Breast Cancer Diagnosis Project

Olivier Paratte

28/01/2021

1 Introduction

1.1 Overview

This project is part of the HarvardX Data Science Capstone course and uses the Breast Cancer Wisconsin (Diagnostic) Data Set to build a prediction model using various machine learning methods.

This report describes the various steps taken to analyse the data, the methods used to build the models and the evaluation of the models performance, it includes 5 sections:

1. Introduction
2. Data exploration and analysis
3. Model creation and validation
4. Results
5. Conclusion

1.2 Projet goal

The goal of this project is to create and test various models to predict whether a sample collected during a biopsy done using the fine needle aspirate (FNA) procedure is cancerous or not and indentify the one(s) with the best performance.

For this project, the main metrics used to measure the performance of the diagnostic predictions are accuracy (proportion of correct prediction) and sensitivity (rate of false negatives).Trying to reach the maximum accuracy is rather obvious. You want the model to make as many correct predictions as possible. Sensitivity was selected second key metrics because of the potentially deadly consequences of a false negative test (not treating someone with breast cancer). Those metrics and others used to evaluate the models will be explained in more details in section 3.

1.3 Background information

Fine needle aspiration (FNA) is a type of biopsy where a small amount of tissue or fluid from a suspicious area is withdrawn (aspirated) with a very thin, hollow needle attached to a syringe and is checked for cancer cells. Fine needle aspiration is relatively non-invasive, generally considered a safe procedure, complications are infrequent and the entire procedure from start to finish generally takes around 30 minutes. The biopsy sample may be examined under a microscope by a pathologist who will make a diagnostic and/or send to a lab for testing.

2 Data exploration and analysis

2.1 Dataset infromation

The features of the Breast Cancer Wisconsin (Diagnostic) data set used for this project are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Data set link:

http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data

Data set description:

https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names

2.2 Dataset dimensions

The Breast Cancer Wisconsin (Diagnostic) data set has 32 columns and 569 rows

[1] "they are no missing values in the data set"

Each row corresponds to one biopsy and the features computed from one digitised image of a fine needle aspirate (FNA) of a breast mass as well as the diagnosis (M = malignant, B = benign). The column "diagnosis" is the outcome we want to predict. Here are the features and their characteristics:

Data set structure:

```
## Classes 'data.table' and 'data.frame':   569 obs. of  32 variables:
##  $ id                     : int  842302 842517 84300903 84348301 84358402 843786 844359 84458202 844
##  $ diagnosis              : chr  "M" "M" "M" "M" ...
##  $ radius_mean            : num  18 20.6 19.7 11.4 20.3 ...
##  $ texture_mean           : num  10.4 17.8 21.2 20.4 14.3 ...
##  $ perimeter_mean         : num  122.8 132.9 130 77.6 135.1 ...
##  $ area_mean              : num  1001 1326 1203 386 1297 ...
##  $ smoothness_mean        : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
##  $ compactness_mean       : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
##  $ concavity_mean         : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
##  $ concave_points_mean    : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
##  $ symmetry_mean          : num  0.242 0.181 0.207 0.26 0.181 ...
##  $ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
##  $ radius_se              : num  1.095 0.543 0.746 0.496 0.757 ...
##  $ texture_se             : num  0.905 0.734 0.787 1.156 0.781 ...
##  $ perimeter_se           : num  8.59 3.4 4.58 3.44 5.44 ...
##  $ area_se                : num  153.4 74.1 94 27.2 94.4 ...
##  $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
##  $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
##  $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
##  $ concave_points_se      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
##  $ symmetry_se            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
##  $ fractal_dimension_se   : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
##  $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
##  $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
##  $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
##  $ area_worst             : num  2019 1956 1709 568 1575 ...
##  $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
##  $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
##  $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
##  $ concave_points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
##  $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
##  $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

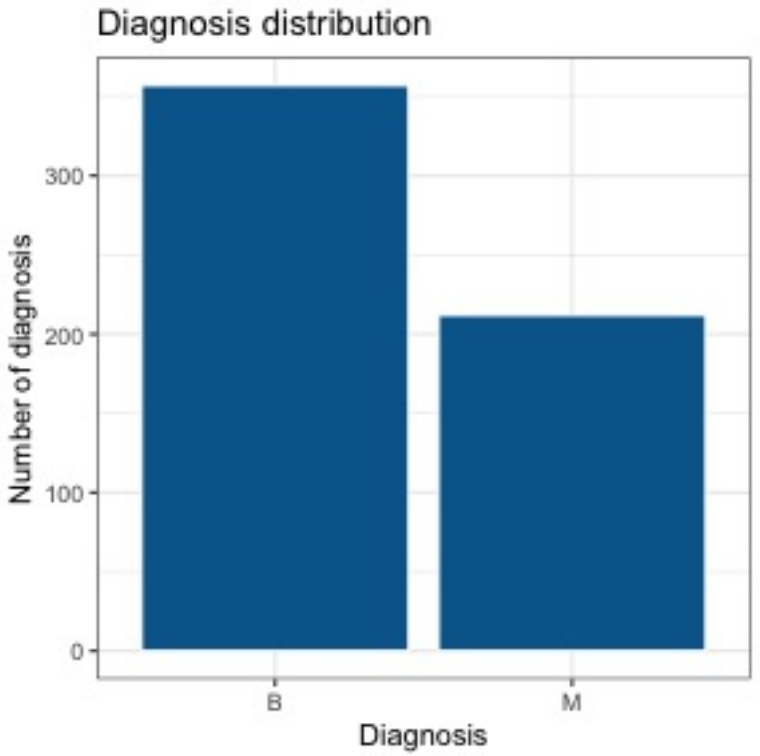
2.3 Data set attribute and features:

Attribute Information:

1. **id**
