1	3	1	1	3	1	1	Benign
430	1276091	2	1	1	1	2	1	2	1	1	Benign
431	1276091	1	3	1	1	2	1	2	2	1	Benign
432	1276091	5	1	1	3	4	1	3	2	1	Benign
463	1276091	6	1	1	3	2	1	1	1	1	Benign
639	1277792	4	1	1	1	2	1	1	1	1	Benign
640	1277792	5	1	1	3	2	1	1	1	1	Benign
434	1293439	3	2	2	3	2	1	1	1	1	Benign
435	1293439	6	9	7	5	5	8	4	2	1	Benign
468	1299596	6	6	6	5	4	10	7	6	2	Malignant
645	1299596	2	1	1	1	2	1	1	1	1	Benign
512	1299924	5	1	1	1	2	1	2	1	1	Benign
553	1299924	3	2	2	2	2	1	4	2	1	Benign
517	1320077	1	1	1	1	1	1	1	1	1	Benign
518	1320077	1	1	1	1	1	1	2	1	1	Benign
561	1321942	5	1	1	1	2	1	3	1	1	Benign
562	1321942	5	1	1	1	2	1	3	1	1	Benign
661	1339781	1	1	1	1	2	1	2	1	1	Benign
662	1339781	4	1	1	1	2	1	3	1	1	Benign
673	1354840	2	1	1	1	2	1	3	1	1	Benign
674	1354840	5	3	2	1	3	1	1	1	1	Benign

Table 5: Instances with missing data from the Breast Cancer Wisconsin (Original) Data Set

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
24	1057013	8	4	5	1	2	NA	7	3	1	Malignant
41	1096800	6	6	6	9	6	NA	7	8	1	Benign
140	1183246	1	1	1	1	1	NA	2	1	1	Benign
146	1184840	1	1	3	1	2	NA	2	1	1	Benign
159	1193683	1	1	2	1	3	NA	1	1	1	Benign
165	1197510	5	1	1	1	2	NA	3	1	1	Benign
236	1241232	3	1	4	1	2	NA	3	1	1	Benign
250	169356	3	1	1	1	2	NA	3	1	1	Benign
276	432809	3	1	3	1	2	NA	2	1	1	Benign
293	563649	8	8	8	1	2	NA	6	10	1	Malignant
295	606140	1	1	1	1	2	NA	2	1	1	Benign
298	61634	5	4	3	1	2	NA	2	3	1	Benign
316	704168	4	6	5	6	7	NA	4	9	1	Benign
322	733639	3	1	1	1	2	NA	3	1	1	Benign
412	1238464	1	1	1	1	1	NA	2	1	1	Benign
618	1057067	1	1	1	1	1	NA	1	1	1	Benign

3.4 Data distribution

3.4.1 Class distribution

It is important to look at the class distribution because studies have shown that the distribution of the class labels can have an effect on classifier learning, and that the natural class distribution does not always give the best classifiers [31][32]. According to [31] there are several explanations for why the minority class generally has a higher error rate than the majority class when using unbalanced data while training a classifier.

Ways to handle the unbalanced data while making a classifier include under-sampling of the majority class and over-sampling of the minority-class. Another way is to tackle this problem is to use a cost-sensitive classifier that gives a heavier weight to misclassifying the minority class. These techniques have their own advantages and drawbacks [33].

When using over- or under-sampling techniques it is important to keep in mind that this will result in a bias in the model, this bias will cause the over-sampled class to be predicted too often. This bias will improve the performance of the classifier on the over-sampled class, but the overall performance will deteriorate due to this bias. To compensate for this bias a correction has to be built into the model, one way of using a correction is shown in [31]. When using these techniques it is also important to keep in mind that class imbalance is a relative problem that does not only depend on the degree of class imbalance, but also on the complexity of the concept representing the data, the size of the training set and also on the classifier involved. When using a classifier that is not susceptible to the class imbalance problem the use of over- and under-sampling could hurt the classifier instead of helping it [32].

In figure 1 the class distribution of the Breast Cancer Wisconsin (Original) Data Set can be seen before and after filtering. It can be seen that the data before filtering as well as the data after filtering seems to be unbalanced, it has a minority and majority class. However the data does not seem to be extremely unbalanced. It can also be seen that the data after the filtering step is slightly more balanced than before filtering.

