Enhanced Knowledge in Sciences and Technology Vol. 1 No. 2 (2021) 98-108


**EKST**

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/ekst
e-ISSN: 2773-6385

# Prediction of Aeroplane Crash Fatalities Using Regularization Regression

**Khuneswari Gopal Pillay**[1*], **Audrey Err Kiat Joo**[1]

[1]Department of Mathematics and Statistics,
Universiti Tun Hussein Onn Malaysia, Pagoh, 84600, MALAYSIA

*Corresponding Author Designation

DOI: https://doi.org/10.30880/ekst.2021.01.02.012
Received 20 May 2021; Accepted 22 June 2021; Available online 29 July 2021.

**Abstract**: Air transportation is extensively used these days, and the safety of air transportation is affected as the number of aeroplane crash is getting increase, therefore it is important to reduce the risk of an aeroplane crash. Hence, this study aims to describe aeroplane crash fatalities based on the factors affecting aeroplane crash fatalities, to compare the model selection between logistic regression and LASSO logistic regression in terms of prediction of the presence aeroplane crash fatalities, and to identify the main factors affecting aeroplane crash fatalities. The number of aeroplane crash fatalities from 1st January 2000 to 31st December 2019 is described by using a bar chart. The performance of model selection by using LASSO logistic regression and logistic regression is compared by using the accuracy and precision obtained from the confusion matrix and the AUC value obtained from the ROC curve. The factors that affect the presence of aeroplane crash fatalities are determined from the best model. Based on the results, LASSO logistic regression shows a better performance compared to logistic regression in the analysis of prediction of aeroplane crash fatalities. In conclusion, three main factors which are the flight phase where aeroplane crash happened, regions of aeroplane crash happened, and causes of an aeroplane crash were concluded to show a significant sign that affects the presence of aeroplane crash fatalities. With this, the airline operators should take more precautions to prevent the presence of aeroplane crash fatalities.

**Keywords**: LASSO Regression, Logistic Regression, Prediction Of Aeroplane Crash Fatalities

## 1.0 Introduction

As air transportation is commonly used these days, the safety of air transportation is affected by the world. The aeroplane crash reported creates fear among travellers on their safety when using air transportation. There is a study found out that there was a significantly high proportion of travellers injured or died due to aeroplane crashes [1]. However, the fatality rate of an aeroplane crash is minimal. In the year 2017, there are 50 fatalities out of 4.1 billion travellers where the fatality rate is equivalent to 12.2 deaths per billion travellers [2].

*Corresponding Email: khuneswari@uthm.edu.my

Nevertheless, there are several aeroplanes' crashes with a high number of fatalities. The cause of aeroplane crashes due to mechanical errors, pilot error [3], weather, and sabotage. Based on the track record, the pilot fault is the major cause of the aeroplane crash [4]. Inaccuracy or misinterpretation of information by pilots might result in an aeroplane crash. The responsibility of the pilot and proper aeroplane maintenance required as an aeroplane is made up of hundreds of distinct systems. A malfunction in a system will affect the performance of another system.

The International Civil Aviation Organization (ICAO) [5] stated that the occurrence of aeroplane crashes rose year by year. The factor that influences aeroplane accident fatalities is different in each accident. There was a rise of 30% in the number of accident fatalities in the year 2019 [6]. Therefore, this paper focusing the prediction of aeroplane crash fatalities by using LASSO Logistic Regression.

## 2.0 Materials and Methods

### 2.1 Description of the Dataset

The dataset was obtained from the website of the Bureau of Aeroplane Accident Archives for the aeroplane accident data between 1st January 2000 and 31st December 2019. The dependent variables and the independent variables are shown in Table 1.

