Multivariate Statistical Analysis

Rui Heinonen
# 482136

Tables of contents

Tables

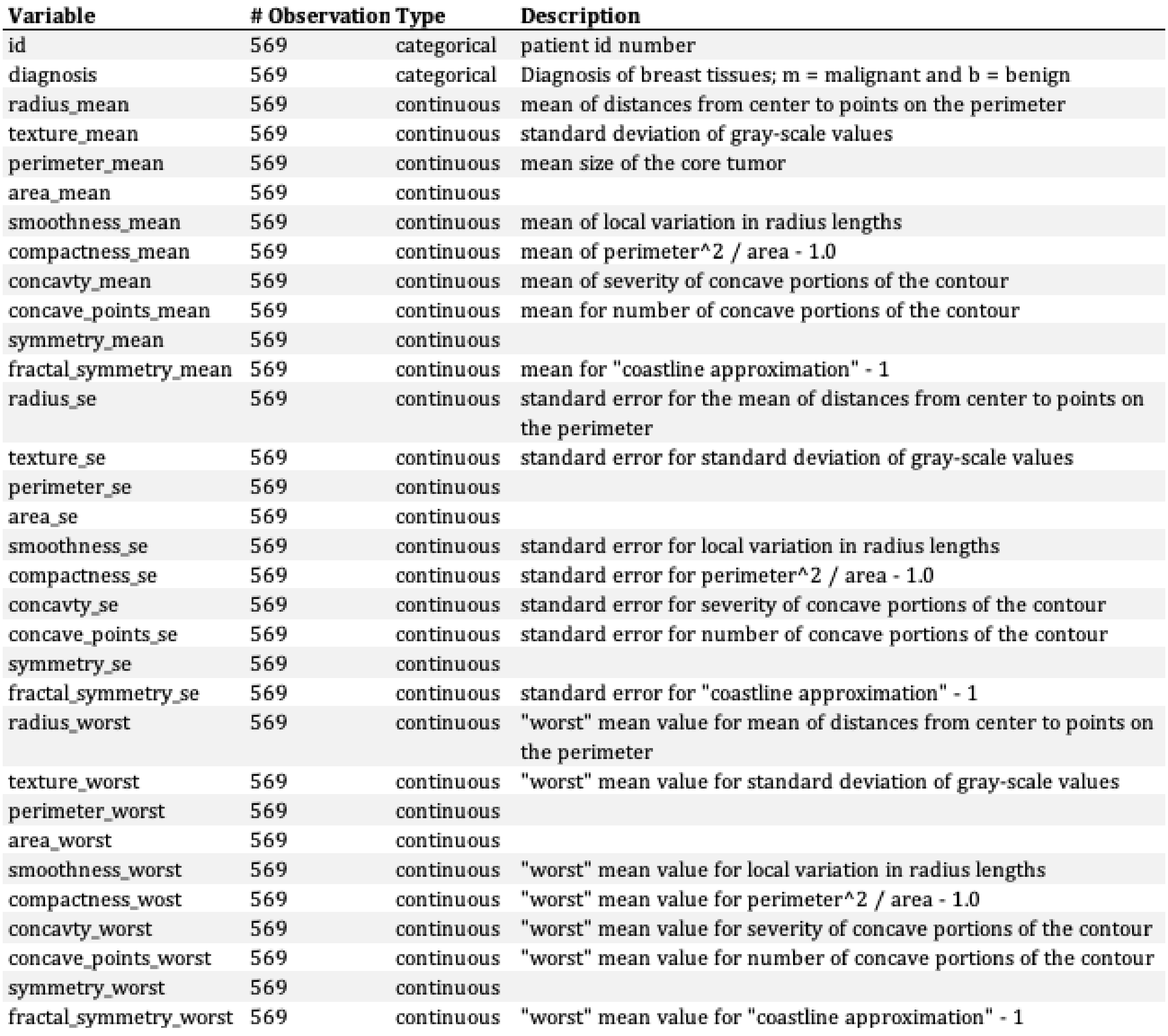
Figures

1.Data description

For this study, I will use the Breast Cancer Wisconsin (Diagnostic) Data Set which I obtained from Kaggle[1]. The features in the data set are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Out of each image, three different attributes were computed for each feature: the mean, the standard error, and the “worst” (mean of three largest values). This results in a data set that consists of 30 distinct features in addition to two categorical variables: id number and diagnosis. The id number is individual patient number and the diagnosis represents the dependent variable with two categories: M for malignant and B for Benign.

Table 1. Data

This table reports the initial data set used in the study. Altogether there are 30 continuous explanatory variables and two categorical variables. The data does not have any missing values.