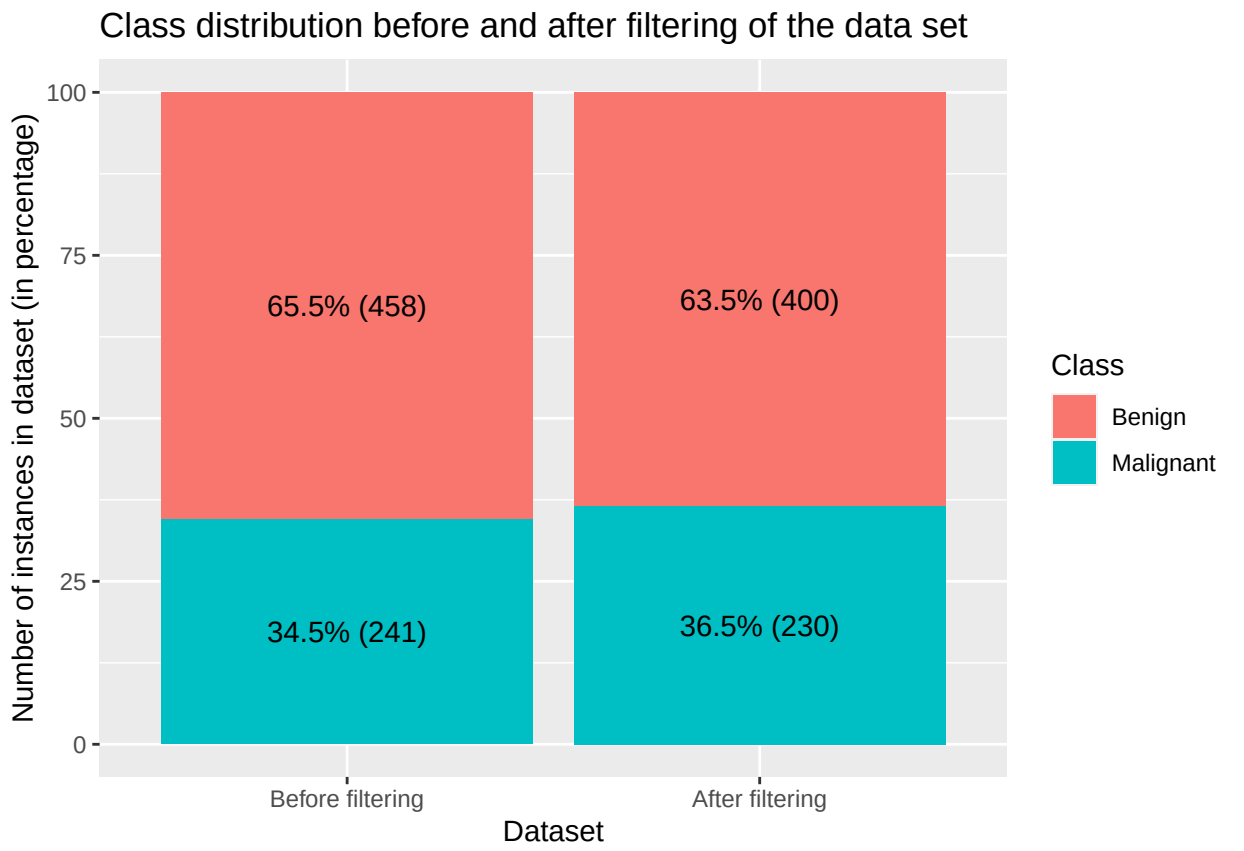

Figure 1: Distribution of the class attribute of the data before and after the filtering steps (the filtering steps are: removing rows with missing data, and than removing duplicated sample code numbers). The numbers between parentheses are the actual number of instances in the data sets.

3.4.2 Attribute distributions

For the attributes to be useful for building a machine learning classifier it is important that the distributions for these attribute are discriminative for the different classes. The literature already states that the nine attributes involved in the *Breast Cancer Wisconsin (Original) Data Set* are significantly different between benign and malignant cases [4]. Figure 2 shows a visual representation of the attribute distributions for benign and malignant samples, as well as the distribution for all the samples together. When looking at this figure it is important to keep in mind that the majority class (the benign instances) have a bigger influence on the overall distribution than the minority class.

All the attributes seem to be very differently distributed between the benign and malignant instances, except for the mitosis attribute. For both the benign and malignant cases the mitosis attribute most often has a score of 1. However when looking at the mitosis distribution of the malignant instances, there is a longer tail towards the higher score than the benign cases show, this might still be a significant difference.

The seemingly (big) difference in distributions for all of these attributes is a positive sign for machine learning, all of these attributes could be useful in differentiating benign and malignant samples from one another.

To verify that the difference is indeed significant for all the attributes a one-sided Mann–Whitney U test is executed for each attribute. The results of these tests and the corresponding p values can be found in table 6. The tests have the following hypotheses:

- Null hypothesis: the two samples (benign and malignant) come from the same population.

- Alternative hypothesis: observations in the malignant sample tend to be higher than observations in the benign sample (the malignant sample is shifted to the right compared to the benign sample).

With $\alpha = 0.05$ the null hypothesis is rejected for all the tests, and the alternative hypothesis is accepted. All of the differences might be significant, when looking at the estimate median of difference (that is the estimated median of the difference between all the observations from one sample and all the observations in another sample, and not the estimated difference in medians between the two samples) it is once again obvious that the difference in mitosis is quite small. The difference in all the other attributes seems quite large, especially the bare nuclei attribute. This could mean that the mitosis attribute is less useful for machine learning than the other attributes.

