**AMCS 243 - Applied Statistics and Data Analysis, Hernando Catequista Ombao - Fall Semester 2017**
**King Abdullah University of Science and Technology (KAUST) - CEMSE Divison**

# Analysis of Diabetic Retinopathy Data via Logistic Regression

Henrique Aparecido Laureano[1], Azza Al-Thagafi[2]

**Abstract**
Diabetic Retinopathy (DR) is the most common diabetic eye disease and is the leading cause of new blindness among the diabetes patients. The exact technique by which diabetes causes this condition is unclear, and it can develop without any serious symptoms. Therefore, the early detection of this disease is crucial. This paper focuses on the analysis of the retina in the diabetes patients via a logistic linear regression model. Moreover, it aims to test, quantity and interpret the variables significance at the differentiation of the patient status (diabetic retinopathy signs disease or not). Test the accuracy of the prediction by using this methodology with different link functions was also a goal of this paper. The data was taken from UCI repository [1], it contains features extracted from the Messidor (Methods to evaluate segmentation and indexing techniques in the field of retinal ophthalmology) image set to predict whether an image contains signs of diabetic retinopathy or not. In a exploratory analysis we saw that practically all the data (99.7%) present a sufficient quality assessment and that more than 90% of the patients present a Severe Retinal Abnormality (SRA). Looking marginally to the means among the groups (patients with and without signs of DR) of the features (1) Euclidian distance of the center of the macula to the center of the optic disc and (2) the diameter of the optic disc, we saw through a $t$-test that their means don't differ significantly, being in reality very closer. With a $\chi^2$-test we saw no relation between this desease status with an AM/FM-based classification. Fitting a logistic regression with all the features the same result was obtained. The logistic link function presented the better results when compared with others. The goodness of fit was satisfactory, having almost all features related with Microaneurism Detection (MD) and Exudates detection as significant. A predicitive model was also trained and good results was obtained, as a AUC of 0.798, a sensitivity of 0.805 and a specificity of 0.686.

**Keywords**
Diabetic Retinopathy; $t$-test; $\chi^2$-test; Logistic Regression; Link functions; Akaike information criterion.

[1] *Ms/PhD Student in Statistics*
[2] *Ms Student in Computer Science*
**Email**: {henrique.laureano, azza.althagafi}@kaust.edu.sa

## 1. Introduction

### 1.1 Background

Diabetes is a disease in which the ability of the body to produce and respond to the hormone insulin is impaired. A number of medical risks are associated with diabetes and many of them stem from damage to the tiny blood vessels in the eyes, called Diabetic Retinopathy (DR) [2]. DR is a condition that happens when the high blood sugar levels cause damage to the blood vessels in the retina that lines the back of the eye [3]. These blood vessels can swell or close and stopping the blood from passing through. And in the most advanced stage, the new abnormal blood vessels

grow on the retina, which can lead to a potential of severe vision loss and blindness to the people with diabetes [4]. The aforementioned features of this condition show up in fundoscopy images in Figure 1.

The number of patients with diabetic retinopathy nowadays increased very rapidly [4], and the complications associated with the long duration of the disease becomes one of the challenges that faced the health care system. During the development of DR, the patients may not notice any changes in their vision, and the DR might be very advanced by the time that patients have visual complaints and experience visual loss eventually [5]. So, to detect DR in an early stage, people with diabetes should get a dilated eye exam at least once a year, thus in case of an early diagnosis, the progression of DR can be reduced by an appropriate therapy. That's mean the early detection, timely treatment, and appropriate follow-up care of diabetic eye disease can protect the people with diabetic against vision loss.

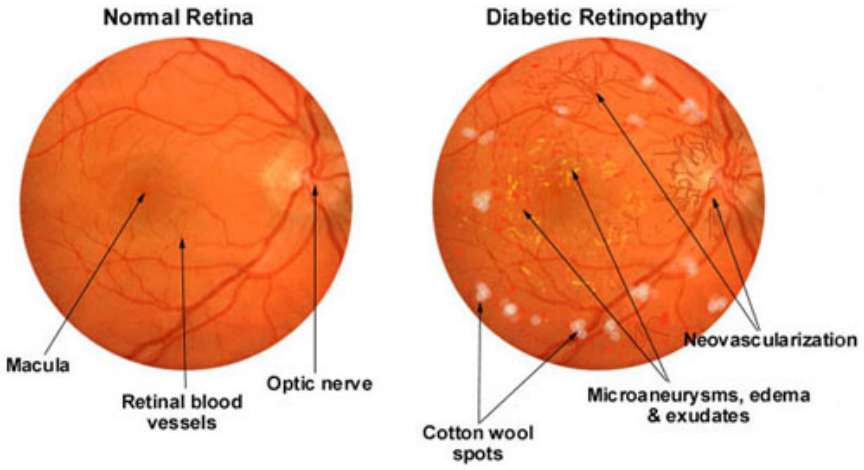


**Figure 1.** Retinal Fundus image.

Automatic Computer-Aided diagnosis system of retinal images is an important field that assist doctors in the interpretation of medical images and to easily check the state of the patient eyes. This type of system uses a wide ranges of data analysis and machine learning techniques to automatically diagnose the vessels, optic disk, and bright lesions, as well as to assess the image quality of the eyes [6].

This paper aims to use statistical techniques to understand which features are related with the response variable (presence or not of signs of DR) and try to predict these responses.

## 1.2 Dataset Description

The Diabetic Retinopathy Dataset was taken from the UCI repository website [1].

### 1.2.1 Dataset Information

This dataset contains features extracted from the Messidor image set and aims to predict whether a particular image contains signs of diabetic retinopathy or not. All the variables represent either a detected lesion, a characteristic feature of an anatomical part or an image-level descriptor.

### 1.2.2 Dataset Characteristics

The dataset characteristics are shown in Table 1.

**Table 1.** Dataset characteristics.

| | | | |
|---|---|---|---|
| Number of instances: | 1151 | Number of attributes: | 20 |
| Attributes characteristics: | Integer, Real | Area: | Life |
| Data denoted: | 03-11-2014 | Associated tasks: | Classification |
| Missing values: | No | Number of Web Hits: | 29802 |

#### 1.2.3 Attribute Information

The attributes data view of each records are shown in Table 2.