**Table 1: Descriptions of data for aeroplane crash**

| Variable | Descriptions | Description |
|---|---|---|
| $Y_1$ | Number of aeroplane crash fatalities | |
| $Y_2$ | Presence of fatalities | |
| $X_1$ | Local time of an aeroplane crash | AM |
| | | PM |
| $X_2$ | Flight phase where an aeroplane crash happened | Flight |
| | | Landing |
| | | Take-off |
| | | Parking |
| | | Taxiing |
| $X_3$ | Types of terrain where an aeroplane crash happened | Airport (less than 10 km from the airport) |
| | | City |
| | | Lake, Sea, Ocean, River |
| | | Mountains |
| | | Plain, Valley |
| | | Desert |
| $X_4$ | Age of aeroplane | 0-9 |
| | | 10-19 |
| | | 20-29 |
| | | 30-39 |
| | | 40-49 |
| | | 50-59 |
| | | 60-69 |
| | | More than 70 |
| $X_5$ | Regions of an aeroplane crash | Asia |
| | | Africa |
| | | North America |
| | | Europe |
| | | South America |
| | | Central America |

| | | |
|---|---|---|
| $X_6$ | Causes of an aeroplane crash | Oceania<br>Human<br>Other reason<br>Sabotage<br>Technical failure<br>Weather |

## 2.2 Descriptive Statistics

Descriptive statistics has been used to present the summary of data by using a number or graph. This could guide researcher to understand the data clearly where the bar charts are used to represent the data.

## 2.3 Multicollinearity

Multicollinearity is a phenomenon in multiple linear regression where there is a correlation between the independent variables [7]. One of the indicators to detect multicollinearity is the Goodman-Kruskal lambda. The multicollinearity between independent variables in this study is detected by using Goodman Kruskal Lambda.

## 2.4 Cook's Distance

Cook's distance proposed by Cook [8] is a method used to detect the influential point in a dataset. The presence of influential points will affect the outcome and the accuracy of a regression model. The influential point is detected by measuring the effect of deleting the given observation. F-distribution is used to determine the cut-off line as a percentile of over 50 indicates a highly influential point. The calculation for Cook's distance is defined as

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_i - \hat{y}_{j(i)})^2}{(p+1)\hat{\sigma}^2} \quad Eq.\ 1$$

## 2.5 Logistic Regression

The logistic regression model is a regression model used to describe the relationship between a binary or dichotomous dependent variable and one or more independent variables [9]. It also can be used to model the probability of an outcome based on individual characteristics in the form of an odds ratio. As the odd ratio increases, the probability of an outcome decreases. In logistic regression, the dependent variable is a logit form where the logit is obtained from the log of odds. The expression for logit function is defined as

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_i x_i \quad Eq.\ 2$$

where $p$ represents the probability of success or event that ranges between 0 and 1 and $\beta_i$ are the regression coefficient associated with the reference groups and $x_i$ indicates the explanatory variables. The range of logit is between $-\infty$ and $+\infty$.

In logistic regression, the significant variables in the model can be determined by using backward elimination. At the initial step of backward elimination, a significance level is selected to decide whether the variables stay in the model. The common significance level is 0.05. Next, all the predictors are fitted with a model and the predictor with the highest $p$-value is removed. This step is repeated until all the predictors that have a $p$-value higher than the significance level are removed. The model is then rebuilt and fitted with the remaining variables.

## 2.6 LASSO Regression

The Least Absolute Shrinkage and Selection Operator (LASSO) is a suitable method for variable selection [10]. The log partial likelihood is minimized, and a constant is bounded by the absolute values of the parameters. The coefficients in the model are shrunk or some of the coefficients become zero if not significant. It aids to reduce the estimated variance which resulted in an interpretable final model. The general estimator for a linear regression model is defined as

$$\min_{\beta_0 \beta}\{\frac{1}{2N}\sum_{i-1}^{N} y_i - \beta_0 - \sum_{j-1}^{p} x_{ip}\beta_p\} \quad Eq.\ 3$$

$$subject\ to \sum_{j=1}^{p} |\beta_0| \leq t \quad Eq.\ 4$$

where *t* is the limit of the sum of absolute values for the parameter estimates and it is a tuning parameter with $t \geq 0$. It used to the estimates to control the amount of shrinkage by shrinking the coefficients towards 0 or some of the coefficients will be equal to 0.

