Logistic regression based next-day rain prediction model

Ejike, Ogochukwu; Ndzi, David L.; Al-Hassani, Abdul-Hadi

Published in:
2021 International Conference on Communication & Information Technology (ICICT)

DOI:
10.1109/ICICT52195.2021.9568483

Published: 26/10/2021

Document Version
Peer reviewed version

Ejike, O., Ndzi, D. L., & Al-Hassani, A-H. (2021). Logistic regression based next-day rain prediction model. In *2021 International Conference on Communication & Information Technology (ICICT)* IEEE. https://doi.org/10.1109/ICICT52195.2021.9568483

Logistic Regression based Next-Day Rain Prediction Model

[1]Ejike Ogochukwu, [1]David L. Ndzi and [2]Abdul-Hadi Al-Hassani

[1]School of Computing, Engineering and Physical Sciences, University of the West of Scotland, Paisley, PA1 2BE, UK.

[2]Iraq University College, Somer Building, Al-Estiqlal St.,Basrah, Iraq

E-mail: B00386504@studentmail.uws.ac.uk, david.ndzi@uws.ac.uk,

hadi.alhassani@iuc.edu.iq

ABSTRACT

Rain prediction is challenging due to the complex nonlinear combination of atmospheric factors. This paper presents the application of logistic regression modelling to predict rain the next day using weather parameters from the previous days. One year of weather data (temperature, pressure, humidity, sunshine, evaporation, cloud cover, wind direction, and wind speed) from Canberra, Australia has been used to develop the logistic regression-based model. Akaike Information Criterion (AIC) Backward, Baysian Information Criterion (BIC) Stepwise, and Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression models have been developed based on input variable selection and prediction. These models are evaluated using Area Under the ROC Curve(AUC) and Hosmer- Lemeshow test to determine the models' adequacies and accuracies to predict rainfall occurrence the next day. The likelihood of rainfall the next day has been interpreted based on the calculated odds ratios with 95% confidence intervals of the selected independent weather parameters. The result showed that the rainfall the next day can be predicted using logistic regression (AIC Backward) with 87% accuracy, provided that the appropriate weather parameters are chosen.

Keywords:-

Rain, rainfall, prediction, logistic regression

1. INTRODUCTION

Climate, weather, and rainfall are non-linear and complex phenomena, which require detailed modelling to obtain accurate predictions. The evolution of climate due to the complex interaction of different factors have resulted in high incidences of severe weather events such as extended droughts, severe floodings, and storms. Rainfall exhibits wide-scale variations in both space and time. It is a stochastic process, whose occurrence depends on the antecedent of other parameters such as temperature, atmospheric pressure, wind, humidity and other atmospheric parameters that require consistent and relevant meteorological and environmental data to predict. Notwithstanding the usefulness of rainfall, it can also be cataclysmic; causing natural disasters like floods and landslides. Therefore the forecasting of extreme weather events is necessary due to the emerging climate change and possible adverse effects on humans[1]. Globally, many studies have been carried out on rainfall. In [2], the trend analysis of rainfall over Jordan using three neighboring locations covering 81 years (1922-2003) has been studied. Researchers in [3] have studied the synoptic

regimes associated with rain and no-rain days in south-eastern Queensland (Australia) and other studies have identified the influence of the El Niño on climates and especially on its rainfall [4] - [7].

2. RAIN PREDICTION ANALYSIS

Weather forecasting is challenging regardless of the amount of available data. There are several rainfall-forecast methods used in weather prediction. Recent studies have developed rainfall prediction using different weather and climate forecasting techniques [8][9]. These methods can be grouped into empirical and dynamical techniques. In the empirical techniques, the historical rainfall data and its relationship to a variety of atmospheric and oceanic variables is analysed. The most common empirical approaches used for climate prediction are Artificial Neural Network (ANN), Fuzzy Logic (FL), and Machine Learning (ML). The dynamical techniques make predictions based on physical models built on systems of equations that calculate the evolution of the global climate system in response to initial atmospheric conditions [10].