2. **diagnosis** (M = malignant, B = benign) Ten real-valued features are computed for each cell nucleus 3-32):
3. **radius**: Nucleus radius (mean of distances from center to points on the perimeter).
4. **texture**: Nucleus texture (standard deviation of gray-scale values).
5. **perimeter**: Nucleus perimeter.
6. **area**: Nucleus area.
7. **smoothness**: Nucleus smoothness (local variation in radius lengths).
8. **compactness**: Nucleus compactness (perimeterˆ2/area - 1).
9. **concavity**: Nucleus concavity (severity of concave portions of the contour).
10. **concave_pts**: Number of concave portions of the nucleus contour.
11. **symmetry**: Nucleus symmetry.
12. **fractal_dim**: Nucleus fractal dimension ("coastline approximation" - 1).

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, column 3 is Mean Radius, column 13 is Radius SE, column 23 is Worst Radius.

We can also see the the graph below that 357 of the diagnosis are benign and 212 are malignant.

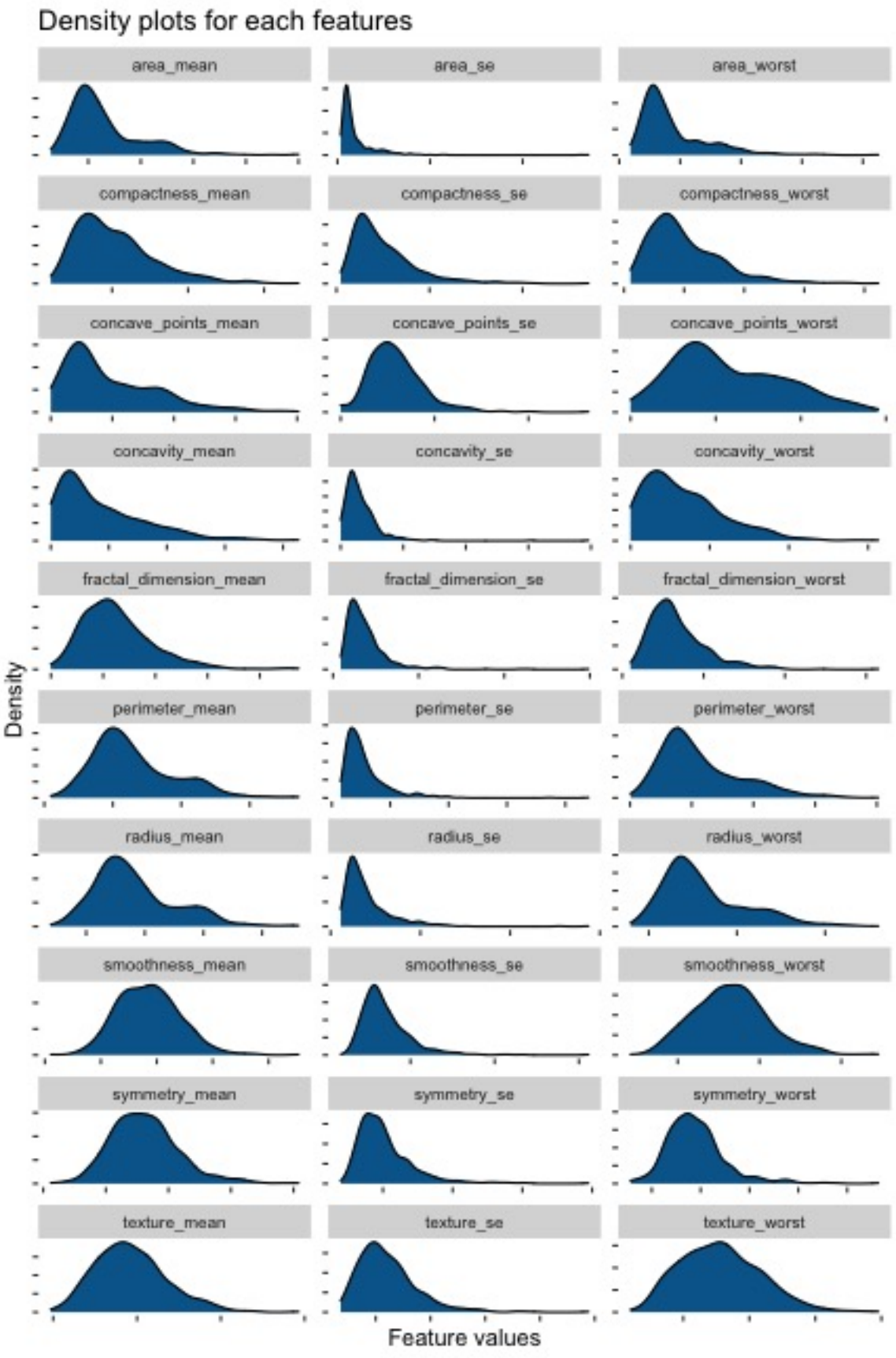


Features summary:

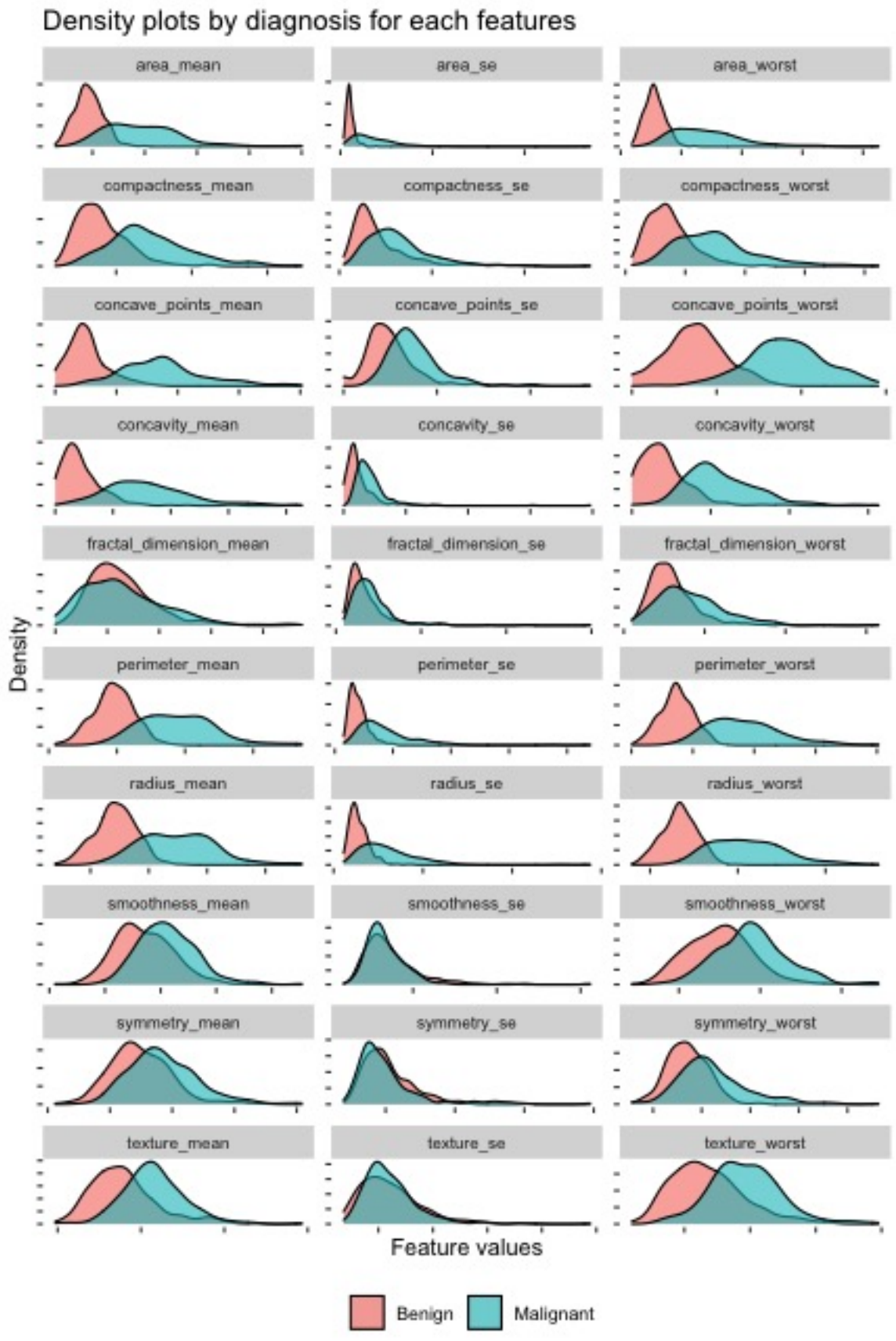
radius_mean	Min. : 6.981	1st Qu.:11.700	Median :13.370	Mean :14.127	3rd Qu.:15.780	Max. :28.110
texture_mean	Min. : 9.71	1st Qu.:16.17	Median :18.84	Mean :19.29	3rd Qu.:21.80	Max. :39.28
perimeter_mean	Min. : 43.79	1st Qu.: 75.17	Median : 86.24	Mean : 91.97	3rd Qu.:104.10	Max. :188.50
area_mean	Min. : 143.5	1st Qu.: 420.3	Median : 551.1	Mean : 654.9	3rd Qu.: 782.7	Max. :2501.0
smoothness_mean	Min. :0.05263	1st Qu.:0.08637	Median :0.09587	Mean :0.09636	3rd Qu.:0.10530	Max. :0.16340
compactness_mean	Min. :0.01938	1st Qu.:0.06492	Median :0.09263	Mean :0.10434	3rd Qu.:0.13040	Max. :0.34540
concavity_mean	Min. :0.00000	1st Qu.:0.02956	Median :0.06154	Mean :0.08880	3rd Qu.:0.13070	Max. :0.42680
concave_points_mean	Min. :0.00000	1st Qu.:0.02031	Median :0.03350	Mean :0.04892	3rd Qu.:0.07400	Max. :0.20120
symmetry_mean	Min. :0.1060	1st Qu.:0.1619	Median :0.1792	Mean :0.1812	3rd Qu.:0.1957	Max. :0.3040
fractal_dimension_mean	Min. :0.04996	1st Qu.:0.05770	Median :0.06154	Mean :0.06280	3rd Qu.:0.06612	Max. :0.09744
radius_se	Min. :0.1115	1st Qu.:0.2324	Median :0.3242	Mean :0.4052	3rd Qu.:0.4789	Max. :2.8730
texture_se	Min. :0.3602	1st Qu.:0.8339	Median :1.1080	Mean :1.2169	3rd Qu.:1.4740	Max. :4.8850
perimeter_se	Min. : 0.757	1st Qu.: 1.606	Median : 2.287	Mean : 2.866	3rd Qu.: 3.357	Max. :21.980
area_se	Min. : 6.802	1st Qu.: 17.850	Median : 24.530	Mean : 40.337	3rd Qu.: 45.190	Max. :542.200
