MATEMATIKA: MJIAM, 2021, Volume 37, Number 1, 33–43

Robust Logistic Regression for Graduate's Employability from Public Universities in Malaysia

Tengku Salbiah Tengku Mohamed and Muhammad Hisyam Lee

Department of Mathematical Sciences
Faculty of Science
Universiti Teknologi Malaysia
81310 Johor Bharu, Johor, Malaysia
Corresponding author: mhl@utm.my

Article history
Received: 30 October 2018
Received in revised form: 25 April 2021
Accepted: 28 April 2021
Published online: 30 April 2021

Abstract The graduate's employability is studied to investigate its factors involving graduates from public universities in Malaysia for the year 2016. This topic has become a concerning topic in Malaysia as the number of graduates getting hired is lower than expected and does not commensurate with the total number of graduates produced every year. The investigation on graduate's employability based on their profiles is done by utilizing a robust method due to the existence of outliers in the data set. The analysis shows that age, CGPA, discipline, educational stage, English grade, entry qualification, gender, state, sponsor and university types are factors that affect Malaysian public university graduates. Overall, robust logistic model fits the data with correctly classified 69.65% and produces area under the Receiver Operating Characteristics (ROC) curve with 0.6961 in separating the groups of employed and unemployed graduates.

Keywords Robust logistic regression modelling, variable selection methods, tracer study, graduate's employability, Malaysia public universities.

Mathematics Subject Classification 62J12

1 Introduction

Employability can be defined as a set of skills, knowledge and personal criteria that could increase the likelihood of people to get employed in their chosen occupation(s) which give benefits to themselves, workforce, community, and the economy [1]. The topic of employability has been disputed from time to time especially towards graduates either in their academic performance, the ability of tertiary schools in producing more employable graduates and skills that are needed by the industry. The employability among graduates has been highly associated with tertiary education in providing graduates who fulfil basic prerequisites made by employers or labour market demands. However, it is generally known that there is a mismatch between skills provided by universities and requirements set by employers such as soft skills, advancement in industry knowledge and work experience [2-4].

The rise of unemployment among graduates is highly related to the population of education, economic crisis, and technology advancement [5]. In particular, the rise in the number of universities over time and an increase in open applicants have boosted the number of graduates each year. Ultimately, graduates will face redundancy and decrease in demand,leading to increased competition among them in securing their first jobs [6]. Moreover, unemployment among graduates can be caused by individual factors, such as academic background, skills, experience, demography, attitude and aptitude. Based on a previous study, CGPA, gender, study field have been repeatedly found to be significant factors for graduate's employability [6,7]. In addition, it has been found that low CGPA is a factor that deprived many graduates of their career goals other than work experience [8].

This topic has become a concerning topic in Malaysia as the number of graduates in getting hired is lower than the expectation, and thus, not commensurate with the number of graduates produced per year.Statistics show that the rate of unemployment in Malaysia has been increasing from the year 2014 until 2016 which were 2.9%, 3.1% and 3.5% respectively. In addition, Human Resource Minister Datuk Seri Richard said that over 200,000 out of 500,000 citizens who were jobless in 2016 were unemployed graduates [9]. A tracer study involving investigation on graduates' employability, is a primary step in measuring the evenness between higher education and industry demand and this may help narrow the gap between them. Tracer study is a follow-up study involving former participants of education programmes. It is a backward look investigation in identifying the impact that has been achieved on students who were given a specific training [10]. This study is crucial especially for universities to discover their programme effectiveness, trace graduates' employability and share information about job opportunities [11,6].

Mmab Modise study indicates that non-traditional research model career workshop is another strategy to reduce mismatch between labour market needs and graduate competencies by assisting students to discover their capability, give inspiration and widen their horizon in holistic approach [12]. In South Africa, a tracer study was done by Regan which depicts that education history, demographics or socio-economic status of students do affect graduate's employability. Besides, policy makers should give attention to institutions that are inadequate in money sources and reduce the gap between graduates and employers by overcoming the supply-side problems [13]. Furthermore, universities and firms should collaborate, communicate well in terms of skills that may be required by the industries. Thus, universities are more aware and will nurture the skills through curriculums or extra programmes [8].