2.1 LOGISTIC REGRESSION

Machine Learning is broadly grouped into supervised and unsupervised machine learning. In supervised machine learning, the prediction is made by training the model based on known output values, while unsupervised machine learning does not have a known set of output values. The supervised learning is further divided into regression, where the model is trained to predict results based on the relationships with the input variables, and classification, where the model is trained to recognise and predict categories.

Logistic Regression is a probabilistic binary classifier from the binomial family of generalised linear models (GLMs), that provides probabilities and categorises new data using different types of datasets It calculates the output of a categorical dependent variable with a mixture of continuous and categorical variables as the independent variables. The logistic regression is used when the probabilities between two classes are required, such as whether it will rain tomorrow or not, is either 1 or 0, true or false, etc. as is the case in this study. Using logistic regression to predict rain occurrence and GLM to predict rainfall volume, [11] demonstrated how to project future rainfall based on future climate scenarios.

The logistic regression aims to ascertain the relationship between the probability that it will rain the next day, p, which is binomially distributed with the covariates. For n independent observations $y_1, y_2, \ldots\ldots, y_n$, y_i(which is the i^{th} observation) is a realisation of the random variable Y_i. If Y_i is binomially distributed i.e.,

$$Y_i \sim \text{Bin}(n_i, p_i) \qquad (1)$$

and Y_i are independent variables, so for $E(Y_i) = p_i$, applying the equation of generalised linear model, $g(E(Y_i))$ is

$$g(p_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_j x_{j,i} \qquad (2)$$

where $\mathbf{x}_i$ is a vector of input variables (weather paremeters), $\boldsymbol{\beta}$ is a vector of regression coefficients and g(.) is the link function. Since p_i lie between 0 and 1, the logit link function is defined as

$$\text{logit}(p_i)=\log\left(\frac{p_i}{1-p_i}\right) = \beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\ldots+\beta_j x_{j,i} \quad (3)$$

which is the logarithm of odds, the ratio of the probability that rain will fall the next day (p_i) over the probability that rain will not fall the next day ($1 - p_i$). The odds that it will rain the next day is written as

$$\left(\frac{p_i}{1-p_i}\right)=\exp(\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\ldots+\beta_j x_{j,i}) \quad (4)$$

This shows the relationship between the covariates and the probability of the response. Therefore,

$$p_i=\frac{\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_j x_{ij})}{1+\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_j x_{ij})} \quad (5)$$

2.2 MODELLING AND VARIABLE SELECTION

The data is made up of 12 variables (weather parameters); MinTemp, Evaporation, Sunshine, WindGustSpeed, WindGustDirection, WindSpeed3pm, Humidity9am, WindSpeed9am, Humidity3pm, Pressure3pm, Cloud9am, Cloud3pm, and a binary output variable RainTomorrorw. The dataset is divided into 70% training and 30% test data sets and used to find the best model that can predict rain the next day with the least number of input variables (weather parameters).

A Full model is built with all the 12 weather variables as input. k-fold cross-validation (CV) which, is a resampling procedure that splits the training data into subsets, in this case k= 10, is used one set at a time to train the model on the remaining 9 subsets and tests it. This helps to tune the full model. WindGustSpeed and Pressure3pm were found to be the significant variables in the full logistic regression model.

Least Absolute Shrinkage and Selection Operator (LASSO) regression [12] is a model that performs L1 regularisation by reducing the model complexity and dimensionality by removing less important variables. The best value of the tuning parameter, using the k-fold cross-validation process on the training set after standardizing the input variables that minimises the deviance is determined. The optimal tuning parameter value is 0.039 giving a final model with 5 input variables; Sunshine, WindGustSpeed, Humidity3pm, Pressure3pm, and Cloud3pm.

Akaike Information Criterion (AIC) backward elimination model is built using an automated backward model selection procedure based on the Akaike Information Criterion [13]. AIC measures how well a model fits the data by calculating the information lost. Backward elimination method builds the model by removing input variables in a stepwise process based on a chosen information criterion. This starts with all 16 weather parameters, and using the step process, removes the insignificant variables based on the AIC

value obtained in each step. The model with the lowest AIC consists of 6 input variables; Sunshine, WindGustSpeed, WindSpeed3pm, Humidity9am, Pressure3pm, and Cloud3pm.