smoothness_se	Min. :0.001713	1st Qu.:0.005169	Median :0.006380	Mean :0.007041	3rd Qu.:0.008146	Max. :0.031130
compactness_se	Min. :0.002252	1st Qu.:0.013080	Median :0.020450	Mean :0.025478	3rd Qu.:0.032450	Max. :0.135400
concavity_se	Min. :0.00000	1st Qu.:0.01509	Median :0.02589	Mean :0.03189	3rd Qu.:0.04205	Max. :0.39600
concave_points_se	Min. :0.000000	1st Qu.:0.007638	Median :0.010930	Mean :0.011796	3rd Qu.:0.014710	Max. :0.052790
symmetry_se	Min. :0.007882	1st Qu.:0.015160	Median :0.018730	Mean :0.020542	3rd Qu.:0.023480	Max. :0.078950
fractal_dimension_se	Min. :0.0008948	1st Qu.:0.0022480	Median :0.0031870	Mean :0.0037949	3rd Qu.:0.0045580	Max. :0.0298400
radius_worst	Min. : 7.93	1st Qu.:13.01	Median :14.97	Mean :16.27	3rd Qu.:18.79	Max. :36.04
texture_worst	Min. :12.02	1st Qu.:21.08	Median :25.41	Mean :25.68	3rd Qu.:29.72	Max. :49.54
perimeter_worst	Min. : 50.41	1st Qu.: 84.11	Median : 97.66	Mean :107.26	3rd Qu.:125.40	Max. :251.20
area_worst	Min. : 185.2	1st Qu.: 515.3	Median : 686.5	Mean : 880.6	3rd Qu.:1084.0	Max. :4254.0
smoothness_worst	Min. :0.07117	1st Qu.:0.11660	Median :0.13130	Mean :0.13237	3rd Qu.:0.14600	Max. :0.22260
compactness_worst	Min. :0.02729	1st Qu.:0.14720	Median :0.21190	Mean :0.25427	3rd Qu.:0.33910	Max. :1.05800
concavity_worst	Min. :0.0000	1st Qu.:0.1145	Median :0.2267	Mean :0.2722	3rd Qu.:0.3829	Max. :1.2520
concave_points_worst	Min. :0.00000	1st Qu.:0.06493	Median :0.09993	Mean :0.11461	3rd Qu.:0.16140	Max. :0.29100
symmetry_worst	Min. :0.1565	1st Qu.:0.2504	Median :0.2822	Mean :0.2901	3rd Qu.:0.3179	Max. :0.6638
fractal_dimension_worst	Min. :0.05504	1st Qu.:0.07146	Median :0.08004	Mean :0.08395	3rd Qu.:0.09208	Max. :0.20750

2.4 Features analysis:

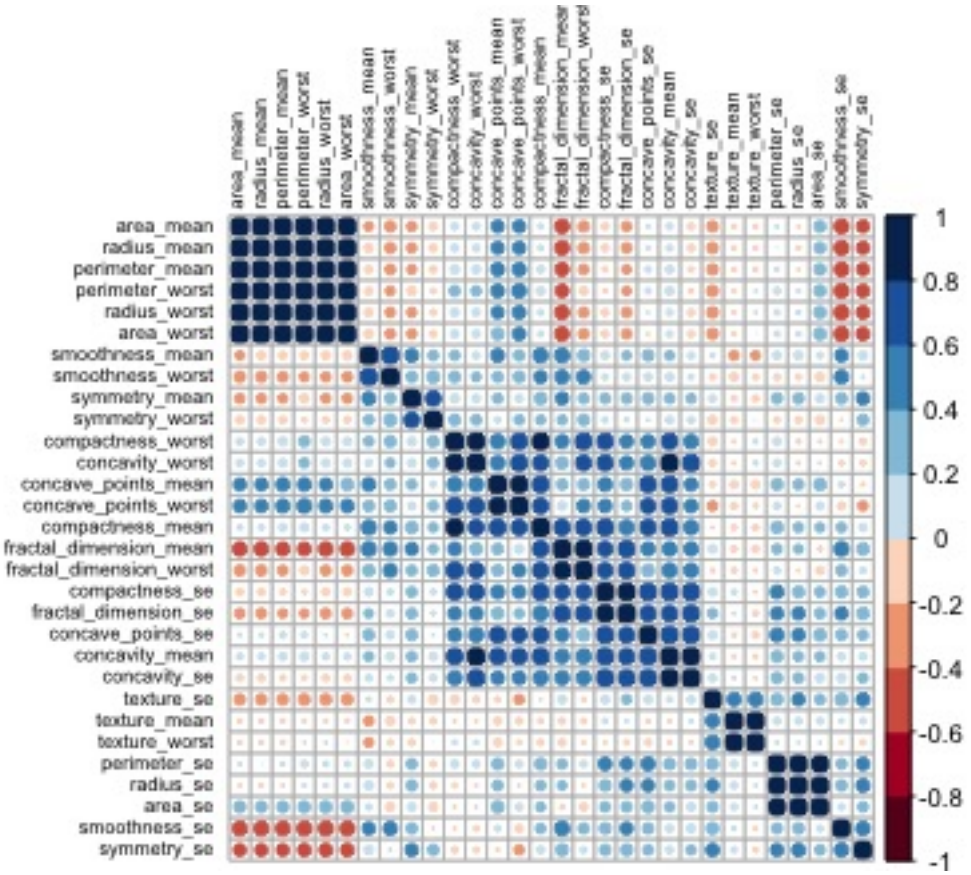
The density plot below gives a general overview on the distribution of the 30 features. We can see that the data seems normally distributed.

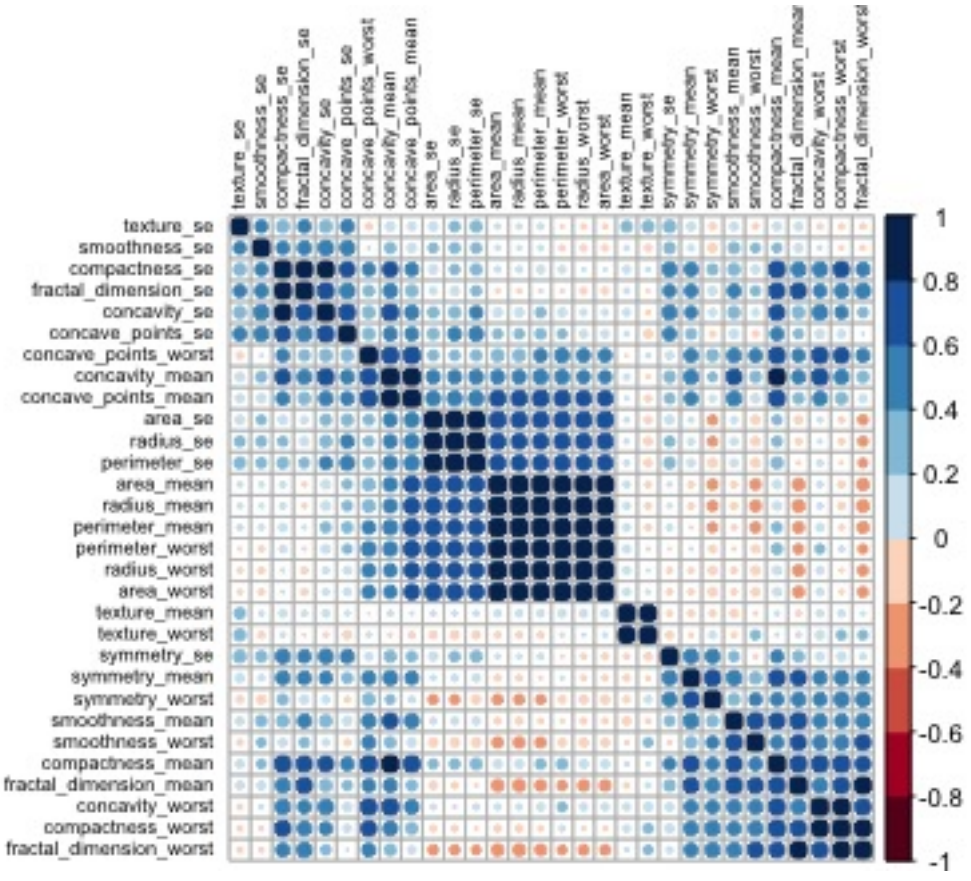


The density plot below shows the distribution of the 30 features grouped by diagnostic. We can see that some features the density curves for benign and malignant are mostly overlapping and for other the 2 density curves are quite disjoint. We can also see that the features for malignant diagnostics seem to have a grater variance. This could be taken into consideration when building the prediction models.