Therefore, in order to determine the factors that affect graduate's employability in Malaysia, such a tracer study data from the Ministry of Education is a genuine and best available data to be analysed, investigating the graduates' employability factors to represent all graduates in Malaysia. Meanwhile, logistic regression is a preferred method in this study as it can serve both probabilistic outcomes along with classifying graduate's employability. Previously there were studies related to graduate's employability using logistic regression method. For instance, a study on predicting factors affecting employability among IT graduates [7], job attainment for Bachelor in Australia [14] and factors affecting employability of people with epilepsy [15]. Moreover, in India logistic modelling was applied to give more understanding on this matter in psychological view [16].

However, the existence of outlier in tracer study data had caused classical logistic regression method no longer suitable. An outlier is a data point that does not follow the trend of the rest of the data [10]. It may occur based on several reasons such as human errors in entering data, natural disaster, fault in machinery or other related reasons that change the norm of the population [17]. It is well known when applying classical method, outliers need to be removed. This is because the classical logistic regression method is not resistant towards outlier [18] and the maximum likelihood estimation could break down into zero [19]. Thus, it loses its stability and sensitivity [20] which cause results to be produced inaccurate due to bias in estimation and significant test statistics. On the other hand, simply removing outliers could give an effect as it downsizes the number of observations, change actual data distribution and might lose some information as data were removed [21]. Therefore, an advance method is required to model the data tracer study and determine factors that are affecting graduate's employability.

Robust regression method is a method used when handling influential outliers. The basic idea is that this method induces a bounded function for the estimator to be robust by losing some of the efficiency and known as Huber M-estimator [22]. Consequently, it minimizes gross error sensitivity and breaks down point as it could influence estimator to be robust towards outliers [21]. Therefore, this method will be used in logistic regression method so that both accuracy in estimating and determining the significant factors that affect graduate's employability can be improved concerning the outlier observations in the data set.

2 Methodology

2.1 Binary Logistic Regression Model

Logistic regression modelling is used to investigate the relationship between dependent variable and independent variables where the dependent variable is dichotomous [23]. Thus, binary logistic regression with the outcome; Y could be either employed or unemployed follows binomial distribution and X is factor that influences graduate's employability can be written as follows

$$Pr\left(Y_i = 1|X_i\right) = \frac{exp\left(X_i^T\beta\right)}{1 + exp\left(X_i^T\beta\right)} \tag{1}$$

where $\beta' = \left(\beta_0, \beta_1, \beta_2, \cdots, \beta_p\right)^T$ is $p \times 1$ vector of parameters to be estimated using the maximum likelihood estimator and shows the relationship between dependent and explanatory variables. The individual testing for coefficients either significantly associated with outcome variable in logistic model, will have the hypothesis testing as follows

$$H_0 : \beta_i = 0$$

$$H_1 : \beta \neq 0$$

with Wald test statistics is derived by

$$W = \left(\frac{\hat{\beta}}{\hat{se}\hat{\beta}}\right)^2 \sim N\left(0, 1\right) \tag{2}$$

where $\hat{se}$ is the standard error of coefficient estimates. The variable is statistically significant if the value of W is more than 2.

2.2 Robust Logistic Regression Model

The limitation of maximum likelihood estimator in the logistic regression framework has led to the use of robust method to overcome the issue in handling outliers from the data set. The likelihood equation of robust has given some extension for the generalized linear model by using Mallows-type estimator and the scale parameter as

$$\boldsymbol{\psi_k(x) = max\left(-k, min(k, x)\right)} \tag{3}$$

where k is constant value showing the loss of efficiency. The estimating equation then is written as

$$g\left(\beta; y\right) = \sum_{i=1}^{n} w\left(x_i \frac{\left\{\psi_k\left(r_i\right) - \alpha\left(\mu_i\right)\right\}}{V_i\left(\mu_i\right)^{\frac{1}{2}}} \frac{\vartheta \mu_i^T}{\vartheta \beta}\right) = 0 \tag{4}$$

where ψ_k is a Huber function with $V\left(\mu_i\right)$ is the variance of function and $r_i = \left(y_i - \mu_i\right)/\sqrt{V}_i$ is a Pearson residual. The equation of $g\left(\beta; y\right) = 0$ under the theory of M-estimation, has a bounded function as $x\left(x_i\right)$ bounding the outlying value in covariates while the response value was bounded by tuning constant k. Thus, in order to be implemented into the logistic regression of in Equation (4), it can be considered as

$$\mu_i = F\left(\mathbf{X}_i^T \boldsymbol{\beta}\right) \tag{5}$$

with

$$F\left(\mu\right) = \frac{exp\left(\mu\right)}{1 + exp\left(\mu\right)} \tag{6}$$

$$V_i\left(\mu_i\right) = \mu_i\left(1 - \mu_i\right) = V_i \tag{7}$$

and

$$\alpha\left(\mu_i\right) = \psi_k\left(\frac{1 - \mu_i}{\sqrt{V}_i}\right)\mu_i + \psi_k\left(-\frac{\mu_i}{\sqrt{V}_i}\right)\left(1 - \mu_i\right) \tag{8}$$