**Table 2.** Description of Diabetic Retinopathy Dataset.

| Feature | Description |
|---|---|
| Quality assessment | Binary result (0 = Bad quality, 1 = Sufficient quality) |
| Pre-screening | Binary result (0 = Lack of SRA, 1 = Severe Retinal Abnormality (SRA)) |
| MD (six features, 0.5 to 1) | Numeric. The results of Microaneurism Detection (MD). Each feature value stand for the number of microaneurisms found at the confidence level $\alpha$ = 0.5, 0.6, 0.7, 0.8, 0.9 and 1 |
| Exudates detection 1 to 8 | Numeric. Number of points in the results of exudates detection in different set of points. The values are normalized by dividing the number of lesions with the diameter of the ROI (Region of Interest) to compensate different image sizes |
| Euclidian distance | Numeric. The euclidean distance of the center of the macula to the center of the optic disc to provide important information regarding the patients condition. The values are normalized with the diameter of the ROI |
| Diameter | Numeric. Diameter of the optic disc |
| AM/FM-based classification | Binary result of the multiscale AM/FM (Amplitude-Modulation/Frequency-Modulation) - based classification (0 = Normal retinal structures, 1 = pathological lesions) |

### 1.3 Scientific Goals and Primary Questions of Interest

The number of patient with DR increased rapidly, and the exact technique by which diabetes causes this disease remains unclear. In addition to that, DR can develop without any severe symptoms. Therefore, there is a high need to improve the methods that can diagnose DR as soon as possible because the early detection and treatment can reduce the risk of blindness by 95%[4]. The project provide a selection and a study of the variables which have a significant impact on the diabetic retinopathy. Knowing these relationships better the patient can receive an early treatment that can limit the potential for significant vision loss.

The principal goal of this study was to check which variables have a difference statically significant between the two levels of the response variable, i.e., between patients without signs of DR, and with signs of DR. Besides verify which variables, we aim to quantify and interpret this difference. Another goal of this study was to check which variables are statistically significant to predict if the patient has or hasn't signs of DR.

## 2. Statistical Methods

### 2.1 Preliminary Data Exploration

From the 1151 patients in the study, 611 (53%) present signs of DR. The three categorical features presented in the dataset are shown in the Figure 2. Practically all the patients (99.7%) have a sufficient quality assessment and more than 90% present a Severe Retinal Abnormality (SRA). Given this disproportionality, this two features will not be used in the statistical analysis. Also in Figure 2 we see that 1/3 of the patients present a positive result AM/FM classification, i.e., 33.6% of the patients present pathological lesions in the retinal structures. A summary with the mean, median and standard deviation for all the numerical variables are presented in Table 3.

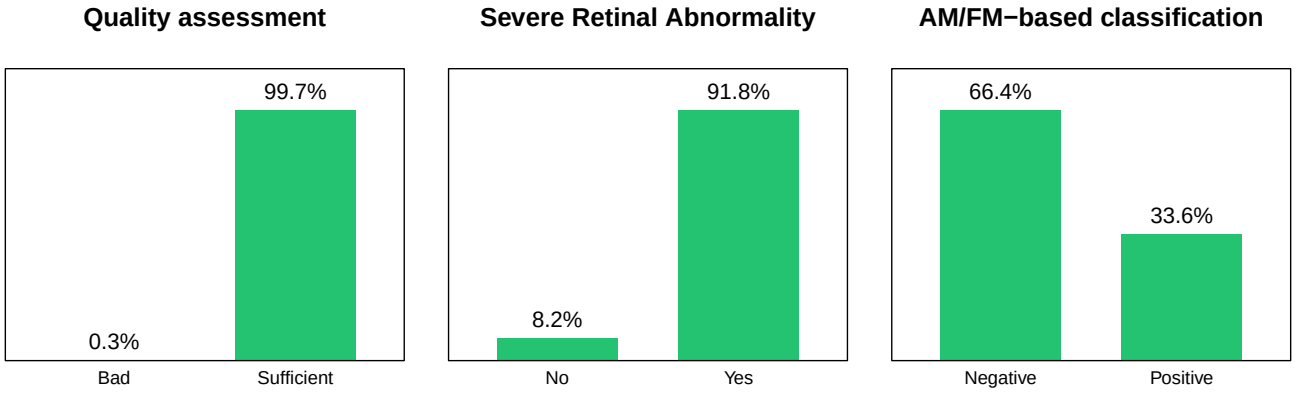


**Figure 2.** Barcharts for the three categorical features: status of quality assessment (left), presence of severe retinal abnormality (center) and result of an AM/FM-based classification (right).

**Table 3.** Summary of the numerical variables in the Diabetic Retinopathy Dataset. Mean, median and standard deviation are presented divided by the presense, or not, of signs of DR. Between the MD and exudates features, the biggest and smallest values are in bold, for easy identification.

| Feature | Mean | | Median | | Standard deviation | |
|---|---|---|---|---|---|---|
| | No sign of DR | Sign of DR | No sign of DR | Sign of DR | No sign of DR | Sign of DR |
| MD: 0.5 | **30.457** | **45.473** | **25.000** | **44.000** | **20.743** | **27.411** |
| MD: 0.6 | 30.083 | 42.943 | **25.000** | 42.000 | 20.473 | 25.444 |
| MD: 0.7 | 29.450 | 40.170 | 24.000 | 39.000 | 20.183 | 23.802 |
| MD: 0.8 | 27.863 | 36.216 | 22.000 | 34.000 | 19.321 | 21.860 |
| MD: 0.9 | 25.394 | 31.710 | 20.000 | 29.000 | 18.317 | 20.058 |
| MD: 1 | **19.098** | **22.966** | **15.000** | **20.000** | **14.257** | **15.598** |
| EXU 1 | **60.489** | **67.285** | **47.577** | **40.526** | **50.765** | **64.418** |
| EXU 2 | 23.077 | 23.098 | 18.988 | 15.297 | 19.719 | 23.156 |
| EXU 3 | 8.234 | 9.121 | 4.576 | 4.368 | 10.565 | 12.380 |
| EXU 4 | 1.402 | 2.221 | 0.435 | 0.575 | 2.794 | 4.670 |
| EXU 5 | 0.185 | 0.893 | 0.011 | 0.051 | 0.555 | 3.335 |
| EXU 6 | 0.042 | 0.363 | **0.000** | 0.004 | 0.156 | 1.427 |
| EXU 7 | 0.007 | 0.155 | **0.000** | **0.000** | 0.035 | 0.537 |
| EXU 8 | **0.003** | **0.067** | **0.000** | **0.000** | **0.016** | **0.241** |
| EUC | 0.523 | 0.523 | 0.523 | 0.523 | 0.029 | 0.028 |
| DIAM | 0.109 | 0.108 | 0.107 | 0.106 | 0.018 | 0.018 |

In the MD measures the biggest mean, median and standard deviations are observed in the patients without signs of DR, for all the confidence levels. We see that conform the confidence level became bigger all the three statistics became smaller with a considerable drop. For the patients without signs of DR, comparing the measures of the 0.5 level with the 1 level the mean decay 37%, the median decay 40% and the standard deviation decay 31%. Already for this patients with signs of DR, the mean decay 49%, the median decay 55% and the standard deviation decay 43%. Similar behavior is seen with the exudates detection measures. Comparing the measures of 1 point with 8 points, the mean falls to practically 0. Only by doubling the number of points, 1 to 2, the mean and the median decay more than 50%, independent if the patient present or not signs of DR. Patients with no signs of DR present bigger mean and standard deviations than the patients with signs of DR, for all number of set of points. For bigger set of points, the opposity behavior is seen with the medians. Also in Table 3, we see that for the euclidean distance of the center of the macula to the center of the optic disc and for the diameter of the optic disc extremally similar values are absorved in the three statistics, for both groups (patients with and without signs of DR).

