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 the accuracy of the prediction by using this methodology with different link functions, to predict whether a particular person has a diabetic retinopathy signs disease or not. 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 preliminary analysis, we see 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 see through a t-test that their means don't differ significantly, being in reality very closer. Fitting a logistic regression with all the features this same result was obtained. In a initial analysis the goodness of fit of the model was satisfactory, having almost all features related with Microaneurism Detection (MD) as significant. In the next steps a selection of variables will be made and others link funcions in the model will be tested. In the end, besides the intrepretation of the coefficients, a predictive model will be trained.

Keywords
Diabetic Retinopathy; t-Test; Logistic Regression; Link functions.

[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

The diabetes is the 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.

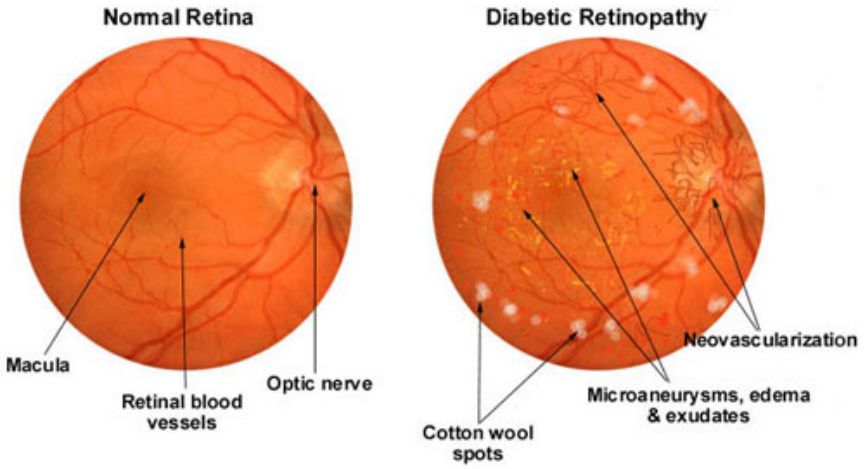

Figure 1. Retinal Fundus image.

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 α = 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 will provide a quick way for an early detection of diabetic retinopathy so the patient can receive an early treatment that can limit the potential for significant vision loss.

The principal goal of this study is verifying 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 is verifying 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., using this instrument 33.6% of the patients present pathological lesions in the retinal structures.

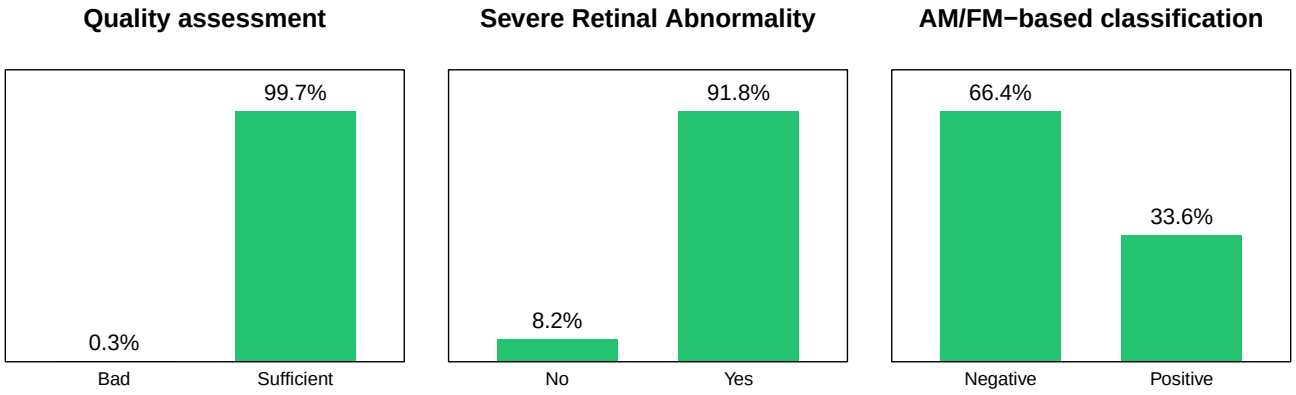

Figure 2. Barcharts for the categorical features.

In the Figure 4 are provided 2x2 scatter plots and the correlations for all the numerical variables. 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.