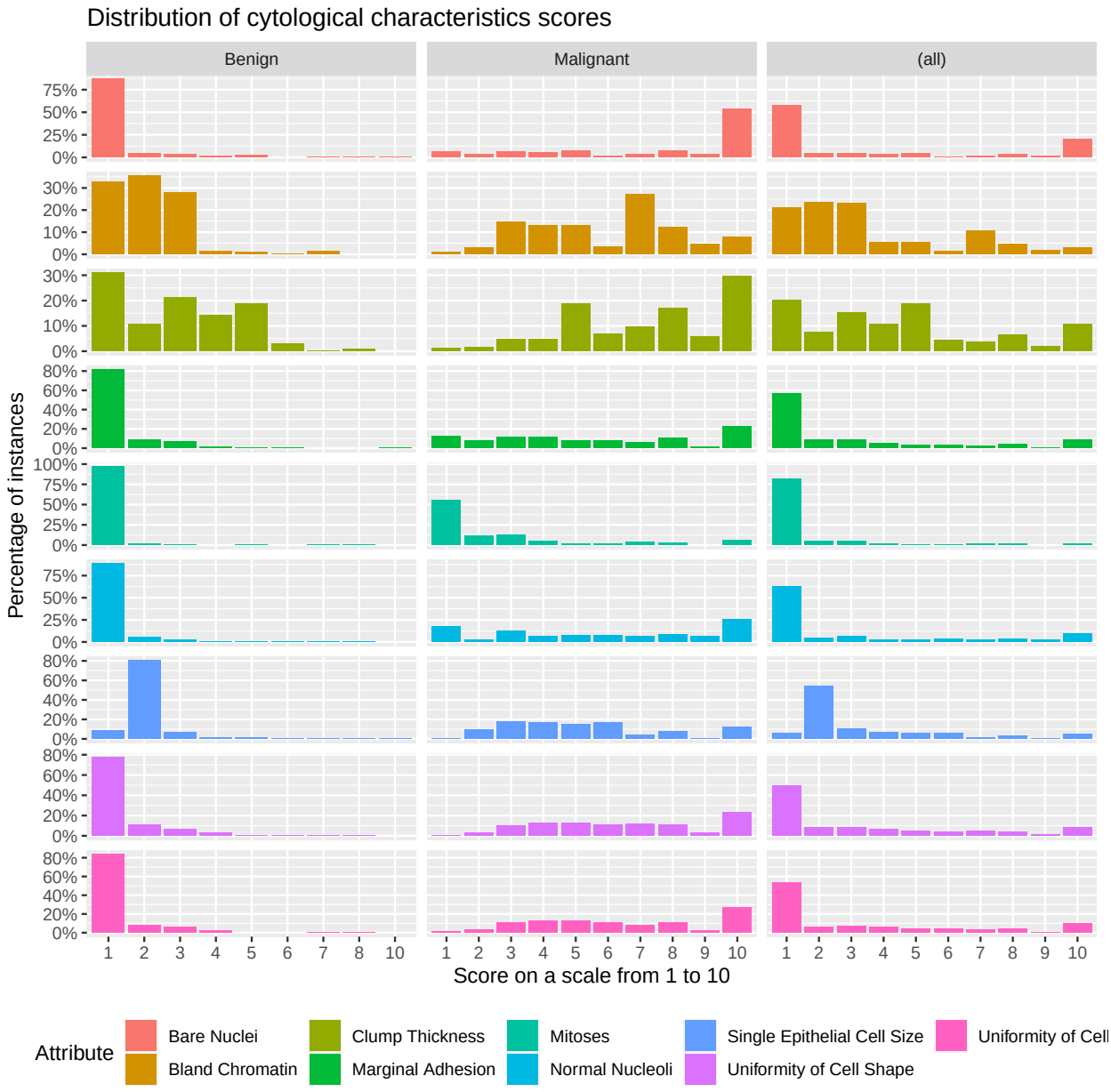

Figure 2: Distribution of data in percentage for 9 different cytological characteristics for benign instances, malignant instances and for all instances together

Table 6: Results of one-sided Mann–Whitney U test for each attribute where the null hypothesis is that the distribution of the malignant samples **is not** higher than that of the benign samples. And the alterernative hypothesis is that the distribution of the Malignant samples **is** higher than that of the benign samples. All of the p-values are well below 0.05 so we reject the null hypothesis for each attribute.

Attribute name	Estimate median of difference	P value
Clump Thickness	4.0000287732	1.736520e-68
Uniformity of Cell Size	5.0000208895	0.000000e+00
Uniformity of Cell Shape	5.0000340393	3.000000e-100
Marginal Adhesion	4.0000478861	1.197571e-77
Single Epithelial Cell Size	2.9999504915	1.339529e-84
Bare Nuclei	7.9999934047	8.290000e-98
Bland Chromatin	4.0000370221	3.375664e-79
Normal Nucleoli	4.9999897428	2.285017e-79
Mitoses	0.0000555406	1.003309e-39

3.5 Correlation between attributes

It is important to investigate the correlation between the attributes as some machine learning algorithms, Naive Bayes for example, can be influenced by this correlation. In figure 3 it can be seen that all the attribute in the data set have a positive correlation score. The correlation score shown in the figure is the Kendall τ_b rank correlation coefficient. The strength of the correlation is moderate (0.33 between mitoses and bland chromatin) to strong (0.82 between uniformity of cell shape and uniformity of cell size). This correlation can be problematic for algorithms that make assumptions about the independence of the attributes. Naive Bayes does assume that the attributes are independent. The calculations done by the Naive Bayes algorithm are based on calculations with conditional probability that are not accurate when the attributes depend on each other. Also important to note is that all the correlation coefficients that were calculated had a p value below 0.05, and can thus be considered significant. These p values can be found in the log file available in the Thema09: Building a classifier with Weka repository [29].