The the 2 plots below shows the correlations between the 30 features for begin diagnostics (top) and malignant diagnostics (bottom). We can see that many variables are significnalty correlated with each others.





We can also see that the correlations between the features differs for begin diagnostics and malignant diagnostics. This could be taken into consideration when building the prediction models.

2.5 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of large data sets. The general idea is to reduce the dimension of the dataset while preserving important characteristics making it easier to explore to visualise. PCA also reduces the computational complexity of the model which makes machine learning algorithms run faster.

Reducing the number of components or variables used to build a model comes at the expense of accuracy but the trick is to balance the trade off between accuracy and simplicity. In other words, the purpose of PCA to reduce the number of variables while preserving as much information as possible.

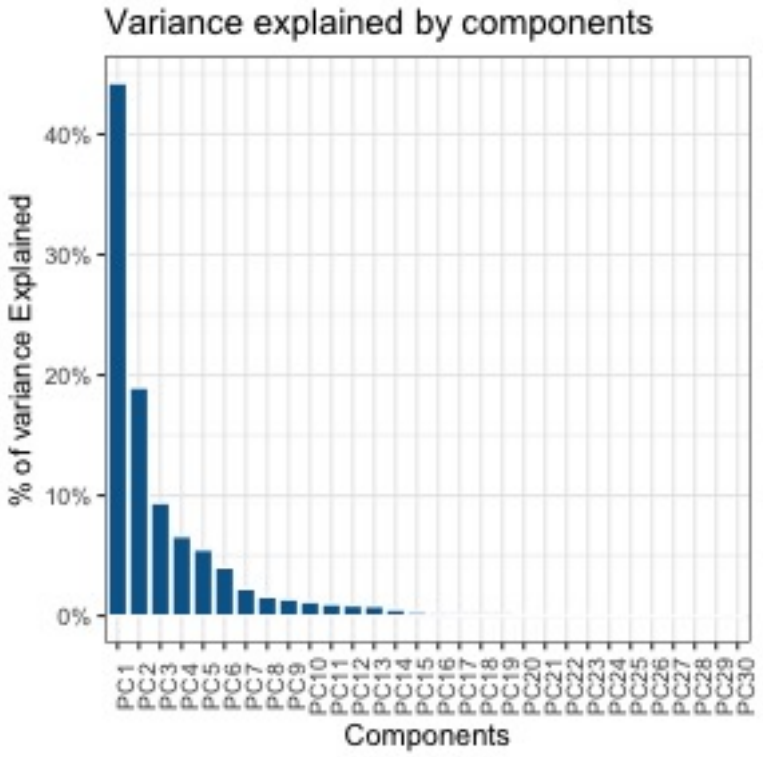
The first component is determined because it accounts for the most variance, and each subsequent component is chosen according to the next greatest portion of variance accounted for. So component one will account for the most variance, component 2 will account for the second most variance, and so forth.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	3.644394	2.385656	1.678675	1.407352	1.284029	1.098798	0.8217178	0.6903746	0.6456739	0.5921938
Proportion of Variance	0.442720	0.189710	0.093930	0.066020	0.054960	0.040250	0.0225100	0.0158900	0.0139000	0.0116900
Cumulative Proportion	0.442720	0.632430	0.726360	0.792390	0.847340	0.887590	0.9101000	0.9259800	0.9398800	0.9515700

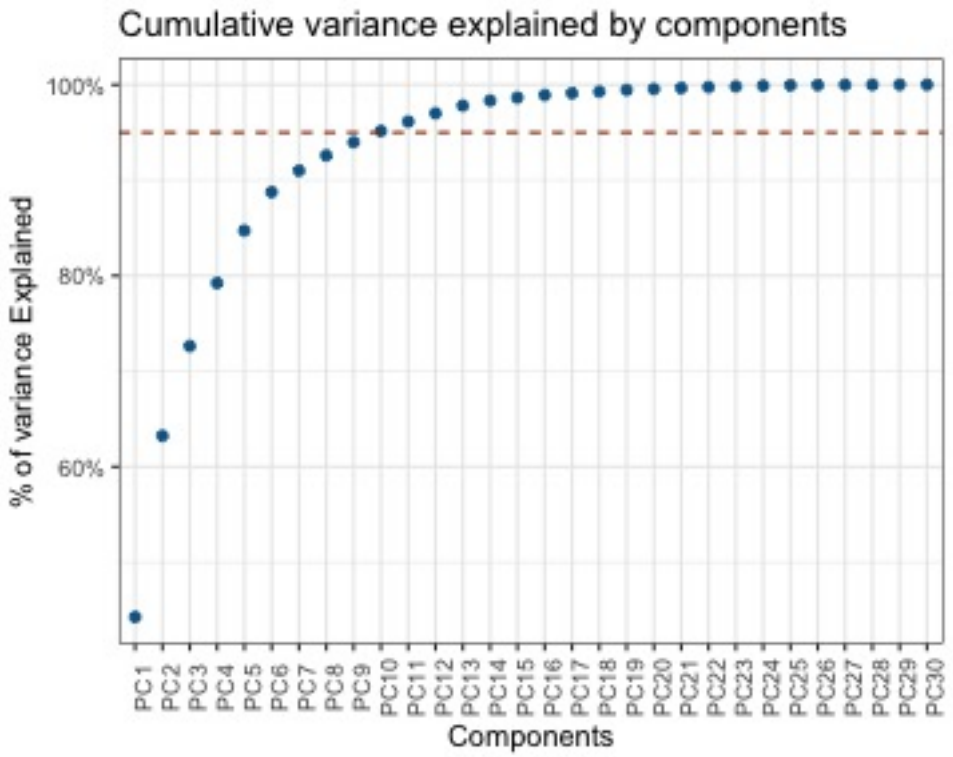
	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20
Standard deviation	0.5421399	0.5110395	0.4912815	0.3962445	0.3068142	0.2826001	0.2437192	0.2293878	0.2224356	0.1765203
Proportion of Variance	0.0098000	0.0087100	0.0080500	0.0052300	0.0031400	0.0026600	0.0019800	0.0017500	0.0016500	0.0010400
Cumulative Proportion	0.9613700	0.9700700	0.9781200	0.9833500	0.9864900	0.9891500	0.9911300	0.9928800	0.9945300	0.9955700

	PC21	PC22	PC23	PC24	PC25	PC26	PC27	PC28	PC29	PC30
Standard deviation	0.1731268	0.1656484	0.1560155	0.1343689	0.1244238	0.0904303	0.083069	0.0398665	0.0273643	0.0115345
Proportion of Variance	0.0010000	0.0009100	0.0008100	0.0006000	0.0005200	0.0002700	0.000230	0.0000500	0.0000200	0.0000000