A 2x2 scatter plots and the correlations for all the numerical variables is provided in Figure 4. For the MD features we see a clear linear relationship. The linear relationship looks more stronger for the patients in blue, without signs of DR. In the left-graph of the Figure 3 we can see better the correlations between the confidence levels of the MD features. The detections are correlated in all confidence levels, with a minimum correlation of 0.86 (between the most far levels). Closer confidence levels are extremely correlated (superior a 0.95). Thus, we see here a pattern. The further away the confidence levels, the lower is the correlation.

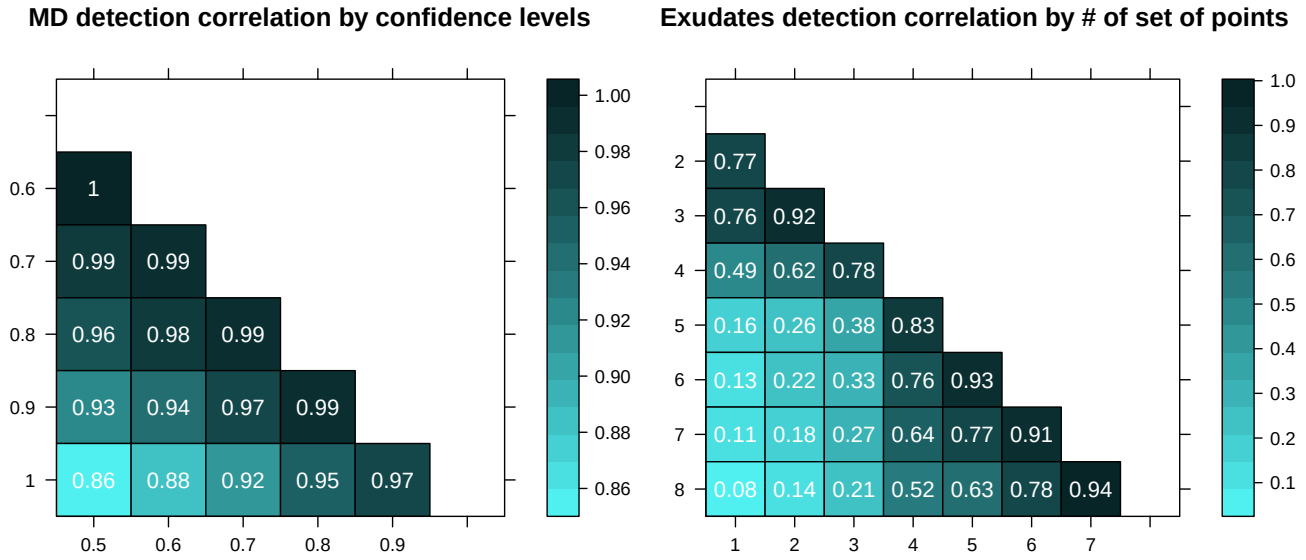


**Figure 3.** Correlations between the different confidence levels of the MD detection, in the left. In the right, correlations between the different numbers (#) of set of points of the exudates detection.

In the scatter plots for the exudates detection by several sets of points, in Figure 4, a linear relationship is observed only for very close numbers of the set of points. Conform the difference between this numbers became larger, the linear behavior disappears, and the correlation goes to less than 0.4 (right-graph of Figure 3). We also see in the scatterplots that, in general, exist much more variability among the values of the patients with signs of DR (in orange). Comparing the euclidian distance and the diameter features with the others, none evident stronger relation is observed.

## 2.2 Modeling Process & Methodology

Before study the effect of all the variables together with the goal of seeing which features are significant to explain the signs of DR and to predict this signs, in the presence of the others, we looked for some of the features individually.

To verify if their means are different from one response group (signs of DR or not) to the other, we used a $t$-test.