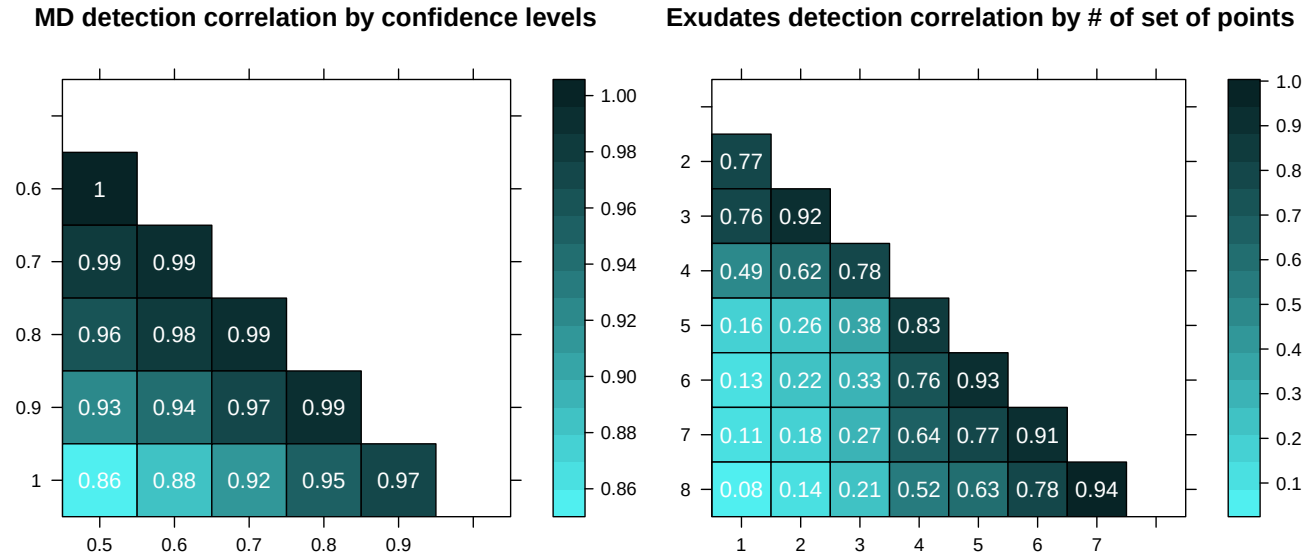

Figure 3. Correlations between the different confidence levels of the MD detection, in the left. In the right, correlations between the different numbers (#) of set of points of the exudates detection.

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.