Variable	# Observation	Type	Description
id	569	categorical	patient id number
diagnosis	569	categorical	Diagnosis of breast tissues; m = malignant and b = benign
radius_mean	569	continuous	mean of distances from center to points on the perimeter
texture_mean	569	continuous	standard deviation of gray-scale values
perimeter_mean	569	continuous	mean size of the core tumor
area_mean	569	continuous	
smoothness_mean	569	continuous	mean of local variation in radius lengths
compactness_mean	569	continuous	mean of perimeter^2 / area - 1.0
concavty_mean	569	continuous	mean of severity of concave portions of the contour
concave_points_mean	569	continuous	mean for number of concave portions of the contour
symmetry_mean	569	continuous	
fractal_symmetry_mean	569	continuous	mean for "coastline approximation" - 1
radius_se	569	continuous	standard error for the mean of distances from center to points on the perimeter
texture_se	569	continuous	standard error for standard deviation of gray-scale values
perimeter_se	569	continuous	
area_se	569	continuous	
smoothness_se	569	continuous	standard error for local variation in radius lengths
compactness_se	569	continuous	standard error for perimeter^2 / area - 1.0
concavty_se	569	continuous	standard error for severity of concave portions of the contour
concave_points_se	569	continuous	standard error for number of concave portions of the contour
symmetry_se	569	continuous	
fractal_symmetry_se	569	continuous	standard error for "coastline approximation" - 1
radius_worst	569	continuous	"worst" mean value for mean of distances from center to points on the perimeter
texture_worst	569	continuous	"worst" mean value for standard deviation of gray-scale values
perimeter_worst	569	continuous	
area_worst	569	continuous	
smoothness_worst	569	continuous	"worst" mean value for local variation in radius lengths
compactness_wost	569	continuous	"worst" mean value for perimeter^2 / area - 1.0
concavty_worst	569	continuous	"worst" mean value for severity of concave portions of the contour
concave_points_worst	569	continuous	"worst" mean value for number of concave portions of the contour
symmetry_worst	569	continuous	
fractal_symmetry_worst	569	continuous	"worst" mean value for "coastline approximation" - 1

[1] [Online] Available from https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. [Accessed 02/04/2019]

2.Research question

Can we predict whether a tumor is benign or malignant based on features that are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

3.Methodology

For this study, I will use Linear Discriminant Analysis (LDA) that is a well-known classification method for predicting categories. Furthermore, as all explanatory variables are continuous, LDA seems to be a natural choice here. Due to the scope of this study, this paper ignores the examination and evaluation of LDA assumptions of multivariate normality, homoscedasticity (covariance matrix) and multicollinearity.

4.Univariate analysis

Table 2. Descriptive statistics[2]

This table reports the key basic statistics of the data. From this table, it is apparent that the variances of some variables are extremely large. By comparing the means, minimums and maximums, one can also observe that there are extreme outliers in the data. Furthermore, by looking at the kurtosis and skewness metrics it is evident that many of the variables are non-normally distributed. To test the underlaying assumption of normality one could apply e.g. with Shapiro-Wilk test[3].

[2] See appendix for univariate histograms.

[3] Shapiro-Wilk requires a large sample. Validating normality in addition with e.g. Q-Q plots is recommended.

	Mean	Standard Error	Median	Standard Deviation	Sample Variance	Kurtosis	Skewness	Minimum	Maximum
radius_mean	706,77	101,88	13,85	2430,24	5906082,83	8,66	3,25	7,76	9904,00
texture_mean	19,29	0,18	18,84	4,30	18,50	0,76	0,65	9,71	39,28
perimeter_mean	91,97	1,02	86,24	24,30	590,44	0,97	0,99	43,79	188,50
area_mean	654,89	14,75	551,10	351,91	123843,55	3,65	1,65	143,50	2501,00
smoothness_mean	4,30	0,88	0,10	21,07	444,14	21,72	4,85	0,05	123,00
compactness_mean	4,84	1,12	0,09	26,83	719,71	43,81	6,35	0,02	277,00
concavity_mean	7,49	1,49	0,06	35,62	1268,71	31,83	5,46	0,00	313,00
concave points_mean	2,37	0,68	0,03	16,16	260,99	61,12	7,71	0,00	162,00
symmetry_mean	16,97	2,26	0,18	53,85	2899,39	7,68	3,02	0,12	304,00
fractal_dimension_m	0,85	0,30	0,06	7,10	50,46	80,45	9,01	0,05	78,00
radius_se	77,14	11,63	0,34	277,33	76910,67	35,32	5,26	0,11	2873,00
texture_se	825,49	34,91	1025,00	832,74	693458,42	0,31	0,65	0,36	4885,00
perimeter_se	2549,98	73,66	2155,00	1757,07	3087309,98	2,49	1,36	0,77	9807,00
area_se	316,23	64,24	25,79	1532,27	2347853,55	28,26	5,46	10,08	9833,00
smoothness_se	0,01	0,00	0,01	0,00	0,00	10,47	2,31	0,00	0,03
compactness_se	0,18	0,07	0,02	1,68	2,82	158,92	12,14	0,00	27,00
concavity_se	1,15	0,73	0,03	17,47	305,23	463,71	20,87	0,00	396,00
concave points_se	0,07	0,03	0,01	0,78	0,62	206,14	14,26	0,00	12,00
symmetry_se	0,23	0,09	0,02	2,11	4,46	123,04	10,74	0,01	31,00
fractal_dimension_se	0,01	0,01	0,00	0,25	0,06	568,87	23,85	0,00	6,00
radius_worst	315,19	69,40	15,15	1655,46	2740545,61	27,09	5,38	7,93	9981,00
texture_worst	25,68	0,26	25,41	6,15	37,78	0,22	0,50	12,02	49,54
perimeter_worst	107,26	1,41	97,66	33,60	1129,13	1,07	1,13	50,41	251,20
area_worst	880,58	23,87	686,50	569,36	324167,39	4,40	1,86	185,20	4254,00
smoothness_worst	10,63	1,56	0,13	37,24	1386,55	9,69	3,36	0,07	185,00
compactness_worst	25,26	4,04	0,23	96,47	9307,04	36,64	5,33	0,03	1058,00
concavity_worst	26,72	4,79	0,25	114,20	13042,56	45,54	6,05	0,00	1252,00
concave points_worst	8,75	1,65	0,10	39,47	1557,56	22,83	4,79	0,00	291,00
symmetry_worst	30,37	3,80	0,29	90,75	8235,21	7,24	2,89	0,16	544,00
fractal_dimension_w	1,96	0,61	0,08	14,46	209,22	68,78	8,09	0,06	173,00