The Bayesian Information Criterion (BIC) [14] is a measure of the fit that is closely related to the AIC. Using the automated stepwise (default) model process, the less significant variables are eliminated based on the BIC value obtained in each step. The model with the lowest BIC is chosen as the final model. This final model has 2 input variables, Sunshine and Pressure3pm, resulting in a model with the least number of input variables.

2.3 MODEL VALIDATION

Hosmer- Lemeshow [15] test which, measures the goodness of fit of a model is used.

Model	Chi-Square Statistics	Degree of Freedom	P-value
FULL	21.981	8	0.005
LASSO	11.046	8	0.199
AIC	9.816	8	0.278
BIC	3.113	8	0.927

Table 1: Hosmer- Lemeshow Goodness-of-Fit Test statistics of the 4 Prediction Models

Table 1 shows the goodness of fit test for the 4 models. The p-values for the LASSO, AIC Backward and BIC are all greater than 0.05, so the null hypothesis is not rejected and confirms that these models are a good fit for the data. The Full model is not a good fit for the data as its p-value is 0.005.

Confusion matrix is used to measure the performance of a classification model. It summarises the performance of the models to accurately predict whether it will rain the next day or not. In Table 2, the BIC and AIC models have the highest accuracy of 87.6%. The Full model has the highest score for classifier exactness. For F1 score, the Full model and AIC model both have the highest F1scores. Since the dataset is imbalanced with approximately 18% of the data showing rain and the remaining 82% when it was not raining, the balance accuracy metric is applied, with the AIC model having the highest balanced accuracy of next day rain prediction of 78.97%.

Table 2: Error metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 score (%)	Balanced_ Accuracy (%)	AUC (%)
FULL	86.67	73.70	60.90	93.90	67.00	77.39	92.78
LASSO	85.71	42.10	66.70	88.17	52.00	77.42	93.39
AIC	87.62	68.40	65.00	92.94	67.00	78.97	93.64
BIC	84.30	40.00	57.10	88.08	47.00	72.61	91.62

Area Under the Receiver Operating Characteristics (ROC) Curve (AUC) [16] is a measure of the total area below the ROC curve. This is used to evaluate model fit and compare the performance of classification models. The ROC Curve [17] is a graphical representation of the performance of a classification model at all classification thresholds.

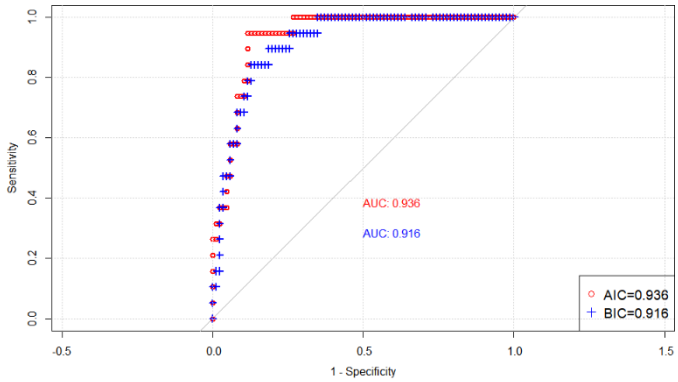

Figure 1: Receiver Operating Curves of the prediction models with the highest and lowest areas

Figure 1 is a plot of the ROC curves of the forecast models with the highest (AIC) and lowest (BIC) AUCs. The AIC model's AUC has the greatest area under the curve of approximately 93.6%, and the BIC model has the lowest with 91.6%. This implies that the AIC model offers the best prediction of rainfall the next-day.

2.4 MODEL ASSUMPTIONS

Logistic regression has some assumptions in common with linear regression. The logistic regression assumptions are tested on the AIC Backward model to ascertain that there is no violation of assumptions and that the reported results are correct and the model is accurately interpreted.

Logistic regression requires the output variable to be nominal or ordinal. For the AIC Backward model, the response variable RainTomorrow is a binary variable of rainfall the next day or not.

For a model with a binary outcome, there is a general requirement of a minimum of 10 EPP (events per variable) means cases with the least frequent outcome for each independent variable in the model [18]. The dataset has 353 samples and exceeds the minimum required sample size.