2.3 Cook's Distance

The Cook's Distance is one of the methods in detecting influential outliers by measuring how residuals would change if data are removed. The Cook's distance method was proposed by Cook [24]. It is calculated by the following equation

$$D_i = \frac{\sum\limits_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{ps^2} \tag{9}$$

Observations are considered influential if and only if its Cook's distance, D_i is greater than a certain threshold in equation (9)

$$D_i > \frac{4}{n - k - 1} \tag{10}$$

where k is the number of variables and n is number of observations

2.4 Checking Model Accuracy

2.4.1 Classification accuracy

Classifying the predicted response variable was carried out to measure the sensitivity and specificity prediction of robust logistic model as well as classification performance. In this approach, instead of predicting the probability of graduate's employability, it estimates in binary way either graduates are employed $Y = 1$ or unemployed, $Y = 0$. In order to achieve a derived outcome presented in dichotomous, a cut point, c is used and compared with the estimated probability whereby those predicted probability exceeding c, will fall into group 1; otherwise it equals to 0 and in this case, cut point $c = 0.5$ is used.

2.4.2 Area under ROC

A classification accuracy for the robust logistic regression model is investigated further by using area under the Receiver Operating Characteristics (ROC) curve using signal detection theory instead of depending on a single cut point to classify the test result. The area of ROC curve is obtained by detecting the positive or negative signal in the presence of noise for an entire range of possible cut points. Table 1 shows the rule of thumb for ROC ranging between 0.5 up to 1.0 to evaluate how well the robust logistic regression model in discriminating graduate's status either employed or unemployed.

Table 1: The Rule of Thumb for Area under ROC

ROC Range	Rule of Thumb
$ROC = 0.5$	No discrimination
$0.5 < ROC < 0.7$	Poor discrimination
$0.7 < ROC < 0.8$	Acceptable discrimination
$0.8 < ROC < 0.9$	Excellent discrimination
$ROC \geq 0.9$	Outstanding discrimination

3 Result Analysis

3.1 Data

A data tracer study for the year 2016 obtained from the Ministry of Higher Education was used in this study. The tracer data includes 55,017 graduates selected from Malaysian public universities consisting of eleven factors have been considered. As part of data preprocessing, several variables have been recoded and revalued. The 20 public universities are divided into three types of universities: research university, comprehensive and specialised-based universities. In addition, the variable sponsorship is divided into two categories which are government and private. The entry qualification with 39 levels were compressed into 8 different levels of qualifications. Meanwhile for CGPA, a range was set consisting of four groups: 3.67 – 4.00, 3.33 – 3.66, 2.50 – 3.32 and 2.00 – 2.40 for groups A, B, C and D respectively. Table 1 shows the data description.

Table 2: Relation of Variables Presented in Tracer Study Data

Variable	Nature	Description
Status	Binary	0: Unemployed 1: Employed
Age	Numeric	In years
CGPA	Ordinal	4 levels
Discipline	Categorical	5 fields
Education Stage	Ordinal	7 levels
English Grade	Ordinal	4 grade levels
Entry Qualification	Ordinal	8 qualifications
Gender	Binary	Male / Female
Malay Grade	Ordinal	4 grade levels
Sponsor	Binary	Government/ Private
States	Categorical	17 states
University	Categorical	3 university types

An inspection of the summary statistics from the data illustrates that there are outliers in the data tracer study as the Cook's distance shows a cut-off value of 9.097732e-05. This implies that 951 observations are detected as influential observations which could give effect on the regression coefficients and the regression line. Figure 1 is the Cook's distance plot and those peaks shown are outliers in tracer study data that exceed the cut-off value.

3.2 Empirical Result

Before building a robust model for the graduates employability, certain reference categories for some non-numeric variables were determined as they will make a difference in evaluating the coefficients for individual predictors. For variables with ordinal data types such as CGPA, education stage, entry qualification, English and Malay grades, the reference groups were selected by choosing their lowest ranks. Whereas Sabah, government and comprehensive are reference groups for state, sponsor and university types respectively while the remaining variables were selected by default.