2.7 Confusion Matrix

The confusion matrix is a method used to evaluate the performance of the model built [11]. Classification problems can be solved by using a confusion matrix as it can be applied in both binary and multiclass classification problems. The actual target values and the predicted target values are presented in the confusion matrix. By using a confusion matrix, the accuracy, misclassification rate, and precision or positive predictive value can be calculated. The equation of accuracy is as

$$Accuracy = \frac{TN+TP}{Total\ number\ of\ prediction} \quad Eq.\ 5$$

where *TN* is the number of predictions correctly predicted as negative and *TP* is the number of predictions correctly predicted as positive. The equation of the precision is as

$$P = \frac{TP}{TP+FP} \quad Eq.\ 6$$

where *FP* is the number of predictions incorrectly predicted as positive.

2.8 ROC Graph

The receiver operating characteristics (ROC) graph is a probability curve that is useful to organize the result of a binary outcome prediction and visualize the performance of the model [12]. It can be used to select the model based on the performance. The area under the curve (AUC) is the area under the ROC curve that is used to represents the degree or measure of separability. Based on the AUC, the ability of the model to distinguish between classes can be shown. The higher the AUC, the better the ability of the model to predict positive as positive and negative as negative.

## 3.0 Results and Discussion

3.1 Description of the Aeroplane Crash Fatalities Data

The number of aeroplane crash fatalities for the local time of an aeroplane crash, flight phase where aeroplane crash happened, types of terrain where aeroplane crash happened, age of aeroplane, regions of an aeroplane crash, and causes of an aeroplane crash are described by using bar-chart.

