INTERNATIONAL CONFERENCE ON ADVANCES IN TECHNOLOGY AND COMPUTING (ICATC-2021)
Faculty of Computing and Technology (FCT), University of Kelaniya, Sri Lanka
18th December 2021

Index No: TL-33-66

Regularization Risk Factors of Suicide in Sri Lanka for Machine Learning

D. M. A. U. Delpitiya (1st Author)
Department of Physical Science,
Faculty of Applied Science,
University of Vavuniya
Vavuniya, Sri Lanka
achinidel@gmail.com

Prabha M. Kumarage(2nd Author)
Department of Software Engineering,
Faculty of Computing and Technology,
University of Kelaniya
Dalugama, Sri Lanka
prabhami_2019@kln.ac.lk

B. Yogarajah(3rd Author)
Department of Physical Science,
Faculty of Applied Science,
University of Vavuniya
Vavuniya, Sri Lanka
yoganbala@yahoo.com

Nagulan Ratnarajah (4th Author)
Department of Physical Science,
Faculty of Applied Science,
University of Vavuniya
Vavuniya,Sri Lanka
rnagulan@univ.jfn.ac.lk

***Abstract*— Indication to World Health Organization, suicide is a major world public health concern that is in the top twenty leading causes of death worldwide. Sri Lanka is a country that has the highest suicidal rates in the globe. The comprehensive study about risk factors for suicide is important because we can prevent or treat the recognized most risky categories of people. The emergence of big data concepts with machine learning techniques introduced a resurgence in predicting models using risk factors for suicide. Regularization is one of the most decisive components in the statistical machine learning process and this technique is used to reduce the error on the training dataset and prevent over-fitting. Comprehensive regularization approaches are presented here to select significant risk factors for age-specific suicide in Sri Lanka and build unique predictive models. The Least Absolute Shrinkage and Selection Operator (LASSO) approach presents regularization along with the feature selection to improve the prediction precision. The dataset collected for the study is rooted in the Sri Lankan people and the factors used for the analysis are, suicide person's gender, lived place, education level, mode of suicide, job, reason, suicide time, previous attempts, and marital status. Further, the riskiest age category of the people, who has exposure to suicide, is identified. Multiple linear regression and Ridge regression were used to evaluate the performance of LASSO. The selected most relevant factors with regularization to predict age-specific suicide prove the effectiveness of the proposed regularization approaches.**

Keywords* - *Suicide, LASSO, Ridge, Machine Learning

I. INTRODUCTION

Suicide means someone ending their own life and it is a way for people to escape pain or suffering. Mental disorders, including depression, bipolar disorder, autism, schizophrenia, personality disorders, anxiety disorders, substance abuse including alcoholism and the use of benzodiazepines, stress, such as from financial difficulties, relationship problems such as breakups, or bullying and there are many more reasons for suicide. According to the World Health Organization (WHO) statistics, suicide is a major world public health concern [1]. It is among the top twenty leading causes of death all around the world, with more deaths due to suicide than to malaria, breast cancer, war, and homicide. There are close to 800 000 people who die by suicide every year [1]. Today, Sri Lanka has recorded one of the highest rates of suicide within the world.

Suicide prevention in developing countries is an essential social and public health objective. For this protection, the most significant risk factors should be identified and build a predictive model for counsellors, and health and human service workers. Many studies [2]-[5] on risk factors for suicide were published in developing countries based on socio-demographic, clinical, and environmental/situational factors. Studies of Sri Lankan suicidal data-based risk factors analysis were presented by the research community, such as studies on risk factors for acute pesticide poisoning [6] and demographic risk factors in pesticide-related suicides [7]. Jeanne Marecek [8] has presented the social ecology of young women's suicide in Sri Lanka and discussed the implications of the cultural, ecological, and psychological factors. Lakmali et al [9] identify the major factors affecting suicide in Sri Lanka using Quasi Poisson and negative binomial regression models. However, these traditional statistical approaches to the prediction of risk factors for suicide attempts have limited accuracy and scale of risk detection [10].