The condition that the errors associated with one variable are not correlated with the errors of any other data variables or samples and each variable needs to be independent of one another. Durbin Watson (DW) test [19], which is based on autocorrelation is applied to check if the assumption holds. From the test carried out on the data, the Durbin Watson (DW) Statistic d =1. 85 indicates a positive autocorrelation. Using the significance level of 0.05 for DW test, p-value of 0.218 is statistically significant, therefore the null hypothesis is not rejected and it is concluded that the residuals in the AIC Backward model are not autocorrelated.

No outlier should significantly influence the results. Outliers can be identified using Cook's distance [20]. The higher the value of the Cook's distance, the more influential the observation.

In Figure 2, the circle size is proportional to the Cook's distance. The data sample 306 (top left-hand corner) appears to have a combination of characteristics with the largest studentised residual and Cook's distance. The change in coefficient by removing the outliers is minimal on the coefficients of the AIC Backward model. Therefore, the assumption holds.

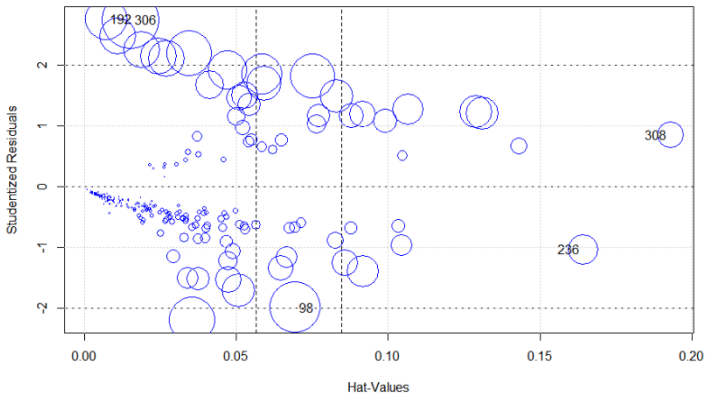

Figure 2: Influential Outliers Plot for AIC model

Logistic regression requires the independent variables to be linearly related to the respective logit response but does not need the output and input variables to be linearly related. The Box-Tidwell transformation test [21] requires that all terms should not be statistically significant for the assumption of linearity to hold. The p-values of all the covariates (weather variables) in the AIC Backward model range from 0.10 to 0.97, which are greater than 0.05. It can be concluded that the variables are linear to their log-odds.

There should be an absence or only moderate multicollinearity between independent variables [22]. The assumption holds if the input variables are not correlated. Pearson correlation coefficient is used to calculate the correlation matrix between the variables. Multicollinearity exists between two variables if their correlation coefficient is $\geq |0.8|$. Figure 3 shows that there are no strongly correlated variables in the AIC model and hence the assumption of no multicollinearity applies. Figure 3 shows that there are no strongly correlated variables in the AIC model and hence the assumption of no multicollinearity applies.

3. INTERPRETING THE INDIVIDUAL VARIABLES

From Equation (3), the formula for the AIC Backward model is given as

$$\text{logit}(p_i) = 174.46 - 0.201*\text{Sunshine}_i + 0.069*\text{WindGustSpeed}_i - 0.058*\text{WindSpeed3pm}_i + 0.035*\text{Humidity9am}_i - 0.177*\text{Pressure3pm}_i + 0.182*\text{Cloud3pm}_i \quad (6)$$

In Equation(6), each coefficient is the expected change in the log odds of rain falling the next day, given a unit increase in the respective covariates.

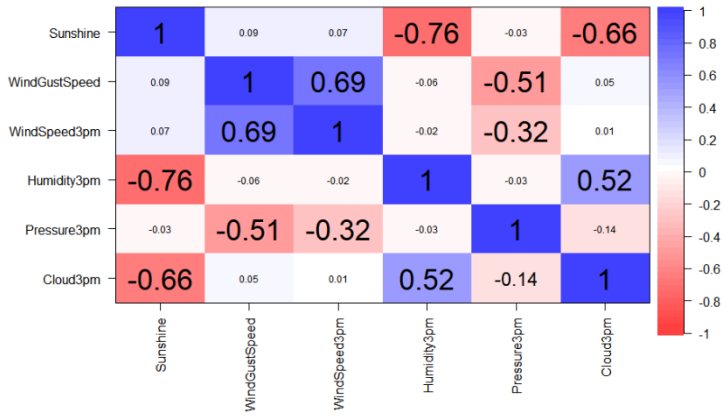

Figure 3: Correlation Plot of the variables in the AIC model

The odds ratio (OR) is used to demonstrate the direction and strength of the relationship between the weather parameters (input variables) and the outcome prediction of rain or no rain the next day. For the OR >1 the odds are increased and for OR<1 the odds are decreased. The confidence interval (CI) is the measure of the level of uncertainty in the odds, as it determines the statistical significance of the input variable.