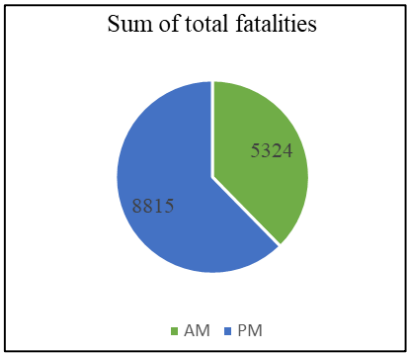


**Figure 1: Number of fatalities based on the local time of an aeroplane crash.**

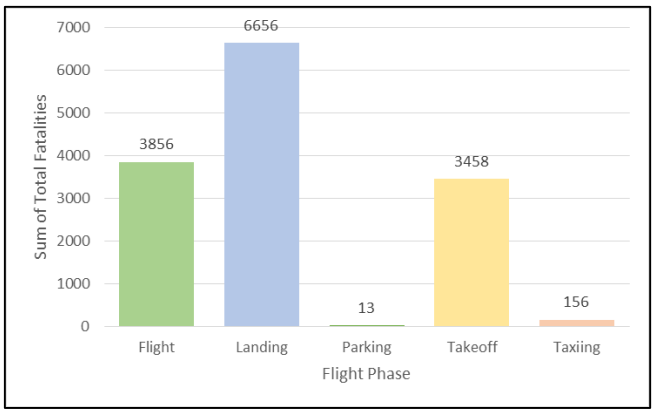


**Figure 2: Number of fatalities based on the flight phase where aeroplane crash happened.**

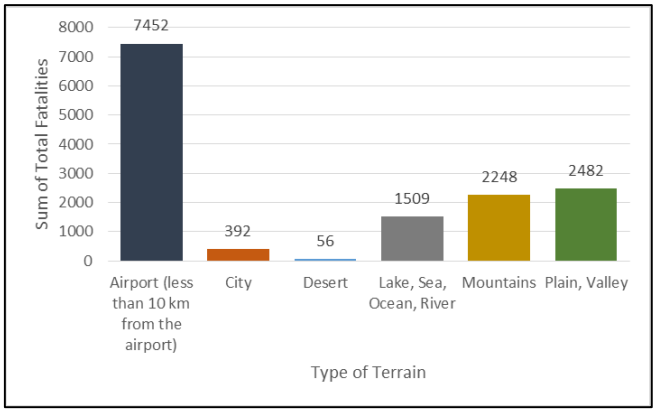


**Figure 3: Number of fatalities based on the type of terrain where the aeroplane crash happened.**

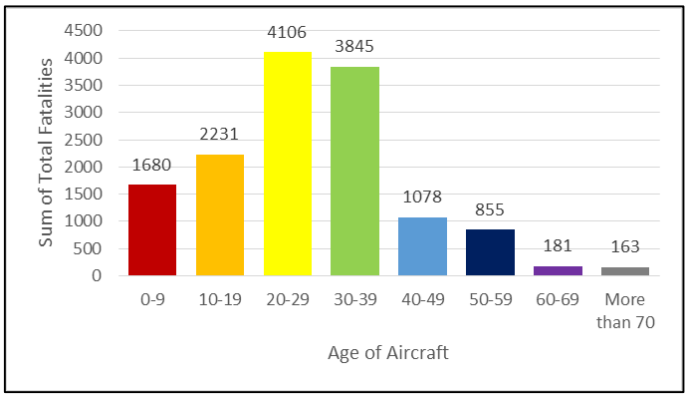


**Figure 4: Number of fatalities based on the age of aeroplane.**

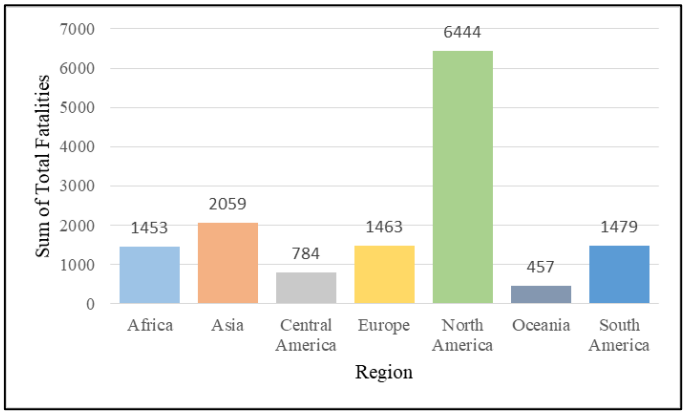


**Figure 5: Number of fatalities based on the regions where the aeroplane crash happened.**

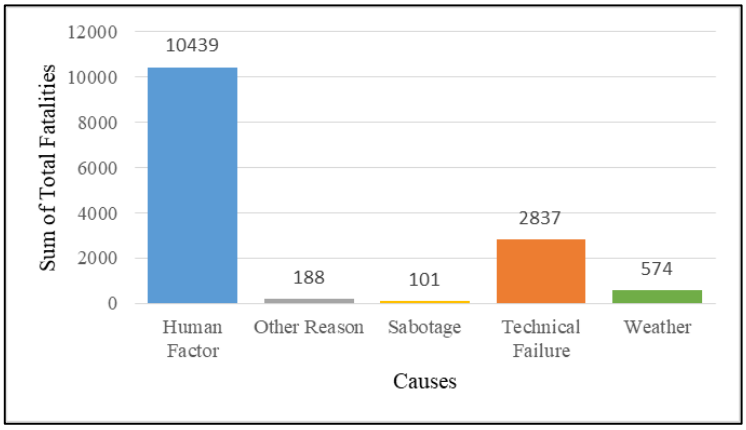


**Figure 6: Number of fatalities based on the causes of an aeroplane crash.**

## 3.2 The Model of prediction of Presence of Aeroplane Crash Fatalities

The dataset with the dependent variable of the presence of aeroplane crash fatalities is used. The dataset is tested with Goodman-Kruskal lambda to detect the multicollinearity between the independent variable. The result for Goodman-Kruskal lambda is shown in Figure 7.