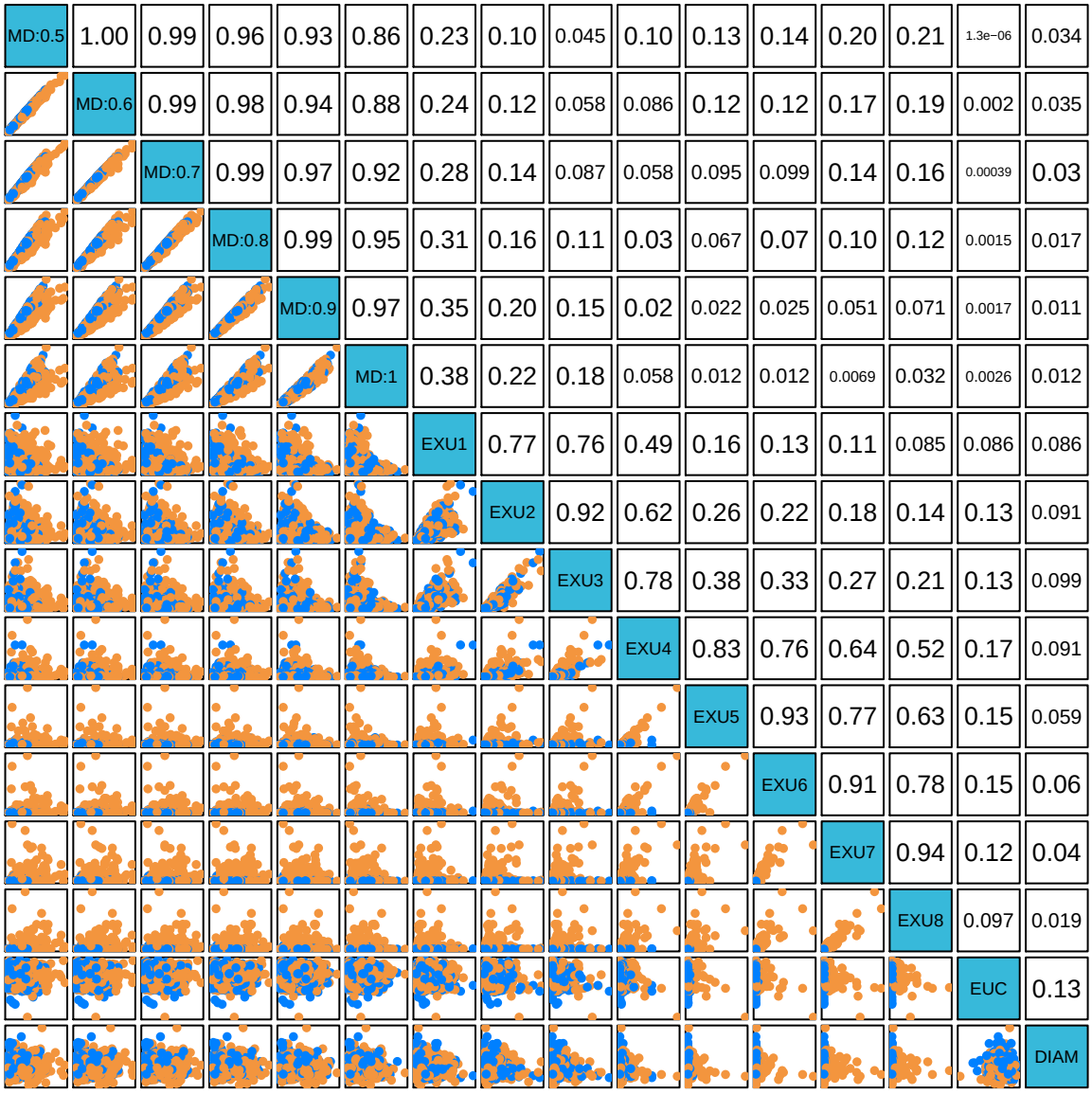


**Figure 4.** Scatter plots and correlations for all numeric features. In blue the pacients without signs of diabetic retinopathy (DR), in orange the pacients with signs of DR.

The formula of the $t$-test statistic is described in the Equation 1, with $W$ being a weight (the sample size of one group divided by the total sample size) for the sample size and with $S^2$ being the estimated sample variance among each group.

$$t_{\text{est}} = \frac{\bar{X}_{\text{No}} - \bar{X}_{\text{Yes}}}{\sqrt{S_p^2 \cdot \left(\frac{1}{n_{\text{No}}} + \frac{1}{n_{\text{Yes}}}\right)}}, \quad \text{with} \quad S_p^2 = W_{\text{No}} \cdot S_{\text{No}}^2 + W_{\text{Yes}} \cdot S_{\text{Yes}}^2. \tag{1}$$

The test statistic $t_{\text{est}}$ follow ($\sim$) a $t$-distribution with $n = n_{\text{No}} + n_{\text{Yes}}$ degrees of freedom.

To test the association between variables we use the $\chi^2$-test, with it formula is described in Equation 2.

$$\chi^2_{\text{est}} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \quad \text{with } i \text{ being the intersection classes of the variables.} \tag{2}$$

For each intersection of the classes of the variables we have the observed counts, expressed by $O$, and the expected counts expressed by $E$. The expected counts can be computed as the product of the marginal totals divided by the total overall. The test statistic $\chi^2_{\text{est}}$ follow ($\sim$) a $\chi^2$ distribution with (number of classes in the first variable - 1) $\cdot$ (number of classes in the second variable - 1) degrees of freedom.

To verify the significance of one feature in the presence of others we used the logistic regression model. The logistic regression is the most famous and used model in medicine and epidemiology, and the reason for this is because this methodology combines simplicity, power and interpretation. Simplicity because isn't a very complex model, powerful because this model is able to provide very good results in a general way and their parameter interpretation can be given in terms of odds ratio. The logistic regression [7] can be understood as finding the values of the $\beta$ parameters that best fit $Y|X \sim$ Bernoulli distributed, with expectation:

$$E(Y|X = x) = P(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

The logistic function is defined by:

$$F(x) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x))}.$$

The inverse of the logistic function, $g$, also called of logit (log odds) is defined by:

$$g(F(x)) = \ln\left(\frac{F(x)}{1 - F(x)}\right) = \beta_0 + \beta_1 x \qquad \Rightarrow \qquad \frac{F(x)}{1 - F(x)} = \exp(\beta_0 + \beta_1 x) = \text{OR}.$$

Where:

- $g$ is the logit function. The equation for $g(F(x))$ illustrates that the logit (i.e., the log-odds) is equivalent to the linear regression expression.
- $F(x)$ is the probability that the response variable equals a case, given some linear combination of the predictors. This is important in that it shows that the value of the linear regression expression can vary from negative to positive infinity and yet, after transformation, the resulting expression for the probability $F(x)$ ranges between 0 and 1.
- $\beta_0$ is the intercept from the linear regression equation (the value of the criterion when the predictor is equal to zero).
- $\beta_1 x$ is the regression coefficient multiplied by some value of the feature.

The terms interpretation can be done by odds ratio (OR).

In the context of generalized linear models for binary data, the logit is the canonical link function and when used the resulting model is called of logistic regression. However, other link function can be used [7] and [8]. These link functions are:

- Probit or inverse Normal function: $g(F(x)) = \Phi^{-1}(F(x))$.

- Complementary log-log function: $g(F(x)) = \log(-\log(1-F(x)))$.
- Cauchit function: $\tan\left(\pi F(x) - \frac{\pi}{2}\right)$.

To estimate the model we used maximum likelihood. To test the significance of the coefficients we used the Akaike information criterion (AIC). Given a collection of models, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection. The AIC value of a given model is the following:

$$\text{AIC} = 2p - 2\log\hat{L}.$$

With $\hat{L}$ being the maximum value of the likelihood function for the model and $p$ being the number of estimated parameters in the model. More details about this technique can be seen in [9].

Thinking in the classification, we separate the data into two parts. One for training the model and other for test. In general between $60 \sim 70\%$ of the data are separeted for the train, and the rest stays for the test.

Looking to the Figures 3 and 4 and thinking in the nature of the data, we see a correlation between the MD variables and the EXU variables. When we have this type of characteristic a famous alternative is the use of principal component analysis (PCA). PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of distinct principal components is equal to the smaller of the number of original variables or the number of observations minus one. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set [10].

To do all the fits and computation we used the R language [11].

### 2.3 Diagnosis & Goodness of Fit

Under the null hypothesis that the model fit is satisfactory, to verify the goodness of fit we used statistics that summarise the concordance among the observed values and the predicted values by the model. In the presence of continuous features, the most popular statistic is the test of Hosmer and Lemeshow [12] and [13]. Beyond this we also used the Pearson and Deviance residuals, the sensitivity, specificity, predict value and the ROC curve [14] and [15].

## 3. Results

As a first step we looked marginally to the two continuous variables that aren't strictly related to the others. We are talking about the (1) euclidian distance of the center of the macula to the center of the optic disc and (2) the diameter of the optic disc. To verify if we have evidence of the difference between the means of each one of these variables in relation to the presence (or not) of signs of DR we used a $t$-test. We tested a null hypothesis $H_{\text{Null}}$ of equality, i.e., that the difference of the means isn't statistically significant, versus an alternative hypothesis $H_{\text{Alt}}$ of significant difference. In the Tables 4 and 5 we present the results of the $t$-test for the two variables.