From Table 3, for one-unit increase in WindSpeed3pm measurement, the likelihood of rain falling the next day decreases by approximately 5.6% (1-0.944 = 0.056), provided that all other variables in the model are held constant. A one-unit increase in Humidity9am gives 3.5% (1.035-1=0.035) increase and for Cloud3pm a 20% (1.2-1=0.2) increase in the odds of rain falling the next day. The variables WindSpeed3pm, Humidity9am, and Cloud3pm are not significant as their 95% confidence interval spans across 1. This is reaffirmed as their p-values are greater than 0.05. Also, the odds of rain falling the next day decreases by 18.2% and 16.2% and increases by 7.1%, and the level of confidence that the true odds lies between 0.69 - 0.96, 0.77- 0.91, and 1.02 – 1.13 for Sunshine, Pressure3pm and WindGustSpeed, respectively, as these variables are statistically significant.

Table 2: Coefficient, Odds Ratio, 95% CI and p-value for AIC Backward model

Covariates (Input_Variable)	Coefficient (β)	Significance (p-value)	Odds (Exp(β))	95% CI for Exp(β)	
				Lower	Upper
Sunshine	-0.201	0.014	0.818	0.697	0.961
WindGustSpeed	0.069	0.008	1.071	1.018	1.127
WindSpeed3pm	-0.058	0.085	0.944	0.884	1.008
Humidity9am	0.035	0.074	1.035	0.997	1.076
Pressure3pm	-0.177	< 0.001	0.838	0.770	0.911
Cloud3pm	0.182	0.108	1.200	0.961	1.498

In general, the AIC Backward model predicted 95% of the days when there was no rainfall the next day correctly and correctly predicted 65% of days when there was next day rainfall, hence having a prediction accuracy of 87.6%, this is also known as the overall correct prediction rate percentage.

4. CONCLUSIONS

Weather prediction is important for many sectors which include agriculture, telecommunications, and environmental agencies. This paper presents logistic regression-based models prediction of next-day rainfall occurrence using weather parameters that can be measured low-cost instruments. The use of low-cost instruments, couple with accurate prediction techniques, is key to the provision of localised high time-space resolution weather forecasting system compared to rain radar and satellite systems. Logistic regression model analysis has been applied to predict rainfall the next day by selecting the appropriate input variables (weather parameters), choosing suitable model building techniques, and validating the best fit model. whilst confirming that relevant assumptions were met and, finally interpreting the results.

Results show that a logistic regression model can be used in predicting next day rainfall occurrence. The AIC Backward model outperforms the BIC Stepwise, LASSO, and the Full models in the discrimination analysis with an accuracy of 87.6%. It has been shown that the weather parameters that are important to predict rainfall the next day are sunshine, wind speed at 3 pm, and atmospheric pressure at 3 pm, which when any of these increases, it decreases the likelihood of rainfall the next day. The important weather parameters also include wind gust speed, humidity at 9 am cloud cover at 3 pm, which when any increases of these increases, it increases the likelihood of rainfall the next day. Although AIC offers the highest prediction accuracy, it is worth noting that AIC Backwards model may select different combinations of weather parameters when run on the same data set. This shows that AIC Backwards model is unstable and a further investigation is required.

The main limitation of the results present in this paper is the short timeline of the data used (one year). This does not allow comparison with other years. The data used in this study was also measured in a year when the Australian weather was experiencing El-Nino.

REFERENCES

[1] U. Ratnayake and S. Herath, “Changing rainfall and its impact on landslides in Sri Lanka,” *J. Mt. Sci.*, vol. 2, no. 3, pp. 218–224, 2005, doi: 10.1007/bf02973195.