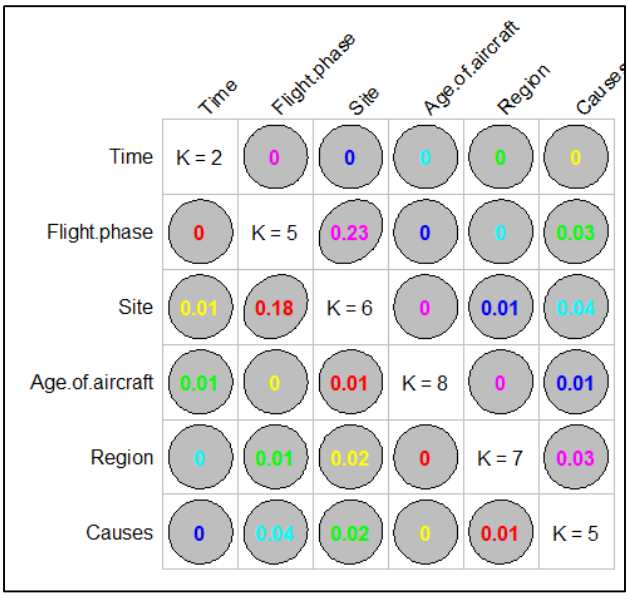


**Figure 7: Goodman-Kruskal lambda of all the independent variables**

The dataset is tested with Cook's Distance to detect the influential point. The result for Cook's Distance is shown in Figure 8.

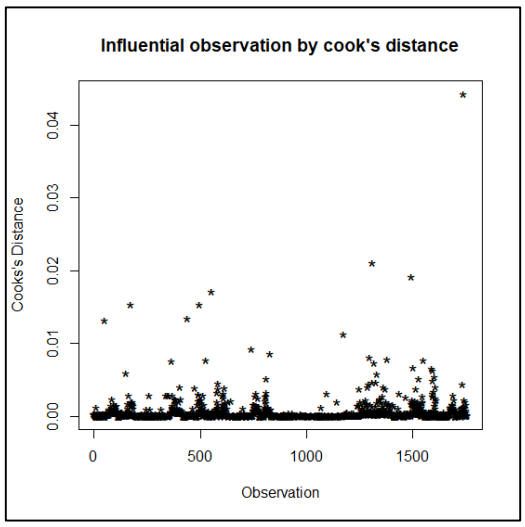


**Figure 8: Influential observation by Cook's Distance**

Figure 8 shows that there is no observation lies outside Cook's distance line where the Cook's distance is over 0.04. Hence, the data with 1756 observations could be used for the analysis. The data is encoded by using the One Hot Coding method. Then, the dataset is divided into a training set and a testing set. The model of logistic regression is built by using the training set and the insignificant variables are removed by using backward elimination. The equation for the prediction of aeroplane crash fatalities using logistic regression is as

$$\begin{aligned}\text{Logit}(p) = {} & -1.2312 + 1.7369\,\text{Flight.phase}_{\text{Flight}} + 1.2776\,\text{Flight.phase}_{\text{Landing}} + 1.7766\,\text{Flight.phase}_{\text{Takeoff}} - \\ & 1.1368\,\text{Flight.phase}_{\text{Less than 10 KM from Airport}} + \quad 1.9794\,\text{Flight.phase}_{\text{City}} - 1.8690\,\text{Site}_{\text{Desert}} + \\ & 1.9503\,\text{Site}_{\text{Mountains}} - 0.2529\,\text{Age.of.aeroplane}_{\text{10-19}} + 0.4280\,\text{Region}_{\text{Europe}} + \\ & 1.7207\,\text{Causes}_{\text{Human}} + 17.0054\,\text{Causes}_{\text{Other Reason}} + 16.9146\,\text{Causes}_{\text{Sabotage}} + \\ & 1.1117\,\text{Causes}_{\text{Technical Failure}}\end{aligned}$$

The testing data is applied to the model fitted to predict the presence of aeroplane crash fatalities. The result of the prediction is presented in a confusion matrix as shown in Table 2.

**Table 2: Confusion matrix of prediction using logistic regression.**

| | Predicted: No | Predicted: Yes | Total |
|---|---|---|---|
| Actual: No | 8 | 11 | 19 |
| Actual: Yes | 88 | 243 | 331 |
| Total | 96 | 254 | 350 |

Based on the confusion matrix of prediction using logistic regression, the accuracy of the model is 0.7171 by using *Eq.5* and the precision of the model is 0.9567. The ROC curve and the AUC value for the result of the prediction are shown in Figure 9.