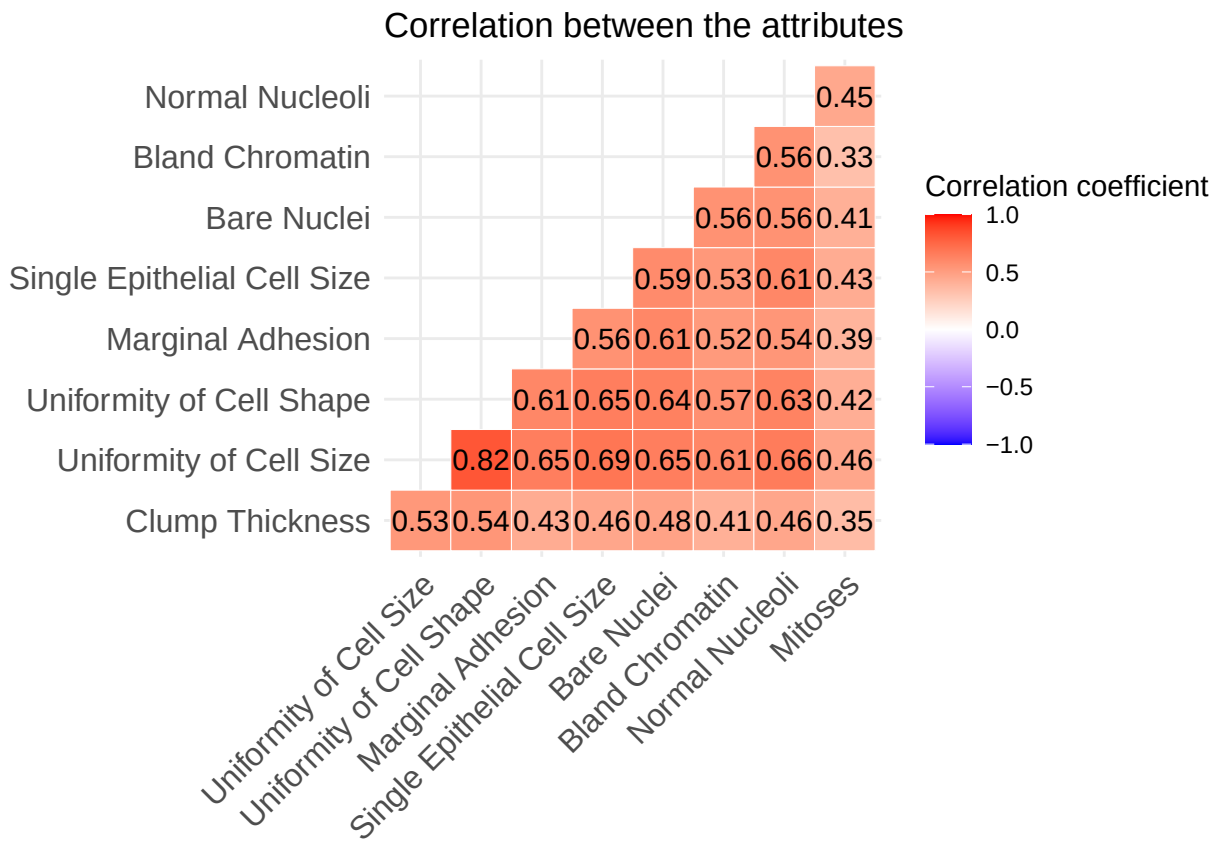

Figure 3: Correlations coefficient (Kendall's τ_b) scored between the attribute Breast Cancer Wisconsin (Original) Data Set in the between the attributes. The p value is calculated, and if the correlation is not significant, the in the plot show an X in the corresponding square.

3.6 Principal component Analysis

To visualize the data in a two dimensional plot a PCA is done. The PCA makes it possible to capture most of the variance of the nine dimensional data in two dimensions, the first and the second principal component (PC1 and PC2). During PCA calculations are done that give the rotation that the the multidimensional data needs to undergo to get the principal components. PC1 is the direction in which the most variance is captured, PC2 will be perpendicular to PC1 and explain the second most variance, PC3 will be perpendicular to PC1 and PC2 and explains the third most variance and so on, in total there are the same amount of PC's as there are dimensions in the data. The rotation matrix that was the outcome of the PCA is shown in table 7. Looking at this table shows how much each attribute contributed to each PC. The higher the absolute value of the rotation on a specific PC by a specific attribute tells how much that attribute contributes to this PC and how much this attributes is correlated with that PC. In the matrix it can be seen that all the attribute contribute roughly evenly to PC 1, except for mitosis. Mitoses contributes less to PC1, but dominates PC2. Mitoses is also the attribute that has a weakest correlation to the other attributes.

In figure 4 a plot is show of PC1 and PC2 of the data set. The arrows show the different attributes, the arrow of mitosis is much more vertical than any of the other attributes, which shows ones again how it dominates PC2, but does not have as much of an influence in PC1. The instances are colored according to their class label. This coloring shows that these groups of instances are fairly distinct, which might indicate that this data set is well suited for machine learning.