To further examine the numerical explanatory variables, I will generate box plots for benign and malignant tumors in three groups: first the “mean” values (Figure 1.), secondly the “standard error” values (Figure 2.) and lastly the “worst” values (Figure 3.). Before plotting the box plots, I will center and scale the explanatory variables to address the broad difference of location and scatter within the variables, as seen from table 2.

Figure 1 Box plots “mean”

This Figure shows the box plots of all the mean values. This box plot shows that the medians and variances for benign and malignant tumors are almost equal for variable fractal_dimension_mean.

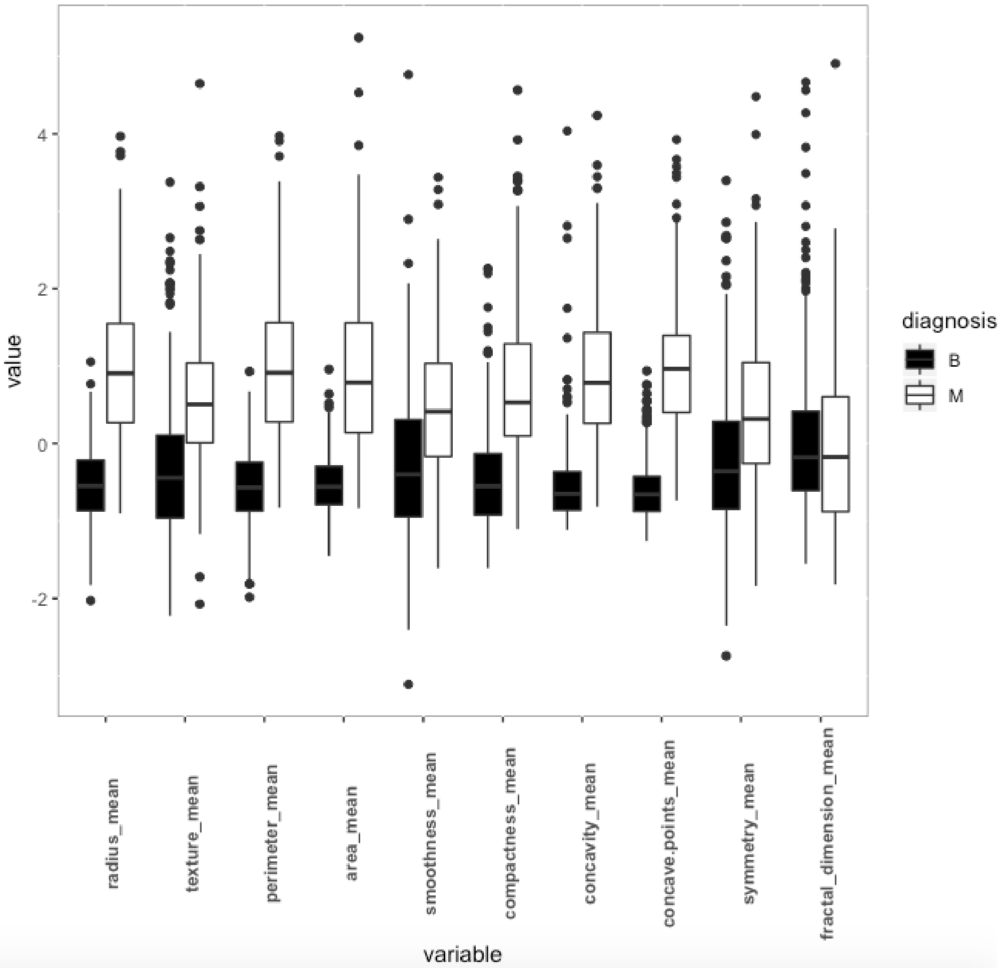


Figure 2. Box plots “standard error”

This Figure shows the box plots of all the standard error values. This box plot shows that the medians and variances for benign and malignant tumors are almost equal for variables texture se, smoothness_se, symmetry_se and fractal_dimension_se.