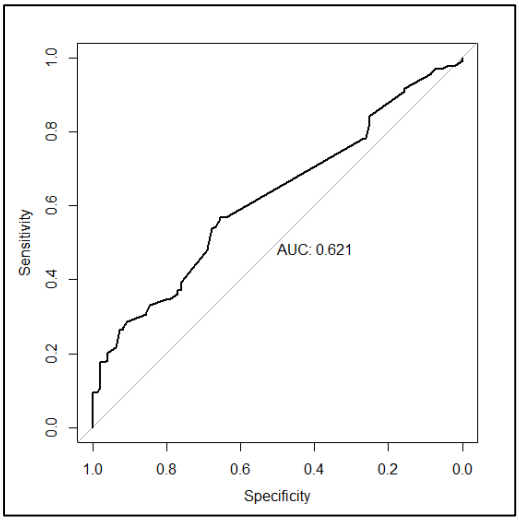


**Figure 9: ROC curve for the result of the prediction of logistic regression**

The training set then used to build a model of LASSO logistic regression. The lambda used for building the model is the *lambda.1se* as it gives the simplest model but the standard error of the optimal value of lambda still lies within one. The equation for the prediction of aeroplane crash fatalities using LASSO logistic regression is as

$$\text{Logit}(p) = 1.4292 - 0.0011\ \text{Flight.phase}_{\text{Landing}} - 0.6652\ \text{Site}_{\text{Less than 10 km from airport}} + 0.4161\ \text{Site}_{\text{Mountains}} - 0.5217\ \text{Causes}_{\text{Weather}}$$

The testing data is applied to the model to predict the presence of aeroplane crash fatalities. The result of the prediction is presented in the confusion matrix as shown in Table 3.

**Table 3: Confusion matrix of prediction using LASSO logistic regression.**

| | Predicted: No | Predicted: Yes | Total |
|---|---|---|---|
| Actual: No | 8 | 4 | 12 |
| Actual: Yes | 88 | 250 | 338 |
| Total | 96 | 254 | 350 |

Based on the confusion matrix of prediction using LASSO logistic regression, the accuracy of the model is 0.7371 by using *Eq. 5* and the precision of the model is 0.9843. The ROC curve and the AUC value for the result of the prediction are shown in Figure 10.

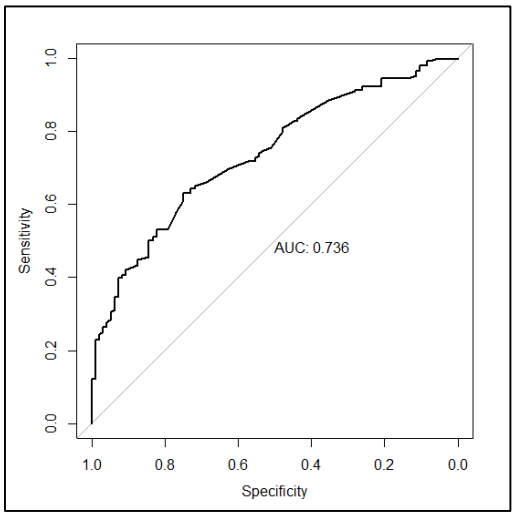

**Figure 10: ROC curve for the result of the prediction of LASSO logistic regression**

The prediction using logistic regression and LASSO logistic regression is compared by using the accuracy and the precision calculated from the confusion matrix and the AUC value obtained from the ROC curve. The result for accuracy, precision, and AUC value for logistic regression and LASSO logistic regression as shown in Table 4.

**Table 4: Accuracy, Precision, and AUC value for logistic regression and LASSO logistic regression**

| | Accuracy | Precision | AUC value |
|---|---|---|---|
| Logistic regression | 0.7171 | 0.9567 | 0.621 |
| LASSO logistic regression | **0.7371** | **0.9843** | **0.736** |

According to Table 4, the accuracy, precision, and the AUC value of the LASSO logistic regression are higher than the logistic regression. Since LASSO logistic regression showed better results on the accuracy, precision, and AUC value, hence the model built by LASSO logistic regression is suggested to use to predict the presence of aeroplane crash fatalities.

3.3 The factors affecting the presence of aeroplane crash fatalities

Overall, based on the model of LASSO logistic regression, the factors that affected the presence of aeroplane crash fatalities are the flight phase where aeroplane crash happened, regions of aeroplane crash happened, and causes of an aeroplane crash. The factors are determined by shrinking the coefficient of the insignificant variables into zero.