A robust logistic regression equation with factors affecting graduates' employability is shown in Equation (11). The model shows the relationship at each factor towards employed graduates. The results of the model are reported in Table 3. Robust logistic model for graduate's employability from public universities in Malaysia was developed without removing the outlier from the data set, taking into consideration of all the possible outcomes of graduates' status according to their profiles. Based on the robust model, age, CGPA, discipline, educational stage, English grade, entry qualification, gender, state, sponsor and university types are factors that affect the employability of graduates at 0.05 significant level and only Malay grade is not significant.

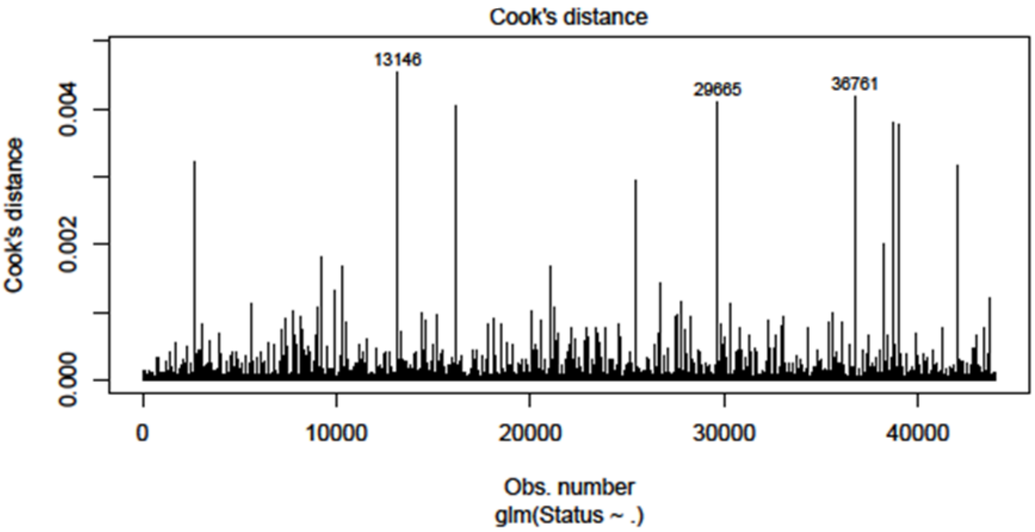

Figure 1: Cook's Distance Plot for Tracer Study Data

$$\begin{aligned}\ln(\mathit{graduates\ employed}) = {} & -4.210 + 0.177(Age) + 0.614(CGPA_A) + 0.517(CGPA_B) + \\ & 0.254(CGPA_C) - 0.181(Science) + 0.103(Techniue) + 0.344(ICT) - \\ & 0.528(Education) + 0.731(Adv.\,Diploma) + 4.163(Ph.D) + 2.276(Degree) + \\ & 2.617(Certificate) + 0.447(Distinction) + 0.357(Pass\,with\,Credit) - \\ & 0.424(STPM) - 0.357(Foundation) + 0.896(Master) - 0.221(Female) + \\ & 0.348(Private) + 1.033(Johor) + 0.679(Kedah) + 0.283(Kelantan) + \\ & 0.993(Melaka) + 1.039(N.Sembilan) + 0.751(Pahang) + 1.219(P.Pinang) + \\ & 0.723(Perak) + 0.472(Perlis) + 1.387(Selangor) + 0.486(Terengganu) + \\ & +0.346(Sarawak) + 1.450(WPKL) + 1.572(WPP) - 0.258(Research) - \\ & 0.160(Specialized) \end{aligned} \tag{11}$$

Based on the model, it shows the age of graduates has a positive impact on graduate status as an older graduate tends to have high possibility in getting a job than the younger ones. While CGPA, those who excel in academic are more likely to get employed with group A 1.848 is more likely followed by those in group B and C with 1.677 and 1.289 respectively. Choosing a study discipline is also crucial to secure graduate's employability with Literature & Social Sciences as group reference, graduates in Information Technology field has higher demand in the industry with odd ratio 1.108 and 1.411 for Technique compared to other fields. Meanwhile, the English language has proven its significance in communicating as the result shows those who excel in English grades are more likely to get employed compared to those who are not.