**Table 4.** Summary of the $t$-test results to the euclidian distance, by sign of DR.

| Sign of DR | Mean: Euclidian distance | $t$-stastistic | Reference distribution | Decision |
|---|---|---|---|---|
| No<br>Yes | 0.52296<br>0.52344 | -0.28699 | 1.64618 | No statistical evidence of diferrence between the means |

We see that for both variables the means are extremely similar in each group (presence or not of signs of DR), and in consequence the reference distribution is bigger than the test statistic in both cases, which means that we don't have enough evidence to reject the null hypothesis.

**Table 5.** Summary of the $t$-test results to the diameter, by sign of DR.

| Sign of DR | Mean: Diameter | $t$-stastistic | Reference distribution | Decision |
|---|---|---|---|---|
| No<br>Yes | 0.10902<br>0.10791 | 1.04682 | 1.64618 | No statistical evidence of diferrence between the means |

Thinking in a regression model with several variables, with this result we can already expect that this two variables, (1) euclidian distance of the center of the macula and the center to the optic disc and (2) the diameter of the optic disc will not be significant to separate the patients between the two groups.

In Figure 2 we saw that using the AM/FM classification 1/3 of the patients present pathological lesions in the retinal structures. In Table 6 we compared the AM/FM classification of the patients with the DR classification.

Among the patients with no signs of DR, 36% (193/540) presented pathological lesions in the retinal structures. Among the patients with signs of DR, 68% presented normal retinal structures. In the cells between parentheses we have the expected counts, where we see a small difference to the observed counts. Performing a $\chi^2$-test we obtained a test statistic of 2.044. For a probability of Type I error of $\alpha = 0.05$ with 1 degree of freedom, the rejection region is determined by the value 3.841, which is bigger than the value of the test statistic. Therefore, we don't have significant statistical evidence to reject a null hyphotesis of lack of association between the result of the AM/FM-based classification and the patient status (with or without signs of DR).

**Table 6.** Comparison of the AM/FM-based classification with the DR situation of the patients, by the observed counts of patients. In parentheses are presented the respective expected counts.

| AM/FM-based classification | No sign of DR | Sign of DR | Total |
|---|---|---|---|
| Normal retinal structures | 347 (358) | 417 (406) | 764 |
| Pathological lesions | 193 (182) | 194 (205) | 387 |
| **Total** | 540 | 611 | 1151 |

To see and understand the behavior of the features in the patients status, we fitted a logistic regression (with logit link function). We start with all features and using as criterium the AIC we arrived in a final model wehre the results can be seen in Table 7. To find the CI we used the follow approach: estimated coefficient $\pm$ standard normal distribution 95% quantil times $(\cdot)$ coefficient standard error.

Looking to the results of this first model, presented in Table 7, we see in the $p$-value column that for the MD features only the values for the 0.9 confidence level weren't significant (for this reason aren't present in the Table), considering a significance level of 10%. For the exudates detection features only two wasn't significant, with three and eight set of points. As we already expected by the results in the Tables 4 and 5, the euclidian distance and the diameter not shown to be significant. Considering a significance level of 10%, the AM/FM-based classification shown to be significant.

To see the significance we don't need to look exclusively to the $p$-value, we can also look to the confidence intervals (CI). If the interval contain the value zero it's a clue that the variable may be not significant. The interval range is also very informative. If it's big, the uncertainty about the coeficient is greater.

With the variables present in this final model obtained with a logit link function we fitted more three models with different link functions. This models are compared by the AIC in Table 8. The better fit is obtained with the logit

link function, this means that we stay with the fitted logistic regression with the estimated coefficients presented in Table 7.

**Table 7.** Summary of the final fitted logistic regression with the estimated coefficients, lower and upper of the 95% confidence interval (CI) and related p-values. In bold, for easy identification, are the significant coefficients at a level of 5% and the CI's that don't include zero.

| Features | Lower CI | Point estimate | Upper CI | *p*-value |
|---|---|---|---|---|
| Intercept | -1.46297 | -0.44403 | 0.57490 | 0.39304 |
| MD: 0.5 | 0.71822 | **0.90967** | **1.10112** | **0.00000** |
| MD: 0.6 | -0.68165 | **-0.43941** | **-0.19717** | **0.00038** |
| MD: 0.7 | -0.49159 | **-0.30424** | **-0.11689** | **0.00146** |
| MD: 0.8 | -0.32245 | **-0.20867** | **-0.09489** | **0.00033** |
| MD: 1 | -0.00097 | **0.04091** | **0.08279** | **0.05555** |
| EXU 1 | 0.00453 | **0.00898** | **0.01343** | **0.00008** |
| EXU 2 | -0.02827 | **-0.01508** | **-0.00190** | **0.02498** |
| EXU 4 | -0.28716 | **-0.15126** | **-0.01537** | **0.02914** |
| EXU 5 | -0.12170 | **0.38254** | **0.88679** | 0.13704 |
| EXU 6 | -4.07110 | -1.80779 | 0.45552 | 0.11747 |
| EXU 7 | 1.62021 | **7.86642** | **14.11262** | 0.01357 |
| DIAM | -14.54667 | -6.29276 | 1.96115 | 0.13510 |
| AM/FM | -0.64448 | -0.29136 | 0.06177 | 0.10585 |

**Table 8.** AIC of the models fitted with different link functions. AIC: small is better. The smallest AIC is in bold, for easy identification.

| **Link function** | Logit | Probit | Complementary log-log | Cauchit |
|---|---|---|---|---|
| **AIC** | **1133.257** | 1135.597 | 1147.505 | 1141.221 |

To verify the final model goodness of fit we used the Hosmer and Lemeshow test that results in a *p*-value of 0.59166. With a *p*-value of this magnitude has no evidence to reject the null hypothesis that the fit of the model is satisfactory.

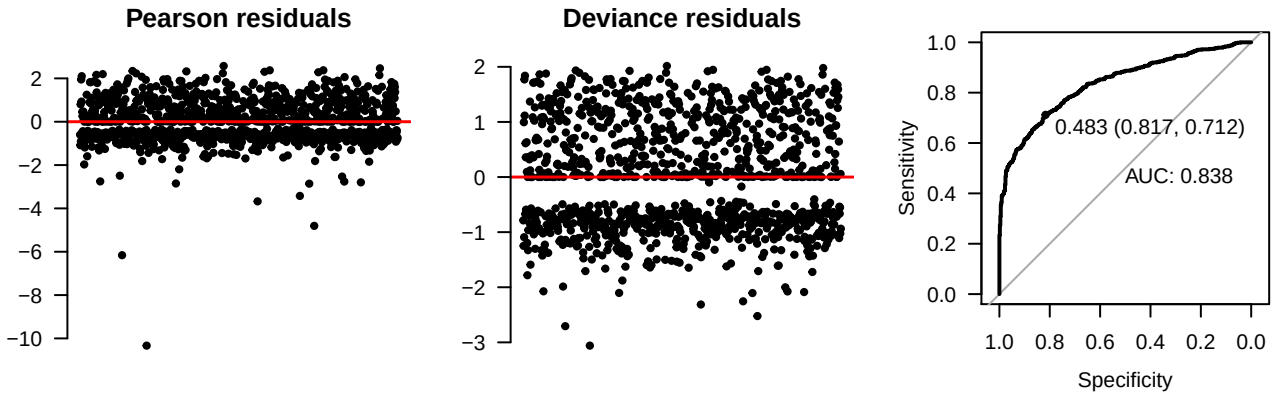


**Figure 5.** Dispersion of the Pearson and Deviance residuals, in the left and in the center, respectively. ROC curve in the right, with AUC value, cutoff, specificity and sensitivity.