Cumulative Proportion	0.9965700	0.9974900	0.9983000	0.9989000	0.9994200	0.9996900	0.999920	0.9999700	1.0000000	1.0000000

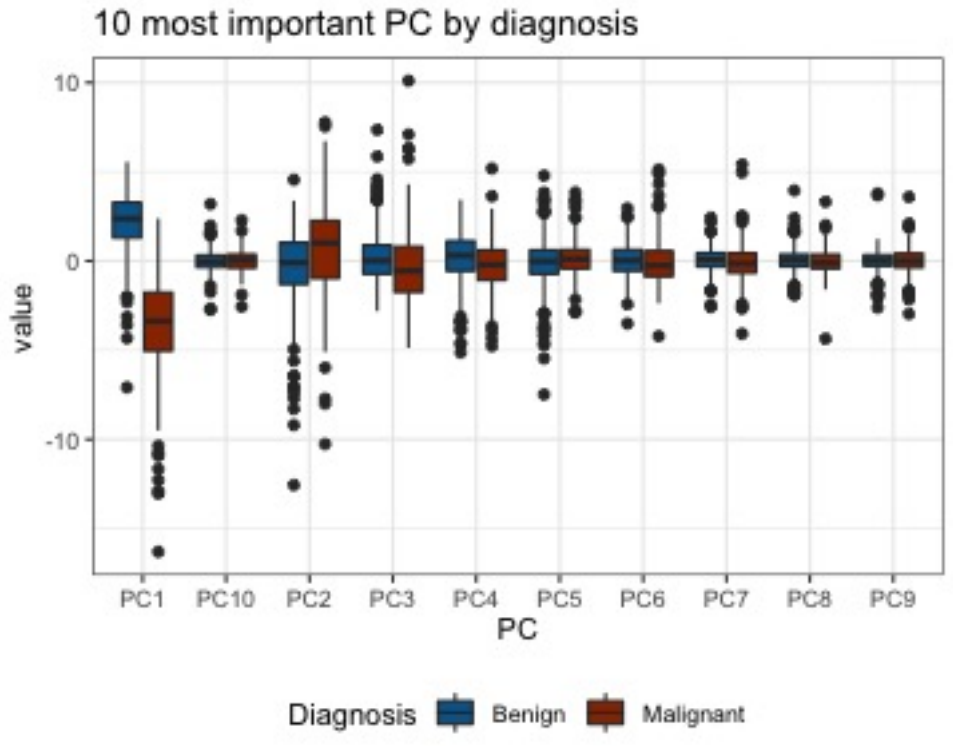
The graph bellow shows of much variability (y-axis) is explained by individual components (x-axis). We can see that values drops exponentially.



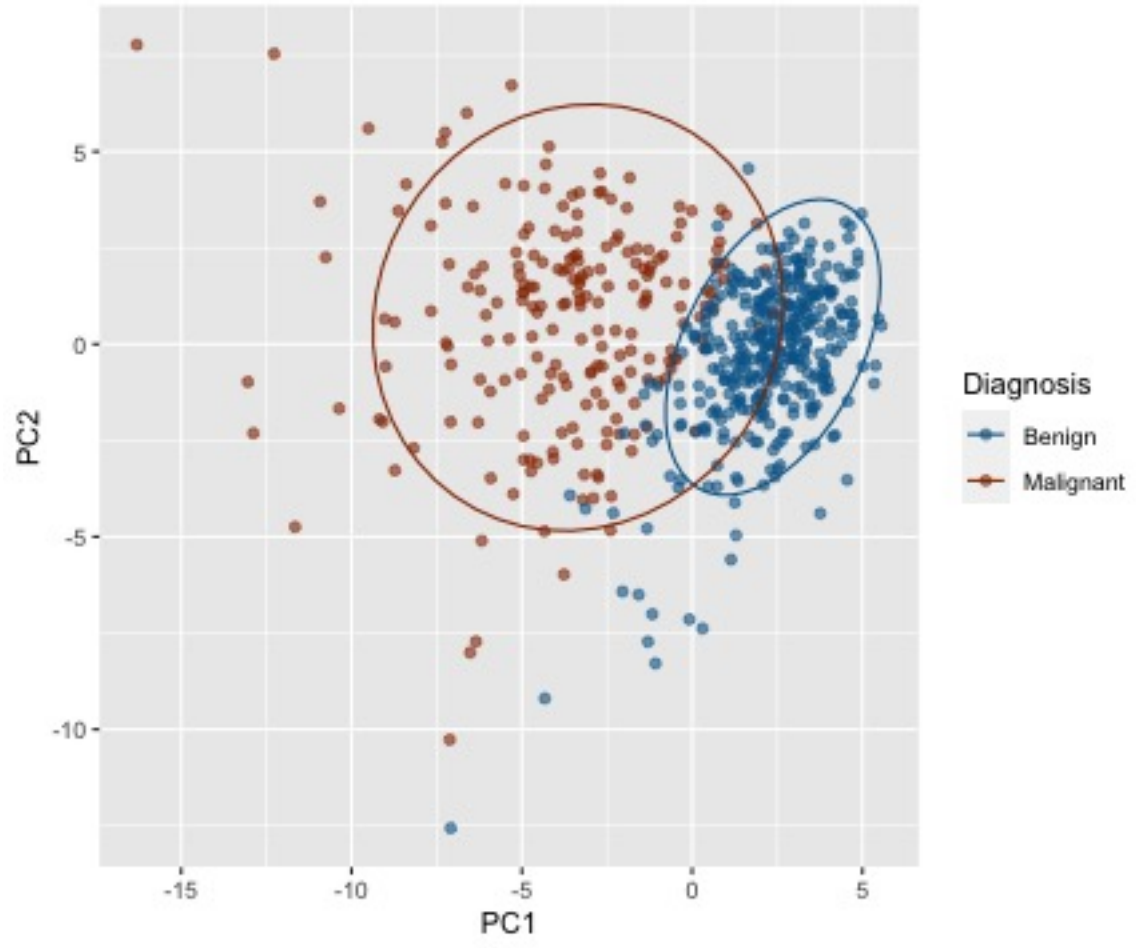
The graph below shows the cumulative variance explained for each subsequent component. We can see than 10 components explain 95% of variance (horizontal line).



The box plots below shows the first 10 principal components grouped by diagnosis. we can see that PC1 is the only component for which the interquartile ranges do not overlap which might one of the reasons why it explain 44.2% of the variance.



The graphs bellow is a two-dimensional scatter plot of the first two principal components. We can see that malignant data-points have a wider spread than the benign ones.



As a result of the principal component analysis, we have seen that there are significant differences between benign and malignant samples, suggesting it should be possible to find prediction models with high performance levels. It theory, using only the first 10 components should be enough to achieve high performance but since our data set is relatively small, using all components in not require a significant amount of computing power.

3 Model creation and validation

3.1 Model performance evaluation

Before we start creating and testing various models, we need to explain how their performance will be evaluated.

The table below, called a confusion matrix, displays the 4 different combinations of predicted vs actual values from our models.

	Diagnosis = Malignant	Diagnosis = Begnin
Prediction = Malignant	True positives (TP)	False negatives (FP)
Prediction = Begnin	False positives (FP)	True negatives (TN)

Accuracy is overall proportion that is predicted correctly:

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

sensitivity the the ability of the model to predict a positive outcome when the actual outcome is positive. In our case, correctly identify if the FNA sample with a malignant.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

specificity is the the ability of the model to predict a positive outcome when the actual outcome is positive. In our case, correctly identify if the FNA sample with a benign.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