Moreover, education stage and entry qualification are also significant towards graduate's employability. However, it shows that a higher education stage does not increase the likelihood for graduates to get employed. Similarly, for entry qualification, STPM and Foundation are less likely to get employed than SPM except graduates with Master entry qualification. On the other hand, gender appears to be significant and reveals that male graduates are more likely to secure their first jobs compared to female graduates. In addition, graduates who are sponsored by the private sectors are more likely to get employed than those who are sponsored by the government with odd ratio 1.416. Furthermore, state does affect graduate's employability in Malaysia showing those living in urban areas are more likely to secure their first jobs. It can be seen that graduates in WPP have the highest possibility to get the jobs followed by WPKL, Selangor, P. Pinang, and Johor. Apart from that, university types are one of the significant factors for graduate's employability. Nevertheless, it shows that both research and specialized universities are less likely to get employed compared to comprehensive university. Table 3 presents the factors that are affecting graduate's employability for robust logistics model.

Table 3: Robust Logistic Regression Results for Graduate's Employability from Public Universities in Malaysia

Factors	Wald	Odd Ratio	Sig.
Constant		0.014	0.000
Age	534.06	1.194	0.000
CGPA	53.60		
D			
A		1.848	0.000
B		1.677	0.000
C		1.289	0.000
Discipline	55.74		
Literature & Social Science			
Science		0.835	0.000
Technique		1.108	0.001
ICT		1.411	0.000
Education		0.589	0.000
Educational Stage	8.06		
Diploma			
Post Graduate Diploma		3.077	0.039
Degree		0.749	0.030
Master		0.190	0.000
Ph.D.		0.103	0.000
English Grade	18.71		
Fail			
Distinction		1.564	0.009
Pass with Credit		1.429	0.035
Entry Qualification	23.43		
SPM			
STPM		0.654	0.002
Foundation		0.700	0.012
Master		2.450	0.036

Gender	81.38		
Male			
Female		0.802	0.000
State	102.30		
Johor		2.808	0.000
Kedah		1.973	0.000
Kelantan		1.328	0.000
Melaka		2.699	0.000
Negeri Sembilan		2.827	0.000
Pahang		2.120	0.000
P. Pinang		3.383	0.000
Perak		2.061	0.000
Perlis		1.603	0.000
Selangor		4.002	0.000
Terengganu		1.625	0.000
Sarawak		1.413	0.000
Wilayah Persekutuan Kuala Lumpur (WPKL)		4.264	0.000
Wilayah Persekutuan Putrajaya (WPP)		4.819	0.000
Sponsorship	9.494		
Government			
Private		1.416	0.002
University Types	33.464		
Comprehensive			
Research		0.772	0.000
Specialized		0.852	0.000

The model performance was then evaluated, and it shows based on all ten significant factors in the robust model, it can correctly classify 69.65% while area under ROC gives 0.6961 in classifying the graduate's employability status. This shows that robust logistic regression with M estimates is poor in classifying the graduates' status as it is quite low for both indicators but almost acceptable discrimination for ROC value.

4 Conclusion

In order to determine the factors affecting graduate's employability, a tracer study is a suitable approach to investigate the graduate employment status. Therefore, to achieve the study objective, the most recent data of tracer study for the year 2016 has been used. However, due to the existence of the outlier that could negatively affect a regression model, a robust method was applied to achieve the study objective. It has been found that age, CGPA, discipline, educational stage, English grade, entry qualification, gender, state, sponsor and university types are factors affecting graduate's employability from the public universities in Malaysia. Moreover, based on the model performance for classification, other factors need to be included for future studies such as graduate skills, economic growth and technology advancement to increase its prediction correctness and accuracy in estimation. Even though the developed model shows quite a poor ROC, the values are quite close to 0.7 which is fair in separating the graduate's status. In addition, another approach should be applied in the future to improve the study results.

Acknowledgement

The authors would like to express appreciation to the Ministry of Education for providing a data source to complete the research study and partially funded through the Research University Grant for University of Technology, Malaysia (grant number 15H64) and also to UTM Zamalah scholarship for their encouragement and support to participate in this activity.

References

[1] Yorke, M. *Employability in Higher Education: What It Is - What It Is Not.* Heslington, York, United Kingdom: The Higher Education Academy United Kingdom. 2006.

[2] Verma, P., Nankervis, A., Priyono, S., Moh, N., Connell, J., & Connell, J. Graduate work-readiness challenges in the Asia-Pacific region and the role of HRM. *Equality, Diversity and Inclusion: An International Journal*. 2018. 37(2): 12–137.