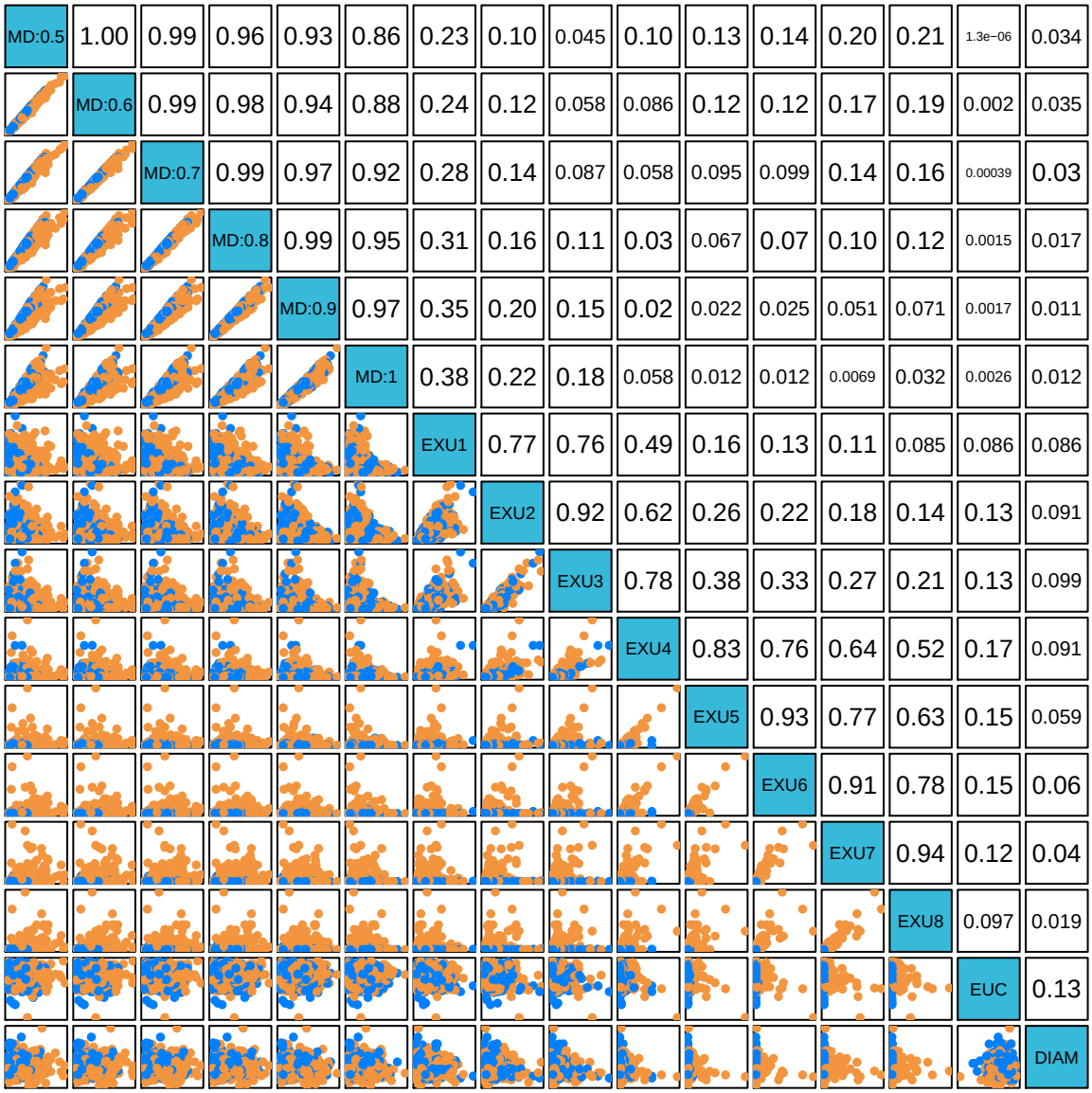

Figure 4. Scatter plots and correlations for all numeric features. In blue the pacients without signs of diabetic retinopathy (DR), in orange the pacients with signs of DR.

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 can look 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 use a t-test.

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.

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

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

To verify the significance of one feature in the presence of others, we choosed to use the logistic regression model. The logistic regression is the most famous and used model in medicine and epidemiology, 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 is given in terms of odds ratio.

The logistic regression [7] can be understood as finding the values of the β parameters that best fit:

$$y = \begin{cases} 1 & \text{if } \beta_0 + \beta_1 x + \varepsilon > 0 \\ 0 & \text{if otherwise} \end{cases}, \quad \text{where } \varepsilon \text{ is an error distributed by the standard logistic distribution.}$$

And the logistic function is defined by:

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

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

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

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.
- β_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.

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

For estimate the model we use maximum likelihood. To test the significance of the coefficients we can use likelihood ratio test or the Wald statistic. More details about this techniques can be seen in [9].

Thinking in the classification, we can 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.

To do all the fits and computation we use the R language [10].

2.3 Diagnosis & Goodness of Fit

Under the null hypothesis that the model fit is satisfactory, to verify the goodness of fit, we can use 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 [11] and [12].

Beyond this we can also use the Pearson and Deviance residuals, the sensitivity, specificity, predict value and the ROC curve [13] and [14].

3. Results

As a first step, we can look marginally for 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_{Null} of equality, i.e., that the difference of the means isn't statistically significant, versus an alternative hypothesis H_{Alt} of significant difference. In the Tables 3 and 4 we present the results of the t-test for the two variables.

Table 3. Summary of the t-test results for the euclidian distance by sign of DR.

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

Table 4. Summary of the t-test results for the diameter by sign of DR.

Sign of DR	Mean: Diameter	t-stastistic	Reference distribution	Decision
No Yes	0.10902 0.10791	1.04682	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.

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 5 we compare the AM/FM classification of the patients with the DR classification. Among the patients with no signs of DR, 36% (193/540) present pathological lesions in the retinal structures. Among the patients with signs of DR, 68% present normal retinal structures.

Table 5. Comparison of the AM/FM-based classification with the DR situation of the patients.

AM/FM-based classification	**No sign of DR**	**Sign of DR**	**Total**
Normal retinal structures	347	417	764
Pathological lesions	193	194	387
Total	540	611	1151

To start, we fitted a logistic regression (logit link function) considering all the features. With this first model, we can start to do variables (features) selection using likelihood ratio test.

Looking now only for this first model we have the Table 6. There we can see in the p-value column that for the MD features, only the values for the 0.9 confidence level weren't significant, considering a significance level of 10%. For the exudates detection features only two was significant, with one and two set of points.

As we already expected by the results in the Tables 3 and 4, 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.

Table 6. Summary of the fitted logistic regression with the estimated coefficients, standard errors and related p-values (p-values less than 0.1 in bold).