False negative rate is the proportion of the true positive (malignant sample) for which the test result is negative (benign).

$$\text{False negative rate} = \frac{\text{FN}}{\text{FN} + \text{TP}}$$

False positive rate is the proportion of true negative (benign sample) for which the test result is positive (malignant).

$$\text{False Positive Rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

As mentioned, the main metrics used to select the best model are accuracy and sensitivity because of the potentially deadly consequences of a false negative test.

3.2 Data Preparation

3.2.1 Create train and test sets

Before going bulling models the wdbc data set was separate into 2 sibsets to ensure that data used to train the prediction models not the same as the data used to test the models and avoid biases over training. The following subsets have been created:

- **wdbc_train** : 80% of the wdbc dataset of used for the data exploration and develop the recommendation model
- **wdbc_test** : 20% of the wdbc dataset used to validate the final prediction models

3.2.2 Data Normalisation

The wdbc data was also normalised before being used to train and test the models. The goal of normalisation is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

3.3 Overview of the models used

Six different models have been evaluated:

1. logistic regression
2. loess regression
3. linear discriminant analysis (LDA)
4. quadratic discriminant analysis (QDA)
5. k-nearest neighbors (KNN)
6. random forest

3.3.1 logistic regression

Logistic regression is widely used approach ti solve classification problems (e.g. B / M). The binary logistic model is used to estimate the probability of a binary response based on one or more predictors or independent variables.

3.3.2 loess regression

Loess (locally estimated scatterplot smoothing) is very flexible regression techniques. It uses a smoothing function that attempts to capture general patterns relationships while reducing the noise and it makes minimal assumptions about the relationships among variables.

3.3.3 Linear discriminant analysis (LDA)

Linear discriminant analysis is a dimensionality reduction technique that provides the highest possible discrimination among various classes. It is used in machine learning to find the linear combination of features, which can separate two or more classes of objects with best performance