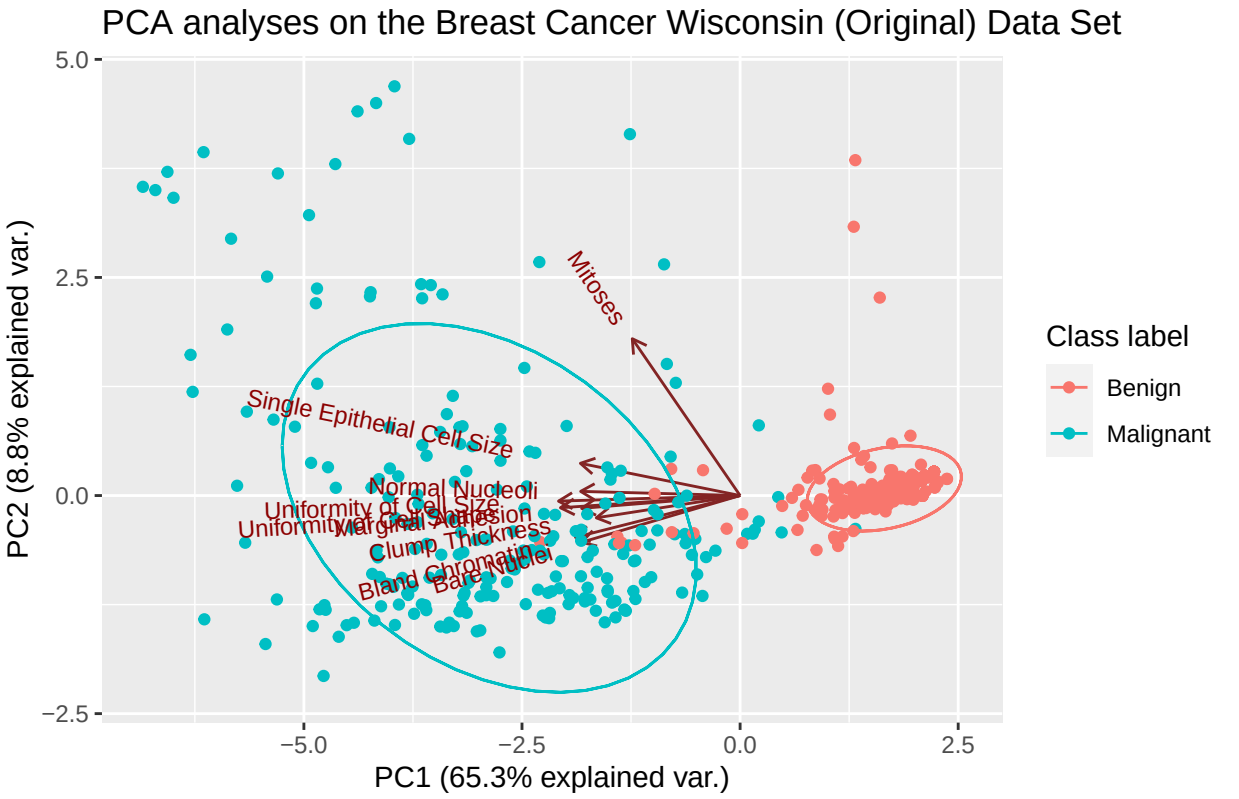

Figure 4: Principal component 1 and principal component 2 of the Breast Cancer Wisconsin (Original) Data Set. The instances are coloured according to their class label. The attributes that are present in the data set are shown with arrows.

Table 7: PCA: Rotation matrix of the 9 cytological characteristics to each principal component

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Clump Thickness	-0.3026067	-0.1277601	0.8730402	-0.0509633	0.0200015	0.2080380	0.0594103	0.2830144	-0.0018746
Uniformity of Cell Size	-0.3821300	-0.0320382	-0.0386545	0.1856231	-0.1336961	0.2422388	-0.1310226	-0.4164032	-0.7415439
Uniformity of Cell Shape	-0.3780364	-0.0701172	0.0287437	0.1593851	-0.0888896	0.1701730	-0.0535618	-0.5990616	0.6537113
Marginal Adhesion	-0.3333947	-0.0734380	-0.3931759	-0.5003666	0.0309380	0.5636711	0.3050514	0.2529206	0.0528526
Single Epithelial Cell Size	-0.3360831	0.1847634	-0.1429394	0.3430707	-0.7025839	-0.1980931	0.1571675	0.3895393	0.0739862
Bare Nuclei	-0.3335249	-0.2678899	0.0273743	-0.5168253	-0.0492067	-0.6806530	0.1963920	-0.1930489	-0.0871259
Bland Chromatin	-0.3465219	-0.2393446	-0.1911542	0.0159296	0.2005512	-0.1029457	-0.7827851	0.3403353	0.0802731
Normal Nucleoli	-0.3355534	0.0240361	-0.1277002	0.4830818	0.6411883	-0.1749484	0.4228656	0.1273752	-0.0195346
Mitoses	-0.2268878	0.8992081	0.0828547	-0.2621999	0.1596186	-0.0873072	-0.1689347	-0.0510517	0.0093337

3.7 The classifier

While building the classifier an assortment of machine learning algorithms were tested on the data set. The test were done in the Weka Experimenter using 10-fold cross validation and 10 repetitions for the tests. An in depth description of which algorithms were tested and their evaluation can be found in the EDA_log_NaomiHindriks.pdf file that is available in the Thema09: Building a classifier with Weka repository [29]. In this log file a description can be found of how the final classifier, that reached an accuracy of over 98%, was build.

The final algorithm is assembled with a cost sensitive learning algorithm (giving 5 times as much weight to the false negative over the false positives) applied to a voting ensemble learner. That ensemble learner lets the following algorithms vote: IBk (with KNN = 2), Naive Bayes (with useSupervisedDiscretization = True) and Random Forest (with default settings in Weka). The quality metrics that were calculated for this algorithm can be found in table 8. In this table it can be seen that the classifier reaches an accuracy of 98.048%, an AUC of 0.994, and an F_2 score of 0.987. While running the test on the data set using 10-fold cross validation on average over 10 runs the classifier only gave 1 false negative result.