The Pearson and the Deviance residuals are presented in the Figure 5. If the model is well adjusted it is expected that these residues follow a standard normal distribution, and in this way the most of the observations have to stay

present in the interval -3 and 3 (99.7% of the data within three standard deviations to the mean, zero). Also in Figure 5 we have the ROC curve. Area Under the Curve (AUC) superior than 0.70 is interpreted as a good fit for the model. All this characteristics are shown in Figure 5.

With this model we also trained a predictive model. For this we randomly splited the dataset in two parts, called in literature of train and test. As the name say, with the train dataset we train the model and test the results doing a prediction in the test dataset. Here we splited in 70 and 30. 70% for the train and 30% for test. In Table 9 are presented the number of observations in each dataset and the percentage of patients with and without signs of DR. We see that the patients are well distributed in the datasets.

**Table 9.** Number (#) of patients and the number (and percentage) of patients with and without signs of DR in each dataset.

| | # of patients | Patients with no signs of DR | Patients with signs of DR |
|---|---|---|---|
| **Train dataset** | 806 | 386 (48%) | 420 (52%) |
| **Test dataset** | 345 | 154 (45%) | 191 (55%) |

The results of the predictive model is seen in Figure 6. The AUC above 0.70 means that the model have a good fit. The better AUC, specificity and sensitivity is found using a threshold of 0.513, very similar to the standard 0.5. Both specificity (true negative rate - proportion of patients without signs of DR that are correctly identified as such) and sensitivity are bigger than 0.65, with a bigger sensitivity (true positive rate - proportion of patients with signs od DR that are correctly identified as such), 0.805.

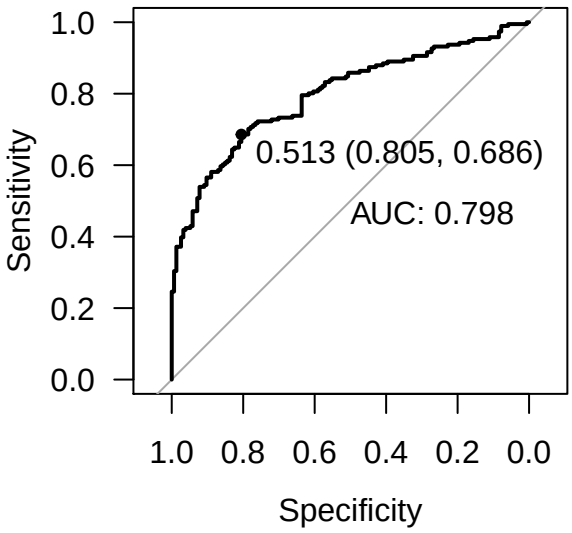


**Figure 6.** ROC curve with the results (AUC, specificity, sensitivity and threshold) of the predictive model. The model was fitted using the train dataset and the prediction was performed in the test dataset.

Other, and final, model that we fitted is a model considering principal component analysis, PCA. Instead of using all the MD and EXU variables together, that present a strong level of correlation (Figures 3 and 4), we performed a PCA and considered as covariables in the model only the first principal component of the MD variables and the first principal component of the EXU variables. The fitted model was a model with five covariables. The two first principal components, the euclidian distance, the diameter and the AM/FM-based classification.

Performing a variable selection by the AIC we finish with a model with only the first components, which means that in the presence of the principal components the euclidian distance, the diameter and the AM/FM-based classification aren't statiscally significant. This model present an AIC of 1507.002, which is much bigger than the

AIC of the models presented in Table 8. This means that the model with the results presented in Table 7 still shown to be better and that the presence of the set of MD and EXU variables aren't detrimental to the model goodness-of-fit.

## 4. Conclusion

Practically all the patients in the study have a sufficient quality assessment and present a SRA. The means of the euclidian distance of the center of the macula to the center of the optic disc are practically the same (without a statistical difference), independent from if the patient present or not signs of DR. The same conclusion can be made for the diameter of the optic disc.

The fitted final model with all the variables presented a satisfactory goodness-of-fit, with a specificity (true negative rate) and sensitivity (true positive rate) superior than 0.70, and with an AUC over 0.80. The estimated cutoff of the probability to classify the patients in one of the two status is very close to 0.5.

About the features, almost all the MD and EUX variables are present in the final model, showing that they are statistically significant to classify the patients. Thinking in the correlation between this variables, a PCA analysis was performed and the first principal components was used as covariables. The resulting model presented a good fit, but much lower than the model with all the variables, as the AIC obtained shown. With the predictive model the results was also very satisfatory. The obtained AUC was greater than 0.70 and the specificity (true negative rate) and sensitivity (true positive rate) was greater than 0.65.

From the Table 7 we can achieve some very interesting interpretations and conclusions about the variables coefficients. In the list below we give the interpretation of each one using odds ratio.

- AM/FM-based classification.
  - The odds ratio ($\widehat{\text{OR}}$), chance, of a patient with positive (pathological lesions) AM/FM-based classification present signs of DR is 0.52 ($\widehat{\text{OR}} = \exp(-0.64448)$) times that of those with negative (normal retinal structures) AM/FM-based classification, with both having all the same values in the others characteristics.
- Microaneurism Detection (MD) at different confidence levels.
  - For each one more microaneurisms found at confidence level of 0.5, the odds ratio, chance, of present signs of DR is 2.05 ($\exp(0.71822)$) times that of those with one less, i.e., the odds increase.
  - For each one more microaneurisms found at confidence level of 0.6, the odds ratio of present signs of DR is 0.51 ($\exp(-0.68165)$) times that of those with one less, i.e., the odds decrease.
  - For each one more microaneurisms found at confidence level of 0.7, the odds of present signs of DR is 0.61 ($\exp(-0.49159)$) times that of those with one less, i.e., the odds decrease.
  - For each one more microaneurisms found at confidence level of 0.8, the odds of present signs of DR is 0.72 ($\exp(-0.32245)$) times that of those with one less, i.e., the odds decrease.
  - For each one more microaneurisms found at confidence level of 1, the odds of present signs of DR is 1 ($\exp(-0.00097)$) times that of those with one less, i.e., the odds decrease.
- Exudates detection in different set of points.
  - Number of points: 1. Exudates mean: 64.097. The odds ratio, chance, of a patient with EXU 1 of 69.097 present signs of DR is 1.02 ($\exp(0.00453 \cdot (69.097 - 64.097))$) times that of those with EXU 1 of 64.097.
  - Number of points: 2. Exudates mean: 23.088. The odds ratio of a patient with EXU 2 of 28.088 present signs of DR is 0.87 ($\exp(-0.02827 \cdot (28.088 - 23.088))$) times that of those with EXU 2 of 23.088.
  - Number of points: 4. Exudates mean: 1.836. The odds of a patient with EXU 4 of 3.836 present signs of DR is 0.56 ($\exp(-0.28716 \cdot (3.836 - 1.836))$) times that of those with EXU 4 of 1.836.