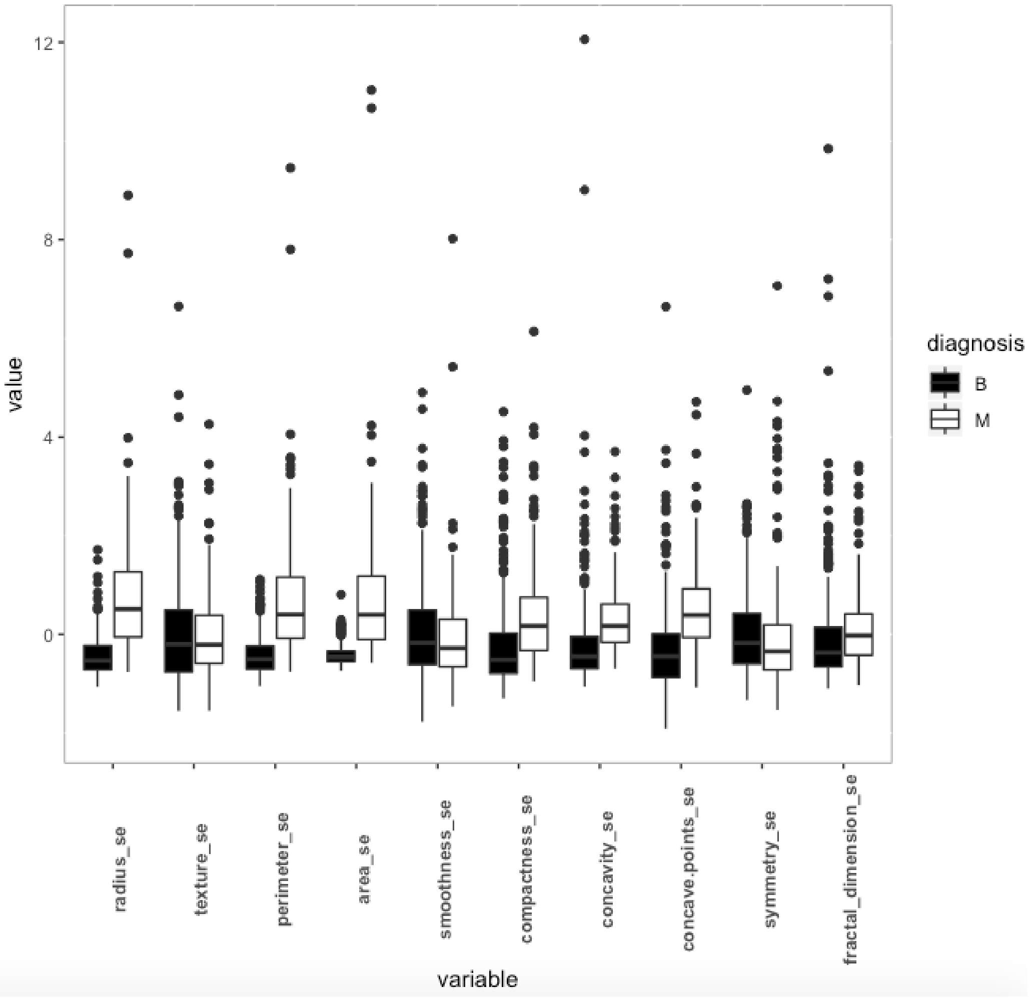
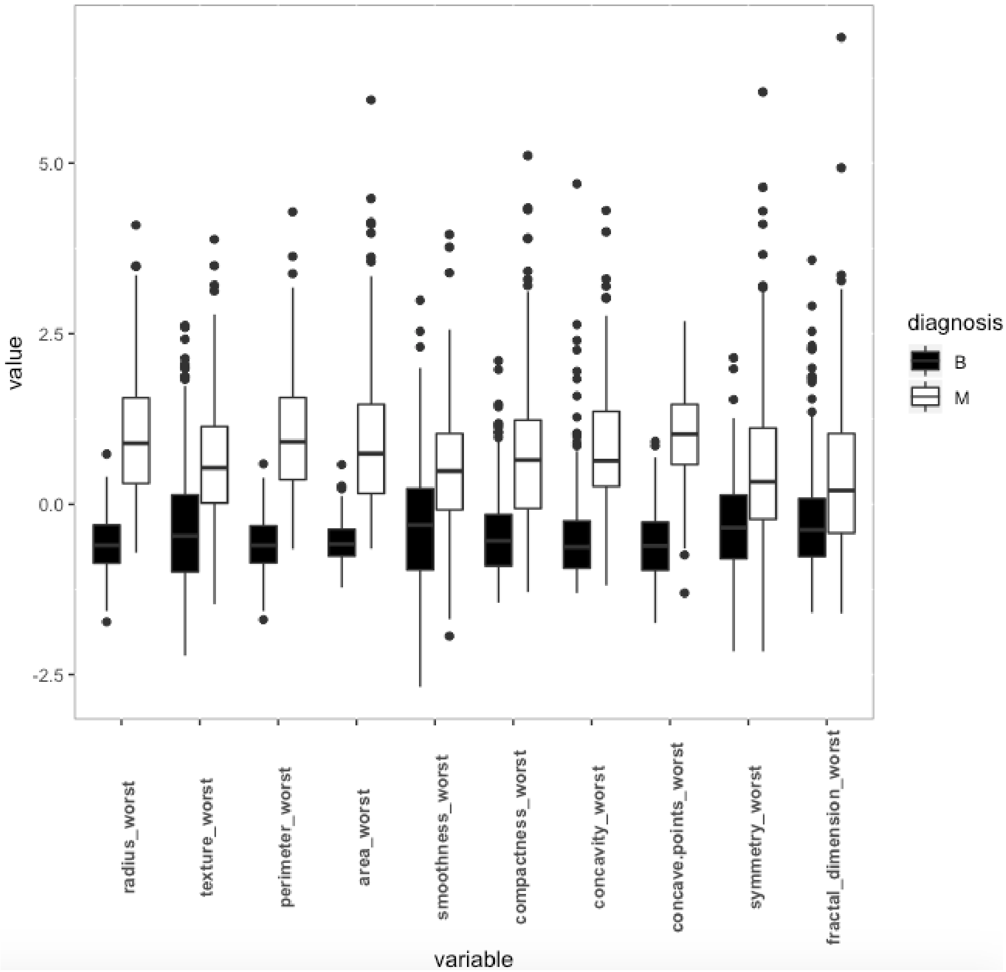


Figure 3. Box plots "worst"

This Figure shows the box plots of all the worst values. This box plot shows that the medians between Malignant and Benign tumor across these variables are quite dissimilar.