In figure 5 the ROC curve of this algorithm can be seen. In this figure it can be seen that the TPR can even get one step higher if the threshold was changed. That one step higher would mean 0 false negatives and a TPR of 1. This is achieved around a true positive rate of around 0.8 (on the x axis around 0.2 false positive rate). Setting the algorithm to this threshold would mean never missing a malignant sample in the filtered Breast Cancer Wisconsin (Original) Data Set. It would also mean that 20 % of benign samples would be classified as malignant.

Table 8: Results of experiment run in Weka Experimenter of the classifier (cost sensitive learning applied to a voting ensemble learner letting the following algorithms vote: IBk (with KNN = 2), Naive Bayes (with useSupervisedDiscretization = True) and Random Forest). The classifier was used on the filtered Breast Cancer Wisconsin (Original) Data Set. The experiment was run using 10-fold cross validation and the iteration was set to 10 repetitions.

	Time			Confusion matrix							
Settings	Training	Testing	ACC	TP	FP	TN	FN	TPR	TNR	AUC	F_2
Voting, cost sensitive learning 1:5	0.2532	0.0243	98.048	229.00	11.30	388.70	1.00	0.996	0.972	0.994	0.987

Column explanation:
ACC = accuracy (%)
TP = true positive
FP = false positive
TN = true negative
FN = false negative
TPR = true positive rate = sensitivity = recall
TNR = true negative rate = specificity
AUC = area under the ROC curve
F_2 = the F_β score with $\beta = 2$

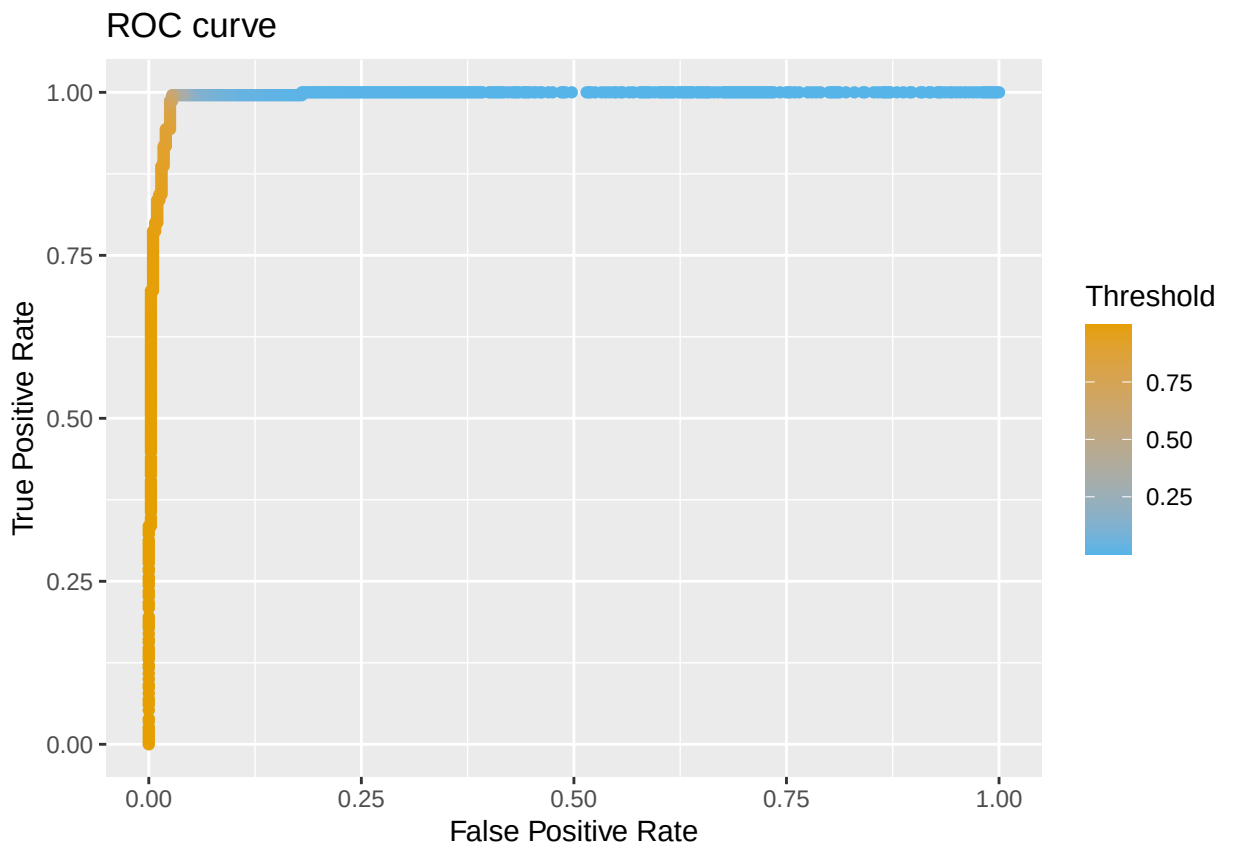

Figure 5: ROC curve of most optimized classifier algorithm use on the filtered Breast Cancer Wisconsin (Original) Data Set. The algorithm that is used is a classifier made and run in the Weka Explorer: the cost sensitive learning algorithm (cost 1:5 for FP:FN) wrapping the voting algorithm. The algorithms that are used for voting are: the IBk (KNN = 2) algorithm, the Naive Bayes (useSupervisedDiscretization = True) and the Random Forest algorithm.