  - Number of points: 5. Exudates mean: 0.561. The odds of a patient with EXU 5 of 2.561 present signs of DR is 0.78 ($\exp(-0.1217 \cdot (2.561 - 0.561))$) times that of those with EXU 5 of 0.561.
  - Number of points: 6. Exudates mean: 0.212. The odds of a patient with EXU 6 of 1.212 present signs of DR is 0.02 ($\exp(-4.0711 \cdot (1.212 - 0.212))$) times that of those with EXU 6 of 0.212.
  - Number of points: 7. Exudates mean: 0.086. The odds of a patient with EXU 7 of 1.086 present signs of DR is 5.05 ($\exp(1.62021 \cdot (1.086 - 0.086))$) times that of those with EXU 7 of 0.086.

- Diameter of the optic disc.

  - Mean: 0.108. The odds of a patient with diameter of the optic disc of 0.358 present signs of DR is 0.03 ($\exp(-14.54667 \cdot (0.358 - 0.108))$) times that of those with diameter of 0.108.

We saw that as we increase the confidence level of the MD detections the odds ratio approach zero, which means that for high confidence levels differences in the MD detection doesn't impact the chance of the patient present signs of DR. For the exudates detection we see that the difference between the odds ratio aren't big, with an exception in the detection with the set of seven points. With this set the odds ratio between the groups is big.

With the diameter of the optic disc the differences between the odds ratio is also not so big. A bigger difference is observed when we compare the odds ratio of the patients by the AM/FM based-classification.

As a final conclusion we can highlight that almost all the variables stay present in the final model, this show that in general all the variables measured are useful when putted together in the model to classify patients about the presence of signs of DR. Looking individually to each coefficient we don't see considerable differences. The final model and the predictive model present very good results with satisfactory measures of accuracy and quality of fit and prediction. The model considering the canonical link function, logistic, presented the best results when compared with others.

## 5. References


[1] Machine Learning Repository. URL: https://goo.gl/9twv8K. Accessed at 5 November 2017.

[2] American Optometric Association. URL: https://goo.gl/rVfqju. Accessed at 5 November 2017.

[3] "American Academy of Ophthalmology, What Is Diabetic Retinopathy?" URL: https://goo.gl/idx3sO. Accessed at 5 November 2017.

[4] "National Eye Institute, Facts About Diabetic Eye Disease." URL: https://goo.gl/sHvKk0. Accessed at 5 November 2017.

[5] HAJAR, S., et all. (2015). Prevalence and causes of blindness and diabetic retinopathy in Southern Saudi Arabia. *Saudi Medical Journal*, 36(4): 449-455. URL: https://goo.gl/4Yt1dE.

[6] Computer-Aided Diagnosis of Retinal Images (CADR). URL: https://goo.gl/Fe7GSW.

[7] McCULLAGH, P. and NELDER, J.A. (1983). *Generalized Linear Models*. Chapman and Hall, Second Edition (1989). Monographs on Statistics and Applied Probability 37.

[8] GUNDUZ, N. and FOKOUE, E. (2017). On the Predictive Properties of Binary Link Functions. *Communications Series A1: Mathematics and Statistics*, 66(1): 1-18. URL (arXiv preprint): https://goo.gl/pGkBDm.

[9] Akaike information criterion. From Wikipedia, the free encyclopedia. URL: https://goo.gl/sRJ16t. Accessed at 27 November 2017.

[10] Principal component analysis. From Wikipedia, the free encyclopedia. URL: https://goo.gl/QPTKwx. Accessed at 10 December 2017.

[11] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[12] Hosmer-Lemeshow test. From Wikipedia, the free encyclopedia. URL: https://goo.gl/8SgkaB. Accessed at 9 November 2017.

[13] JAY, M. (2017). generalhoslem: Goodness of Fit Tests for Logistic Regression Models. R package version 1.3.0. URL: https://goo.gl/7VG9Ke.

[14] Receiver operating characteristic. From Wikipedia, the free encyclopedia. URL: https://goo.gl/ret2fX. Accessed at 9 November 2017.

[15] ROBIN, X., et all. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, p.77. DOI: 10.1186/1471-2105-12-77. URL: https://goo.gl/fBG1We.

# 6. Appendix

## 6.1 Appendix A: R code for $t$-test implementation

Here we present the code for the $t$-test with the euclidian distante variable. For the diameter variable the procedure is exactly the same.

```
# dataset: da ================================================================ #
# response variable: dr (two levels, "yes" and "no"). 17th column of da == #
#                    yes = present signs of DR ============================== #
#                    no = shows no signs of DR ============================== #

## means ======================================================================= ##
# euc: euclidean distance of the center of the macula to the center of the
#      optic disc
euc.no <- da[da$dr == "no", 17] ; euc.yes <- da[da$dr == "yes", 17]

## weights ===================================================================== ##
# wei: weights
wei.no  <-  (length(euc.no)-1) / ((length(euc.no)-1) + (length(euc.yes)-1))
wei.yes <- (length(euc.yes)-1) / ((length(euc.no)-1) + (length(euc.yes)-1))

## pooled variance estimate ==================================================== ##
sp2 <- wei.no * var(euc.no) + wei.yes * var(euc.yes)

## numerator and denominator of the test statistic ============================= ##
num <- (mean(euc.no) - mean(euc.yes))
deno <- sqrt(sp2 * (1/length(euc.no) + 1/length(euc.yes)))

## test statistic and value of the reference distribution ====================== ##
tt <- num / deno
ref <- qt(.95, nrow(da) - 2) # probability of type I error of \alpha = 0.05
                             # degrees of freedom: number of observations
                             #                     (nrow(da))-1
# if tt > ref => reject the null hyphotesis ================================== ##
```