3.3.4 Quadratic discriminant analysis (QDA)

QDA is a variant of LDA in which an individual covariance matrix is estimated for every class of observations. QDA is particularly useful if there is prior knowledge that individual classes exhibit distinct covariances.

3.3.5 K nearest neighbors

The k-Nearest neighbour model (kNN) is a simple approach to supervised machine learning that assumes proximity equates to similarity it compares the closeness between observations that already exist in a data set and the ones that are newly formed. The machine does the math itself and selects the number of neighbours that need to be compared (k). It limits the occurrence of data underfitting and overfitting.

3.3.6 Random forest

Random forest is an expansion of decision tree. It works by first constructing decision trees with training data, then fitting new data within one of the trees as a "random forest" Put simply, random forest averages your data to connect it to the nearest tree on the data scale.

4 Results

The performance of the 6 models used is sumarised in the table below:

Model	Accuracy	Sensitivity	Specificity	False Negative Rate	False Positive Rate
Logistic regression	0.9565217	0.9767442	0.9444444	0.0232558	0.0555556
Loess	0.9826087	0.9767442	0.9861111	0.0232558	0.0138889
LDA	0.9913043	0.9767442	1.0000000	0.0232558	0.0000000
QDA	0.9565217	0.9302326	0.9722222	0.0697674	0.0277778
K nearest neighbors	0.9652174	0.9069767	1.0000000	0.0930233	0.0000000
Random forest	0.9913043	0.9767442	1.0000000	0.0232558	0.0000000

The model with the best performance is LDA with an accuracy of 99.1% and a sensitivity of 97.6%. The specificity of the LDA model is 100%. the loess model in the second best performing with an accuracy of 98.2% and a sensitivity of 97.6%.

Comparing the accuracy of both models with the original peer-reviewed analysis of this dataset, which achieved the accuracy of 97.5% (Wolberg et al., 1995) we can say that the performance achieved in this project seems credible.

5 Conclusion

The goal of this project was to create and test various models to predict whether a FNA sample was benign or malignant. The performance achieved satisfactory for this project but could be further improved using more advanced models, for example boosted algorithms that would increase the odds of correctly classifying samples. Using larger data sets, if available, to train the models would also improve their performance.