## 4.0 Conclusion

The number of aeroplane crash fatalities from the year 2000 to the year 2019 was described by using a bar-chart according to the local time of an aeroplane crash, flight phase where aeroplane crash happened, types of terrain where aeroplane crash happened, age of aeroplane, regions of an aeroplane crash and causes of an aeroplane crash. The category with the highest number of aeroplane crash fatalities as shown in the bar chart, however, the information is quite limited where it only presents the quantities of each category.

LASSO logistic regression has higher accuracy and precision compared to logistic regression by proving through the prediction of the presence of aeroplane crash fatalities. Besides, the ability of the LASSO logistic regression model to distinguish between classes is higher than logistic regression as it showed a higher AUC value compared to logistic regression. Hence, the best model for the prediction of the presence of aeroplane crash fatalities is the model built by LASSO logistic regression. Based on the model of LASSO logistic regression, the main factors affecting can be identified which are the flight phase where aeroplane crash happened, regions of aeroplane crash happened, and causes of an aeroplane crash.

There are some limitations to this research. The first limitation is the smaller number of variables involved in this study. As more factors are included in the study, a more accurate prediction can be obtained. Besides, more precautions can be taken by the airline to ensure a safer flight. The criteria for comparison between logistic regression and LASSO logistic regression is limited as the criteria used are the accuracy and the precision calculated from the confusion matrix and the AUC value obtained from the ROC curve. More criteria such as bias-corrected AIC [13] and Brier score [14] can be used for future study. Furthermore, more methods can be used for the prediction of aeroplane crash fatalities such as neural networks and genetic algorithms [15].

**Acknowledgment**

The authors would also like to thank to the Faculty of Applied Sciences and Technology, University Tun Hussein Onn Malaysia for its support.

**References**

[1] K. S. Guptill, S. W. Hargarten and T. D. Baker, "American travel deaths in Mexico, Causes and prevention strategies", Western Journal of Medicine, 154(2), p. 169, 1991.

[2] L. Behan, "Building a Framework for promoting adult motivation in mandatory manual and patient handling training", (Doctoral dissertation, Dublin, National College of Ireland), 2018.

[3] G. Li, "Pilot-related factors in aeroplane crashes: a review of epidemiologic studies", Aviation, Space, and Environmental Medicine, 1994.

[4] U. Kumar and H. Malik, "Analysis of fatal human error aeroplane accidents in IAF", Ind J Aerospace Med, 47(1), p. 30-36, 2003.

[5] International Civil Aviation Organization. Safety Report (2019). Retrieved on March 20, 2020, from https://www.icao.int/safety/Pages/Safety-Report.aspx.

[6] M. Goldstein, "Aviation Safety In 2019: Fewer Deaths but More Fatal Accidents", Retrieved on February 20, 2020.

[7] R. K. Paul, "Multicollinearity: Causes, effects, and remedies", IASRI, New Delhi, p. 8-65, 2006.

[8] R. D. Cook, "Detection of influential observation in linear regression", Technometrics, 19(1), p. 15-18, 1977.

[9] D. W. Hosmer Jr, S. Lemeshowa and R. X. Sturdivant, "Applied logistic regression", John Wiley & Sons, vol. 398, 2013.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso", Journal of the Royal Statistical Society: Series B (Methodological), 58(1), p. 267-288, 1996.

[11] O. Caelen, "A Bayesian interpretation of the confusion matrix", Annals of Mathematics and Artificial Intelligence, 81(3-4), p. 429-450, 2017.

[12] T. Fawcett, "An introduction to ROC analysis", Pattern recognition letters, 27(8), p. 861-874, 2006.

[13] C. M. Hurvich and C.L. Tsai, "Bias of the corrected AIC criterion for underfitted regression and time series models", Biometrika, 78(3), p. 499-509, 1991.

[14] C. A. T. Ferro, "Comparing probabilistic forecasting systems with the Brier score", Weather and Forecasting, 22(5), p. 1076-1088, 2007.

[15] A. Altay, O. Ozkan and G. Kayakutlu, "Prediction of aeroplane failure times using artificial neural networks and genetic algorithms", Journal of Aeroplane, 51(1), p. 47-53, 2014.