### 6.2 Appendix B: R code for $\chi^2$-test implementation

```r
# variable: am/fm-based classification ================================== #

## expected counts ======================================================= ##
# exp: expected count
# neg: negative am/fm classification
# pos: positive am/fm classification
# nodr: shows no signs of DR
# dr:   present signs of DR
# amfm: am/fm-based classification
exp_neg.nodr <-
  nrow(da[da$amfm == "neg", ]) * nrow(da[da$dr == "no", ])  / nrow(da)
exp_neg.dr   <-
  nrow(da[da$amfm == "neg", ]) * nrow(da[da$dr == "yes", ]) / nrow(da)
exp_pos.nodr <-
  nrow(da[da$amfm == "pos", ]) * nrow(da[da$dr == "no", ])  / nrow(da)
exp_pos.dr   <-
  nrow(da[da$amfm == "pos", ]) * nrow(da[da$dr == "yes", ]) / nrow(da)

## test statistic and value of the reference distribution ================ ##
# obs: observed count
obs_neg.nodr <- nrow(da[da$amfm == "neg" & da$dr == "no",  ])
obs_neg.dr   <- nrow(da[da$amfm == "neg" & da$dr == "yes", ])
obs_pos.nodr <- nrow(da[da$amfm == "pos" & da$dr == "no",  ])
obs_pos.dr   <- nrow(da[da$amfm == "pos" & da$dr == "yes", ])

neg.nodr <- ((obs_neg.nodr - exp_neg.nodr)**2) / exp_neg.nodr
neg.dr   <- ((obs_neg.dr   - exp_neg.dr  )**2) / exp_neg.dr
pos.nodr <- ((obs_pos.nodr - exp_pos.nodr)**2) / exp_pos.nodr
pos.dr   <- ((obs_pos.dr   - exp_pos.dr  )**2) / exp_pos.dr

# chi: \chi^{2} test statistic
chi <- neg.nodr + neg.dr + pos.nodr + pos.dr
ref <- qchisq(.95, 1) # probability of type I error of \alpha = 0.05
                      # degrees of freedom: number of classes of one
                      #                     variable (amfm) minus (-) 1
                      #                     times the number of classes of
                      #                     the other variable (dr) - 1.
                      #                     2-1 x 2-1 = 1
# if chi > ref => reject the null hyphotesis ============================ ##
```

## 6.3 Appendix C: Variable selection using the AIC as a criterium

The variable selection was made with the `stepAIC()` R function. All the steps are presented in Table 10.

In the first step we see that the smallest AIC (with the AIC the lower the better) is obtained in the model without the variable EXU 8 (looking to more decimal places). In the second step the better AIC is obtained when we take out the variable EUC. In the third step we take out the EXU 3 variable, in the fourth step the variable MD: 0.9 and in the fifth step the better AIC is obtained when all the variables are present, i.e., we don't need to take out more variables, this is the best model by the AIC criterium.

**Table 10.** Selecion of variables in five steps. In each step is presented the AIC of the model without the respective feature. In bold is presented the smallest, better, AIC at each step.

| **1st step** | | **2nd step** | | **3rd step** | | **4th step** | | **5th step** | |
|---|---|---|---|---|---|---|---|---|---|
| **Feature** | **AIC** | **Feature** | **AIC** | **Feature** | **AIC** | **Feature** | **AIC** | **Feature** | **AIC** |
| MD: 0.5 | 1262.5 | MD: 0.5 | 1260.5 | MD: 0.5 | 1258.5 | MD: 0.5 | 1257.2 | MD: 0.5 | 1257.1 |
| MD: 0.6 | 1151.5 | MD: 0.6 | 1149.5 | MD: 0.6 | 1147.5 | MD: 0.6 | 1145.6 | MD: 0.6 | 1144.3 |
| MD: 0.7 | 1148.9 | MD: 0.7 | 1146.9 | MD: 0.7 | 1144.9 | MD: 0.7 | 1143.0 | MD: 0.7 | 1141.7 |
| MD: 0.8 | 1144.8 | MD: 0.8 | 1142.9 | MD: 0.8 | 1141.0 | MD: 0.8 | 1139.1 | MD: 0.8 | 1144.6 |
| MD: 0.9 | 1139.1 | MD: 0.9 | 1137.1 | MD: 0.9 | 1135.2 | **MD: 0.9** | **1133.3** | - | - |
| MD: 1 | 1142.5 | MD: 1 | 1140.5 | MD: 1 | 1138.6 | MD: 1 | 1136.8 | MD: 1 | 1135.0 |
| EXU 1 | 1153.3 | EXU 1 | 1151.3 | EXU 1 | 1149.5 | EXU 1 | 1149.0 | EXU 1 | 1147.7 |
| EXU 2 | 1141.8 | EXU 2 | 1139.8 | EXU 2 | 1137.8 | EXU 2 | 1137.9 | EXU 2 | 1136.4 |
| EXU 3 | **1138.5** | EXU 3 | 1136.6 | **EXU 3** | **1134.6** | - | - | - | - |
| EXU 4 | 1141.3 | EXU 4 | 1139.3 | EXU 4 | 1137.4 | EXU 4 | 1137.9 | EXU 4 | 1136.5 |
| EXU 5 | 1141.4 | EXU 5 | 1139.7 | EXU 5 | 1137.7 | EXU 5 | 1135.9 | EXU 5 | 1134.7 |
| EXU 6 | 1140.7 | EXU 6 | 1139.2 | EXU 6 | 1137.2 | EXU 6 | 1135.4 | EXU 6 | 1134.1 |
| EXU 7 | 1140.9 | EXU 7 | 1145.3 | EXU 7 | 1143.3 | EXU 7 | 1141.5 | EXU 7 | 1140.2 |
| **EXU 8** | **1138.5** | - | - | - | - | - | - | - | - |
| EUC | **1138.5** | **EUC** | **1136.5** | - | - | - | - | - | - |
| DIAM | 1140.7 | DIAM | 1138.7 | DIAM | 1136.7 | DIAM | 1134.8 | DIAM | 1133.5 |
| AM/FM | 1141.3 | AM/FM | 1139.3 | AM/FM | 1137.4 | AM/FM | 1135.4 | AM/FM | 1133.9 |
| Model | 1140.5 | Model | 1138.5 | Model | 1136.5 | Model | 1134.6 | **Model** | **1133.3** |

### 6.4 Appendix D: Other computations in R

```r
## generic code to fit the model ========================================== ##
model <- glm(response ~ variables # variable1 + variable2 + ...
             , family = binomial(link = logit) # or probit, cloglog, cauchit
             , data_frame)
AIC(model) ## getting the AIC ============================================= ##

stepAIC(model) ## performing the model selection by the AIC criterium === ##

## computing confidence intervals ========================================= ##
confint.default(model) # estimated coefficient plus/minus 95% normal quantil
                       # times coefficient standard error

## performing the Hosmer-Lemeshow test ================================== ##
# logitgof function is in the generalhoslem package
logitgof(model$response, fitted(model))

## computing the Pearson and Deviance residuals ========================== ##
residuals(model, type = "pearson")
residuals(model, type = "deviance")

## computing the ROC curve ================================================ ##
# roc function is in the pROC package
roc(model$response, fitted(model))
# use this inside the plot.roc() function with the arguments
# print.auc = TRUE and print.thres = TRUE to plot the ROC curve ========= ##
```