[2] M. M. Smadi and A. Zghoul, “A Sudden Change In Rainfall Characteristics In Amman, Jordan During The Mid 1950s,” *Am. J. Environ. Sci.*, vol. 2, no. 3, pp. 84–91, 2006, doi: 10.3844/ajessp.2006.84.91.

[3] L. Wilson, M. J. Manton, and S. T. Siems, “Relationship between rainfall and weather regimes in south-eastern Queensland, Australia,” *Int. J. Climatol.*, vol. 33, no. 4, pp. 979–991, 2013, doi: 10.1002/joc.3484.

[4] R. Allan, J. Lindesay, and D. Parker, *El Nino Southern Oscillation and climatic variability*. 1996.

[5] F. H. S. Chiew, T. C. Piechota, J. A. Dracup, and T. A. McMahon, “El Nino/Southern Oscillation

and Australian rainfall, streamflow and drought: Links and potential for forecasting," *J. Hydrol.*, vol. 204, no. 1–4, pp. 138–149, 1998, doi: 10.1016/S0022-1694(97)00121-2.

[6] R. J. B. Fawcett and R. C. Stone, "A comparison of two seasonal rainfall forecasting systems for Australia," *Aust. Meteorol. Oceanogr. J.*, vol. 60, no. 1, pp. 15–24, 2010, doi: 10.22499/2.6001.002.

[7] L. Firth, M. L. Hazelton, and E. P. Campbell, "Predicting the onset of Australian winter rainfall by nonlinear classification," *J. Clim.*, 2005, doi: 10.1175/JCLI-3291.1.

[8] D. R. Sikka, "Some aspects of the large scale fluctuations of summer monsoon rainfall over India in relation to fluctuations in the planetary and regional scale circulation parameters," *Proc. Indian Acad. Sci. - Earth Planet. Sci.*, vol. 89, no. 2, pp. 179–195, 1980, doi: 10.1007/BF02913749.

[9] F. W. Zwiers and H. Von Storch, "On the role of statistics in climate research," *Int. J. Climatol.*, vol. 24, no. 6, pp. 665–680, 2004, doi: 10.1002/joc.1027.

[10] K. Kar, N. Thakur, and P. Sanghvi, "Prediction of Rainfall Using Fuzzy Dataset," *Int. J. Comput. Sci. Mob. Comput.*, vol. 8, no. 4, pp. 182–186, 2019.

[11] C. C. Cheung, A. M. Hart, and M. R. Peart, "Projection of future rainfall in Hong Kong using logistic regression and generalized linear model," in *5th International Workshop on Climate Informatics*, 2015, pp. 24–25.

[12] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[13] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contr.*, 1974, doi: 10.1109/TAC.1974.1100705.

[14] G. Schwarz, "Estimating the Dimension of a Model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978, doi: 10.1214/aos/1176344136.

[15] D. W. Hosmer and S. Lemesbow, "Goodness of fit tests for the multiple logistic regression model," *Commun. Stat. - Theory Methods*, vol. 9, no. 10, pp. 1043–1069, 1980, doi: 10.1080/03610928008827941.

[16] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, 1982, doi: 10.1148/radiology.143.1.7063747.

[17] C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.*, vol. 8, no. 4, pp. 283–298, Oct. 1978, doi: 10.1016/S0001-2998(78)80014-2.

[18] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis," *J. Clin. Epidemiol.*, vol. 49, no. 12, pp. 1373–1379, Dec. 1996, doi: 10.1016/S0895-4356(96)00236-3.

[19] J. Durbin and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression: I," *Biometrika*, vol. 37, no. 3/4, p. 409, Dec. 1950, doi: 10.2307/2332391.

[20] R. D. Cook, "Detection of Influential Observation in Linear Regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, Feb. 1977, doi: 10.1080/00401706.1977.10489493.

[21] G. E. P. Box and P. W. Tidwell, "Transformation of the Independent Variables," *Technometrics*, vol. 4, no. 4, pp. 531–550, Nov. 1962, doi: 10.1080/00401706.1962.10490038.

[22] D. W. Hosmer and S. Lemeshow, *Applied logistic regression. 2nd Edition*. 2000.