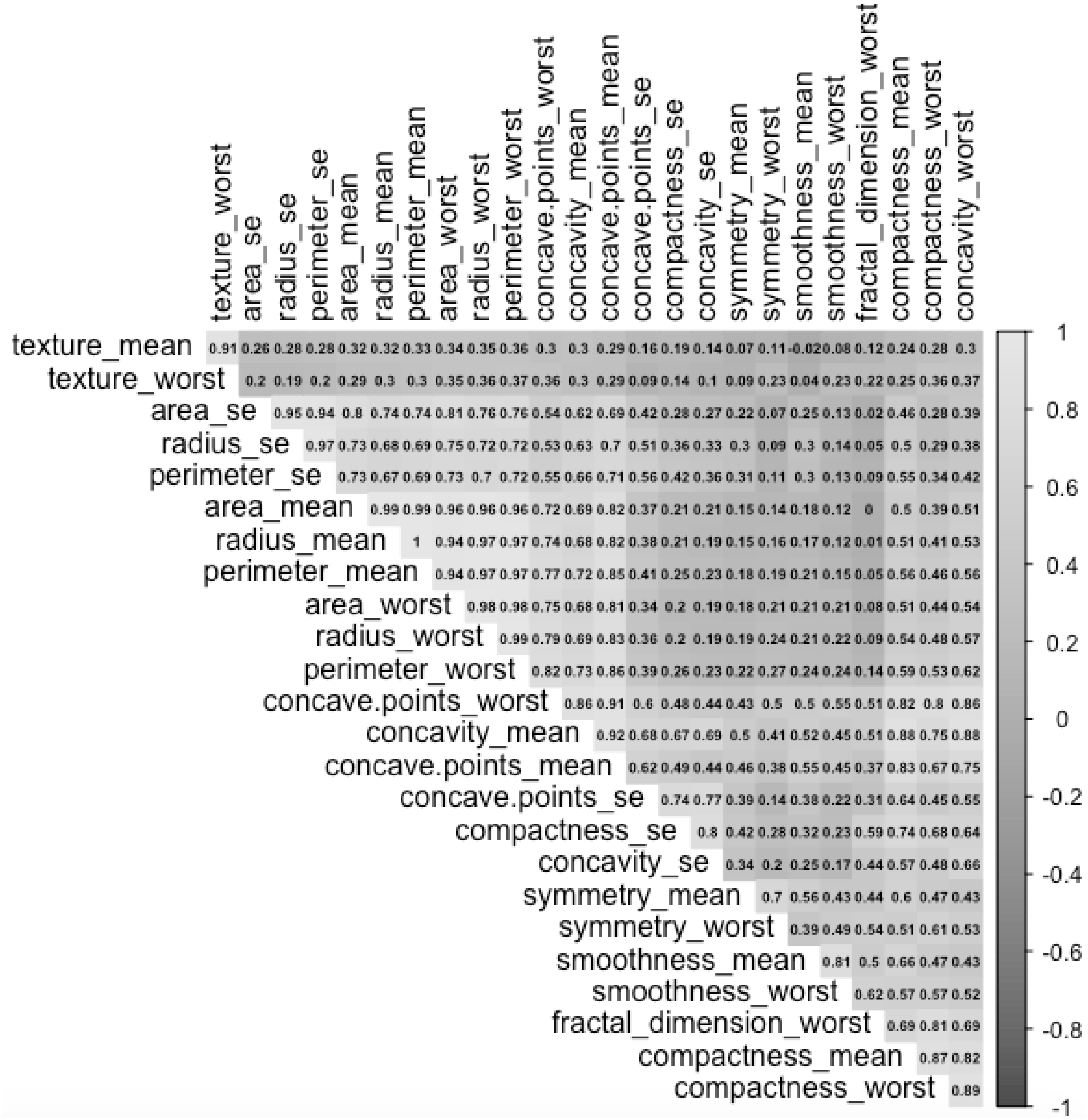
By examining the box plots (figure1, figure 2, figure 3), I am able drop insignificant features, which include fractal_dimension_mean, texture se, smoothness_se, symmetry_se and fractal_dimension_se. These variables do not give good information for classification as their medians for both benign and malignant tumors are almost equal, and hence I will exclude these variables from further examination of the data. Furthermore, as the nature of the features are somewhat similar, one could also explore the variances of the variables and select features based on those variables that exhibit most variation. However, I will not apply this approach in this study, instead, next I will perform feature selection based on examination of the pairwise correlation coefficients.

What is also worth noticing is that values for malignant tumors tend to be higher for all variables in comparison to the corresponding benign values.