Features	Estimate	Standard Error	p-value
Intercept	-0.10510	1.60807	0.94789
MD:0.5	0.90624	0.09795	**0.00000**
MD:0.6	-0.43858	0.12350	**0.00038**
MD:0.7	-0.30495	0.09571	**0.00144**
MD:0.8	-0.17570	0.07023	**0.01236**
MD:0.9	-0.04016	0.04970	0.41905
MD:1	0.05247	0.02616	**0.04486**
EXU1	0.00882	0.00234	**0.00016**
EXU2	-0.01739	0.00964	**0.07117**
EXU3	0.00849	0.02972	0.77512
EXU4	-0.17521	0.10752	0.10318
EXU5	0.38029	0.27296	0.16355
EXU6	-1.69756	1.28095	0.18509
EXU7	7.30653	5.29844	0.16790
EXU8	0.86691	7.02143	0.90174
EUC	-0.63078	2.72319	0.81682
DIAM	-6.32060	4.28378	0.14009
AM/FM	-0.30395	0.18116	**0.09338**

To verify the goodness of fit we use the Hosmer and Lemeshow test, that results in a p-value of 0.80146. With a p-value of this magnitude has no evidence to reject the null hypothesis that the fit of the model is satisfactory.

The Pearson and the Deviance residuals are presented in the left and the center of Figure 5. If the model is well adjusted, it is expected that these residues follow a standard normal distribution, and in this way that the most of the observations stay present in the interval -3 and 3. In the right-graph of Figure 5, we have the ROC curve. Area Under the Curve (AUC) superior than 0.70 can be interpreted as a good fit for the model.

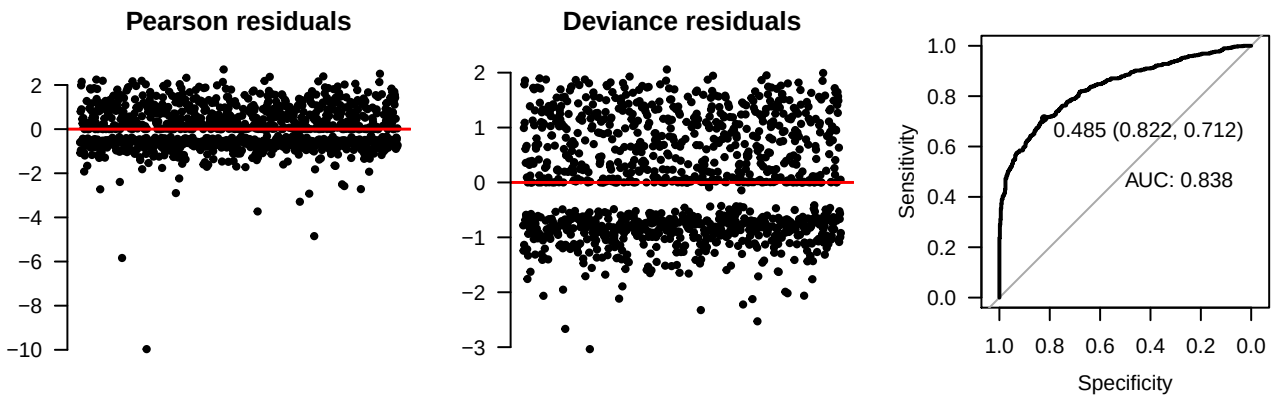

Figure 5. Dispersion of the Pearson and Deviance residuals, in the left and the center, respectively. ROC curve in the right, with the AUC value, cutoff, specificity and sensitivity.

4. Conclusion and next steps

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. With the model fitted the same result is observed. When compared the results of the AM/FM-based classification with the DR signs status, big differences are observed. To see if this difference is statistically significant, a t-test can be performed.

The model fitted with all the variables present 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 statuses in very close to 0.5.

About the features, in general, the features related to the MD results shown to be more significant. Thinking in the next steps of the analysis, a selection of variables can be performed, and a Principal Component Analysis (PCA) can be used to group the MD features in one, and the exudates features in one too. Others link functions can be also tested and a predictive model based on the best model can also be trained.

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] Logistic regression. From Wikipedia, the free encyclopedia. URL: https://goo.gl/ZXY6qE. Accessed at 7 November 2017.
[10] 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/.
[11] Hosmer-Lemeshow test. From Wikipedia, the free encyclopedia. URL: https://goo.gl/8SgkaB. Accessed at 9 November 2017.
[12] JAY, M. (2017). generalhoslem: Goodness of Fit Tests for Logistic Regression Models. R package version 1.3.0. URL: https://goo.gl/7VG9Ke.
[13] Receiver operating characteristic. From Wikipedia, the free encyclopedia. URL: https://goo.gl/ret2fX. Accessed at 9 November 2017.
[14] 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.