[3] Ang, M. C. H. Graduate employability awareness: a gendered perspective. *Procedia - Social and Behavioral Sciences*. 2015. 211(September): 19–198.

[4] Suleman, F. The employability skills of higher education graduates?: Insights into conceptual frameworks and methodological options, *High Education.* 2018. 76: 263–278.

[5] Malec, L., & Kiráová, A. Evaluating competencies of graduates in Tourism as a prerequisite for future employability,*Prague Economic Papers.* 2018. 27(2): 19–214.

[6] Sapaat, M. A., Mustapha, A., Ahmad, J., Chamili, K., & Muhamad, R. A classification-based graduates employability model for tracer study by MOHE. *Communications in Computer and Information Science*. 2011. *188 CCIS*(PART 1): 27–287.

[7] Piad, K. C., & Ballera, M. A. Predicting IT Employability Using Data Mining Techniques, *Third International Conference on Digital Information Processing, Data Mining, and Wireless Communication.* 2016. 2–30

[8] Yusof, N., & Jamaluddin, Z. Graduate employability and preparedness: A case study of University of Malaysia Perlis (UNIMAP), Malaysia. *Malaysian Journal of Society and Space*. 2015. 11(11): 12–143.

[9] Carvalho, M., Sivanandam, H., Rahim, R., & Yunus, A. 500.000 currently jobless-Nation: The Star Online. 2017, March 23. Retrieved May 6, 2019, from https://www.thestar.com.my/news/nation/2017/03/23/500000-currenly-jobless-riot-number-comsidered-low-going-by-international-benchmarks

[10] Zhang, C. X., Ji, N. N., & Wang, G. W. Randomizing outputs to increase variable selection accuracy. *Neurocomputing.* 2016. 21(8): 9–102.

[11] Noko, P., & Ngulube, P. A vital feedback loop in educating and training archival professionals: a tracer study of records and archives management graduates in Zimbabwe. *Information Development.* 2013. 31(3): 270–283.

[12] Mmab Modise, O. Career workshops as a non-traditional research model for enhanced relationships between higher education and the labour market. *International Journal of Training and Development*. 2016. 20(2): 15–163.

[13] Regan, M., & Reynolds, J. Schooling inequality , higher education and the labour marke?: Evidence from a graduate tracer study in the Eastern Cape , South Africa. *Labour Market Intelligence Partnership (LMIP) Project*. (2015). 363(7 June).

[14] Jackson, D. Factors influencing job attainment in recent Bachelor graduate?: Evidence from Australia. 2014. *High Education*. 13–153.

[15] Chen, M., Wo, M., Seang, K., Yuen, W., & Tin, C. Factors affecting the employability in people with epilepsy. *Epilepsy Research*. 2016. 128: 6–11.

[16] Pandit, S. A., G., P., Wallack, D. C., & Vijayalakshmi, C. Towards understanding employability in the Indian context: A preliminary study. *Psychology and Developing Societies*. 2015. 27(1): 81-103

[17] Ohyver, M., Moniaga, J. V., Yunidwi, K. R., & Setiawan, M. I. Logistic regression and growth charts to determine children nutritional and stunting status: A review. *Procedia Computer Science*. 2017. 11(6): 232–241.

[18] Pregibon, D. Resistant fits for some commonly used logistic models with medical aplications. *Biometric*s. 1982. 38: 486-498.

[19] Valdora, M., & Yohai, V. J. Robust estimators for generalized linear models. *Journal of Statistical Planning and Inference*. 2014. 14(6): 3–48.

[20] Acua, E., & Rodriguez, C. On detection of outliers and their effect in supervised classification. *Department of Mathematics, University of Puerto Rico,Mayaguez*. 2014. (June).

[21] Bellio, R., & Ventura, L. An Introduction to Robust Estimation with R Functions. *Research Gate*. 2005. 1–57

[22] Huber, P. J. *Robust Statistics*. Wiley. 1981

[23] Lopes, M. B., Verssimo, A., Carrasquinha, E., Casimiro, S., Beerenwinkel, N., & Vinga, S. Ensemble outlier detection and gene selection in triple-negative breast cancer data. 2018. *BMC Bioinformatics*, 19(1): 168.

[24] Cook, R.D. Detection of influential observation in linear regression. *Technometrics*. 1977. (19): 15-18.