In figure 6 a couple of learning curves of the algorithm can be seen. It shows how much the quality metrics increase when the algorithm is trained on more data. It can be seen that accuracy, AUC and F_2 score pretty high scores even when training with only 25% of the training data, but that the FNR score seems to have a clear downward trend even between 50 to 100%. This probably reflect that using less data will miss more of the border cases that are malignant. Since the algorithm does not take a large amount of time to train, or to test data (see table 8), it is probably wise to use all the data for training and maybe even collect more data for training to enhance the performance of the algorithm even more. Using more data will probably make the algorithm slower in testing, due to the usage of the IBk algorithm. This algorithm will check the distance to each training instance for a newly presented instance. But since diagnosing breast cancer won't stand or fall with an extra minute, hour or even day of testing time, it could be interesting how the algorithm would perform when presented with more training data.

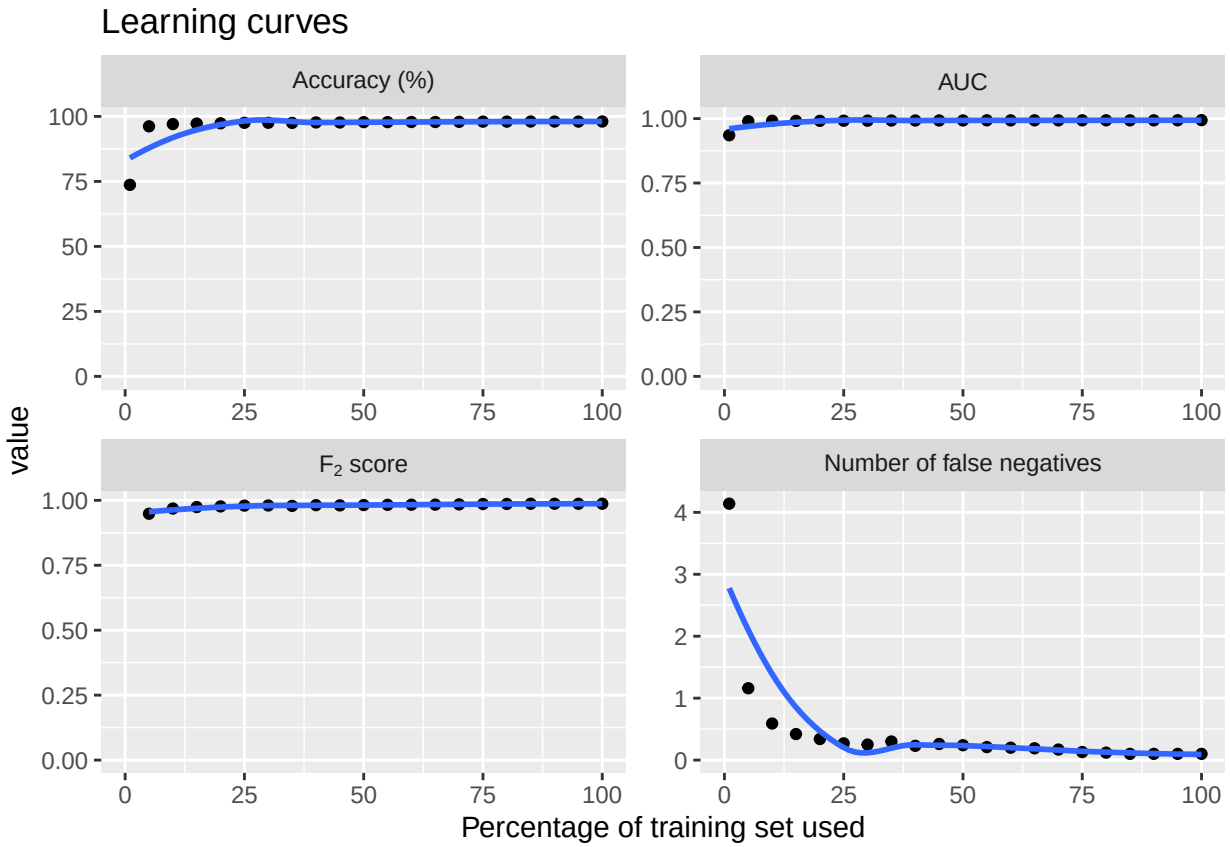

Figure 6: Different quality metrics scored for different percentages of the trainingset used. Results are from the optimal algorithm run in the Weka Experimenter using 10-fold cross validation and number of repititions set to 10

4 Discussion & Conclusion

During the analyses of the data it turned out that there were a lot of weird duplications going on. It is important that these duplications were filtered out for the validity of the classifier that was build, if they were not filtered out some samples might have had more influence on the classifier than they deserved. When looking at the distribution of the attributes it was shown that the malignant distribution significantly differed from the benign distribution. A moderate to strong correlation was found between the different attributes. During the PCA a PC plot (pf PC1 and PC2) was made that clearly showed two distinct clusters of data.