5.Bivariate analysis

Figure 4. Correlation table [4]

This figure presents all pairwise Pearson's correlation coefficients of the data. Exploration of the correlation matrix shows that there are only positive correlations among the data, and that there are multiple high positive correlation coefficients.



1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

	texture_worst	area_se	radius_se	perimeter_se	area_mean	radius_mean	perimeter_mean	area_worst	radius_worst	perimeter_worst	concave.points_worst	concavity_mean	concave.points_mean	concave.points_se	compactness_se	concavity_se	symmetry_mean	symmetry_worst	smoothness_mean	smoothness_worst	fractal_dimension_worst	compactness_mean	compactness_worst	concavity_worst
texture_mean	0.91	0.26	0.28	0.28	0.32	0.32	0.33	0.34	0.35	0.36	0.3	0.3	0.29	0.16	0.19	0.14	0.07	0.11	-0.02	0.08	0.12	0.24	0.28	0.3
texture_worst		0.2	0.19	0.2	0.29	0.3	0.3	0.35	0.36	0.37	0.36	0.3	0.29	0.09	0.14	0.1	0.09	0.23	0.04	0.23	0.22	0.25	0.36	0.37
area_se			0.95	0.94	0.8	0.74	0.74	0.81	0.76	0.76	0.54	0.62	0.69	0.42	0.28	0.27	0.22	0.07	0.25	0.13	0.02	0.46	0.28	0.39
radius_se				0.97	0.73	0.68	0.69	0.75	0.72	0.72	0.53	0.63	0.7	0.51	0.36	0.33	0.3	0.09	0.3	0.14	0.05	0.5	0.29	0.38
perimeter_se					0.73	0.67	0.69	0.73	0.7	0.72	0.55	0.66	0.71	0.56	0.42	0.36	0.31	0.11	0.3	0.13	0.09	0.55	0.34	0.42
area_mean						0.99	0.99	0.96	0.96	0.96	0.72	0.69	0.82	0.37	0.21	0.21	0.15	0.14	0.18	0.12	0	0.5	0.39	0.51
radius_mean							1	0.94	0.97	0.97	0.74	0.68	0.82	0.38	0.21	0.19	0.15	0.16	0.17	0.12	0.01	0.51	0.41	0.53
perimeter_mean								0.94	0.97	0.97	0.77	0.72	0.85	0.41	0.25	0.23	0.18	0.19	0.21	0.15	0.05	0.56	0.46	0.56
area_worst									0.98	0.98	0.75	0.68	0.81	0.34	0.2	0.19	0.18	0.21	0.21	0.21	0.08	0.51	0.44	0.54
radius_worst										0.99	0.79	0.69	0.83	0.36	0.2	0.19	0.19	0.24	0.21	0.22	0.09	0.54	0.48	0.57
perimeter_worst											0.82	0.73	0.86	0.39	0.26	0.23	0.22	0.27	0.24	0.24	0.14	0.59	0.53	0.62
concave.points_worst												0.86	0.91	0.6	0.48	0.44	0.43	0.5	0.5	0.55	0.51	0.82	0.8	0.86
concavity_mean													0.92	0.68	0.67	0.69	0.5	0.41	0.52	0.45	0.51	0.88	0.75	0.88
concave.points_mean														0.62	0.49	0.44	0.46	0.38	0.55	0.45	0.37	0.83	0.67	0.75
concave.points_se															0.74	0.77	0.39	0.14	0.38	0.22	0.31	0.64	0.45	0.55
compactness_se																0.8	0.42	0.28	0.32	0.23	0.59	0.74	0.68	0.64
concavity_se																	0.34	0.2	0.25	0.17	0.44	0.57	0.48	0.66
symmetry_mean																		0.7	0.56	0.43	0.44	0.6	0.47	0.43
symmetry_worst																			0.39	0.49	0.54	0.51	0.61	0.53
smoothness_mean																				0.81	0.5	0.66	0.47	0.43
smoothness_worst																					0.62	0.57	0.57	0.52
fractal_dimension_worst																						0.69	0.81	0.69
compactness_mean																							0.87	0.82
compactness_worst																								0.89

[4] Due to the limitations set by the large data set and for the purposes of this study further bivariate analysis e.g. scatter and mosaic plots are left out of this paper (in terms of bivariate analysis I am only interested in correlations).

In order to select the most important features for the model, I follow Ferrar and Glauber (1967) and constrain pairwise correlations between any two explanatory variables to be less than 0,9. Hence, will exclude texture_worst, radius_se, perimeter_se, radius_mean, perimeter_mean, area_worst, radius_worst, perimeter_worst and concave.points_mean from the data on this point forward. To test the underlaying assumption of multicollinearity, one could generate variance inflation factors to examine the factor by which the variance is inflated. [5]

6.Multivariate analysis

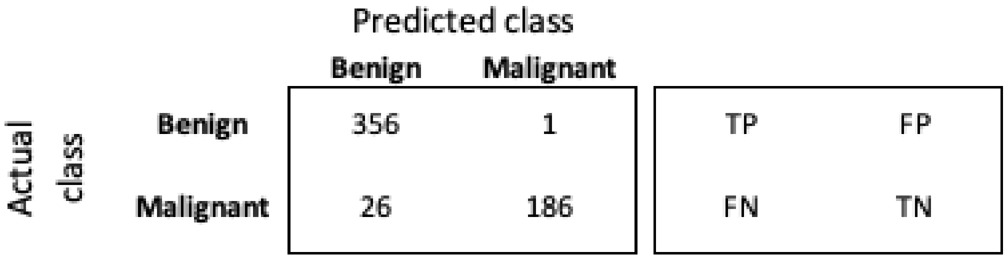
Results and evaluation of the model

For this study, I applied LDA on a Breast Cancer Wisconsin (Diagnostic) Data Set. After feature selection, the whole data matrix shrunk from 32 columns to 18 columns and no rows were deleted. For the formulation of the LDA model, I divide the data into 75% training set and 25% test set. The training set is used to construct the rules of the model and the test data will serve as a proxy for new data. In addition, the dependent variable is quite balanced with 357 benign and 212 malignant tumors. Hence oversampling or undersampling methods were not required.

Before jumping any further, I should point out that this model did not properly validate the assumptions of LDA model. While the independency of variables was examined, multivariate normality, multicollinearity and homoscedasticity (covariance matrix) were not properly addressed. Violation of these assumptions can skew the results of the model and reduce its overall performance.

Table 3. Confusion matrix

This table shows the misclassification rates of benign and malignant tumors. According to this table, the LDA model misclassifies 26 malignant tumors as benign and 1 benign tumor as malignant.



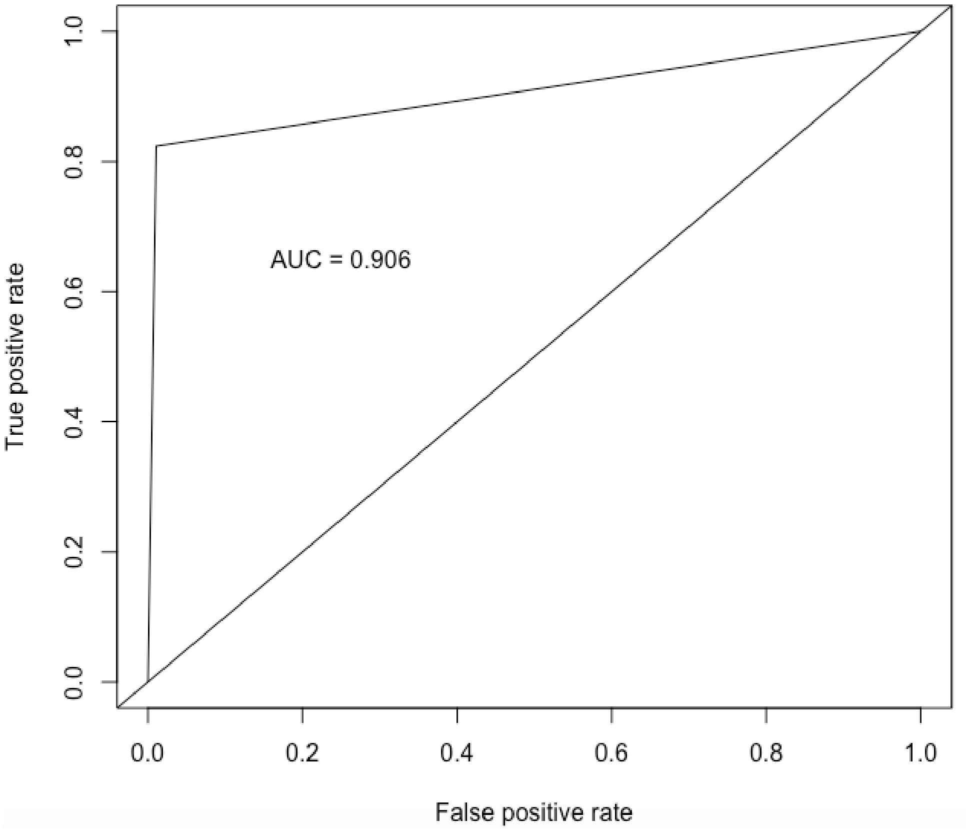
		Predicted class			
		Benign	Malignant		
Actual class	Benign	356	1	TP	FP
	Malignant	26	186	FN	TN

The results indicate that this model is quite accurate in predicting whether a tumor is malignant or benign. However, in 26 out of 27 cases, the model classifies a malignant tumor as a benign one, and on the contrary, in one (1) out of 27 cases the model classifies a benign tumor as a malignant one. In this particular context, false negatives can be extremely costly because classifying a malignant tumor as a benign one can be crucial for the patient. Hence, in this case false positives would be more preferred than false negatives.

Figure 5. AUROC curve

[5] A VIF score equal or larger than 10 would imply that multicollinearity is likely present, which would require further feature selection by dropping the most inflated variables.

This figure reports the true positive rate against the false positive rate and the AUC score (0,906). According to the AUC score, the suggested LDA model will be able to distinguish positive and negative classes with a 90.6% probability. The linear line represents an indicative function that does not differ from a simple 50-50 guess.