The final algorithm is set to a threshold so that it is very close to the (0, 1) coordinate, which means that it only miss classifies 1 malignant sample as benign out of the whole data set. This and the other quality metrics of the classifier might seem very pleasing, but it is important to keep in mind that while diagnosing cancer the lives of actual people are at stake. Miss diagnosing 1 in 230 malignant cases in the small Breast Cancer Wisconsin (Original) Data Set might seem very good, but miss diagnosing this rate of women out of the 2,261,419 new cases of diagnosed female breast cancer per year would mean missing 9832 cases of breast cancer, quite possibly resulting in premature death of this massive amount of women. This means that it is ALWAYS important to not use this classifier as a sole diagnostic tool, but always be aware of other possible signs of malignancy.

Even though an assortment of algorithms have been tested on this data set, there are a lot more machine learning algorithms, and settings for these algorithms that might work better on this data set, or can be added to the voting algorithm used to increase the performance of this classifier. The search for the optimal algorithm used for the classifier was not an exhaustive one, merely an indication of how good these nine cytological characteristics can be in diagnosing breast cancer.

In the data used in this report there was no record of different types of cancer. It would be interesting to see if the type of cancer could be predicted by these nine cytological characteristics as well. Another interesting attribute would be if the benign samples would later turn into malignant samples, and how long it would take to turn malignant, this could especially be useful for the border cases, if it turns out that most of the border cases would later turn out to become malignant early intervention could even prevent these women from developing cancer. It would also be interesting to see if other cytological characteristics of the FNA samples could be found that are also predictors for malignancy and could enhance the performance of this classifier further.

5 Minor proposal

For the minor Application Design it could be interesting to build a web application with which this classifier can be used. Right now there is only a command line application available to use this classifier. A lot of people that are not bioinformaticians might not be so comfortable to use a command line interface. For them it would be useful to make a web application with a clear and friendly user interface where they can classify their own instances without being scared away by technical stuff. This would also have the advantage that the users of the web application do not have to download an application, they could simply visit a website. This could especially be useful for doctors involved in the diagnoses of breast cancer.

In this web app the instance could be filled in with a simple text field or an ARFF file could be submitted (just like the command line application). Then the website should return the classification of the instance(s) and report on the probability that this classification is correct. When an ARFF file is submitted more statistics of the classified instances can be shown on the web application (for example the percentage of instances in the file that is classified as benign or malignant). Another feature that could be added to this web application is that users are able to make an account on the website and save results, so they can be viewed later or shared with other people.

6 References

[1] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[2] R. Etzioni *et al.*, "THE case for early detection," *Nature Reviews Cancer*, vol. 3, no. 4, pp. 243–253, 2003, doi: 10.1038/nrc1041.

[3] M. Wu and D. E. Burstein, "Fine needle aspiration," *Cancer Investigation*, vol. 22, no. 4, pp. 620–628, 2004, doi: 10.1081/CNV-200027160.

[4] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 23, pp. 9193–9196, 1990, doi: 10.1073/pnas.87.23.9193.

[5] R. Merzouki, *User manual breast cancer diagnosis web user interface*, 1st ed. Available: https://www.rai-light.com/docs/BCD_User_Manual_v01.pdf

[6] W. H. Wolberg, "UCI machine learning repository: Breast cancer wisconsin (original) data set." https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29

[7] R Core Team, *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021.Available: https://www.R-project.org/

[8] I. H. Witten, E. Frank, M. Hall, and C. Pal, *The weka workbench. Online appendix for "data mining: Practical machine learning tools and techniques"*. Morgan Kaufmann, 2016.

[9] K. Arnold, J. Gosling, and D. Holmes, *The java programming language*. Addison Wesley Professional, 2005.

[10] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[11] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[12] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

[13] K. P. F. R. S., "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901, doi: 10.1080/14786440109462720.

[14] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63–91, 1993.

[15] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh conference on uncertainty in artificial intelligence*, 1995, pp. 338–345.

[16] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," vol. 95, nos. 1-2, pp. 161–205, 2005.

[17] M. Sumner, E. Frank, and M. Hall, "Speeding up logistic model tree induction," in *9th european conference on principles and practice of knowledge discovery in databases*, 2005, pp. 675–683.

[18] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in kernel methods - support vector learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998.Available: http://research.microsoft.com//~jplatt/smo.html

[19] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to platt's smo algorithm for svm classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637–649, 2001.

[20] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," in *Advances in neural information processing systems*, 1998, vol. 10.

[21] D. Aha and D. Kibler, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

[22] R. Quinlan, *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Thirteenth international conference on machine learning*, 1996, pp. 148–156.

[25] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[26] L. I. Kuncheva, *Combining pattern classifiers: Methods and algorithms*. John Wiley; Sons, Inc., 2004.

[27] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[28] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.

[29] N. J. Hindriks, "Thema09: Building a classifier with weka," *GitHub repository*. https://github.com/naomihindriks/thema09; GitHub, 2022.

[30] N. J. Hindriks, "Breast cancer classifier," *GitHub repository*. https://github.com/naomihindriks/java_wrapper; GitHub, 2022.

[31] G. M. Weiss and F. Provost, "The effect of class distribution on classifier learning: An empirical study," Rutgers University, 2001. doi: 10.7282/t3-v9kt-9510.

[32] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002, doi: 10.3233/IDA-2002-6504.

[33] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.