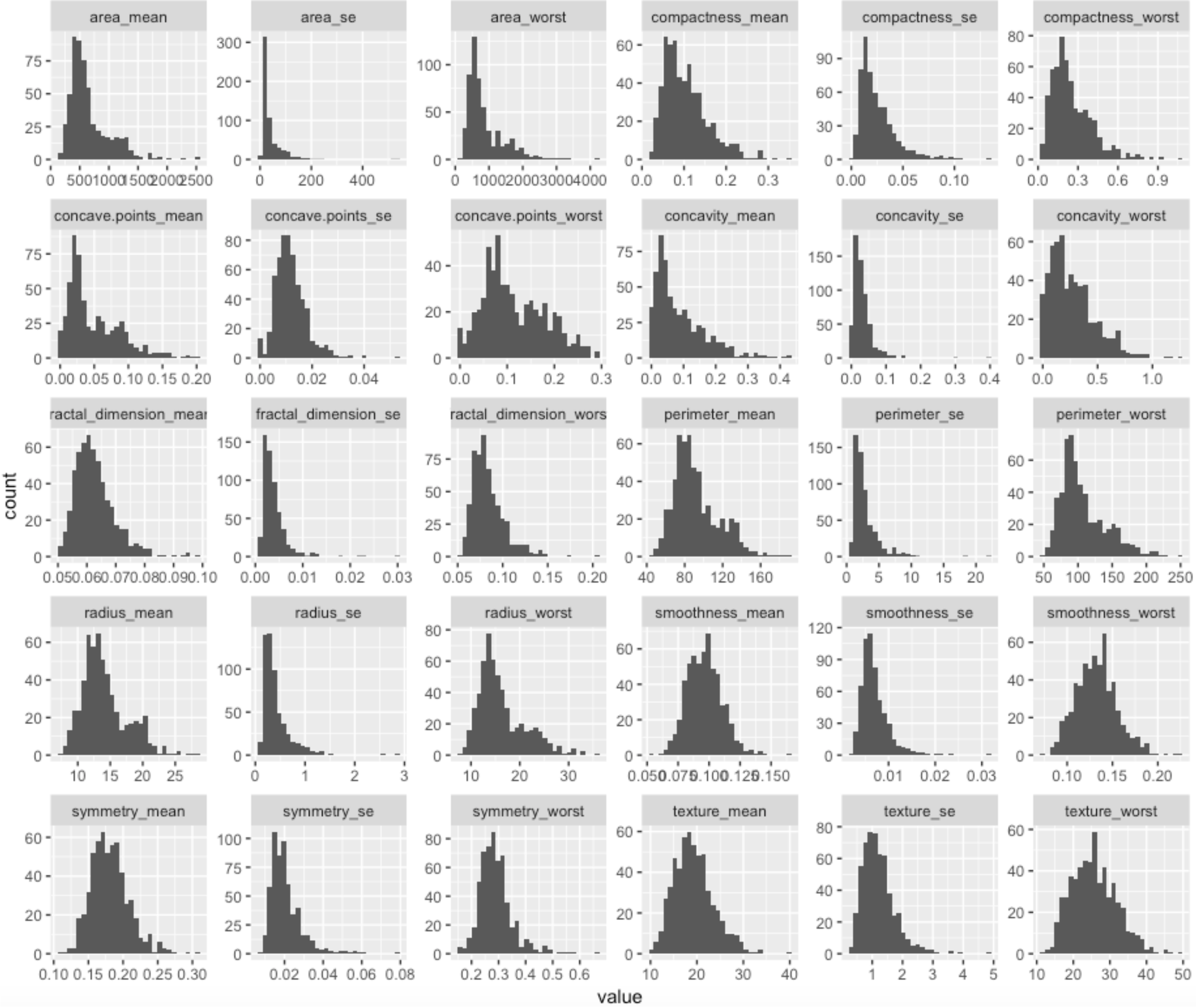
Discussion and Conclusion

In this research, I have examined benign and malignant tumors and the predictive power of 17 different explanatory variables associated to breast cancer. The aim of this study was to find out how well LDA performs in classifying benign and malignant tumors based on features that were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. As it turns out, in general the LDA performed relatively well.

Regardless of the great overall performance of the model, there is still room for improvement in terms of the false negative rate (see table3). As the false negatives can be extremely costly, one could try improving the model in order to find a better fitting model in the sense of that its erroneous predictions origin from false positives rather than false negatives. Classifying more benign tumors as malignant can exhaust more medical resources and cause personal disruption, but the cost of misclassifying a malignant tumor as benign overweighs those costs.

For future implications, the improvement for this classification problem can be achieved by testing alternative classification models, such as logistic regression, support vector machine or gradient boosting machines to name but a few. One could also consider validating different feature selection approaches, such as deriving the features from PCA before the construction of the model or by selecting features based on variation, as outlined earlier.
Appendix 1. Histograms of all independent variables

These histograms demonstrate that in general the standard error values have high kurtosis values, and that there exists positive skewness within the variables. However, some of the “xx_mean” and “xx_worst” histograms look quite nice in terms of normality. For further examination of normality, one could use for example Shapiro-Wilks test or Kolmogorov-Smirnov test and Q-Q plots.



Resources

Farrar, D.E. and Glauber, R.R., 1967. Multicollinearity in regression analysis: the problem revisited. The Review of Economic and Statistics, pp.92-107.

Ilmonen, P., and Kantala, (X). (2018) *Discriminant analysis and classification* [Lecture] MS-E2112: Multivariate Statistical Analysis. Aalto University, 11th March