Recently, the advancement of computer power, the large amount of data, and more importantly development of more advanced machine learning algorithms such as deep learning [11], are utilized to analyze any complete data and predict models. Machine learning-based techniques, such as Support Vector Machines (SVM), Artificial Neural Networks (ANN) and deep learning-based algorithms such as Recurrent Neural Network (RNN), provide systems with the ability to self-learn and improve from practice. This process of learning begins with observations or data, such as examples, direct experience, or instruction and considers patterns of those data. Then it can make better decisions eventually based on the examples that we have given. The main desire of this concept is to allow computers to self-learning without human aid and adjust actions appropriately. Supervised learning is one of the machine learning tasks, which is a learning model, built to make predictions, given an unforeseen input instance. Supervised learning algorithms take a well-known input dataset and its known responses to the data of output to learn the regression or classification model. These learning algorithms then train a model to generate a prediction for the response to a new dataset. Supervised learning uses classification algorithms and regression techniques to develop predictive models.

The goal of machine learning is to model the pattern and ignore the noise. A machine learning algorithm is trying to fit the noise in addition to the pattern, it is called over-fitting. The model gets a low accuracy if it is over-fitted. Thus, the avoiding over-fitting issues are one of the main phases of training a machine-learning model. Over-fitting occurs when the model is tracking heavily to capture the noise in the training dataset. Most commonly, cross-validation is used for avoiding over-fitting that guides in examining the error over a dataset, and in determining which variables are working best for the model. Regularization is a form of regression, which regularizes or shrinks the coefficient estimates close to zero. Particularly, this approach discourages learning a more complex or flexible model to avoid the risk of over-fitting. LASSO is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model [10]. Some of the studies in the literature focused on predicting the risk of suicide attempts through machine learning without using proper regularization approaches [12]-[14].

II. OBJECTIVES

This research paper aims to detect the significant variables, the risk factors for suicide, to create a unique predictive model for each age group to use in the machine learning purpose. eal world dataset of suicides collected from the Sri Lankan general hospitals is used to detect the risk factors for suicide according to age. The LASSO model is used to apply the suicide dataset for variable selection, regularization and prediction. In addition, the riskiest age group for the suicide attempt is identified in this study. The results were evaluated using a comparison with multiple regression analysis and Ridge regression with LASSO. The LASSO model provides excellent prediction precision, guides to extend the model intelligibility by removing negligible factors which do not relate to the responding variable and reduces over-fitting. The initial and important part of the machine learning process is accomplished here to choose the risk factors and build prediction models.

III. METHODOLOGY

A. Suicide Dataset

Mainly the suicide dataset was collected from the Forensic Medical Office, District General Hospital-Nawalapitiya in February 2020. We have conducted an interview with the forensic medical officer and gathered the dataset and got approval to use the dataset for research purposes. Furthermore, we have collected more data from Teaching Hospital-Peradeniya by an interview with the forensic medical officer of the hospital. The dataset consists of 120 data that is containing the details of persons, who have committed suicide from February 2016 to February 2020 in the Nawalapitiya and Kandy districts.

Originally, the dataset contains 14 risk factors of individual information. Data pre-processing was made to ensure that the collected data is correct, consistent, and usable by identifying any errors in the data. The 10 risk factors were extracted based on the pre-processing step (eliminating incomplete, noisy and inconsistent data) and the suggestion from the medical officer, which make machine learning algorithms work well, better represent the underlying problem to the predictive models and improve the accuracy of the model. It contains the data of the suicide person's age, gender, lived place, education level, mode of suicide, job, reason, suicided time, marital status, and the details about if there are previous attempts to suicide. By using the responding variable as age categories, we have done our analysis to get the most risk factors of suicide for using in the machine learning approach and to identify the riskiest age category for committing suicide. Here, we have decided to gather all the suicide persons' data of all age ranges.

B. Variables Selection Using Regression Models

The variable (factor) selection method is used to select the most significant variables from the above exposure variables to predict the responding variable. Thus, we have to use the multiple linear regression, Ridge, and LASSO methods with making use of the R software [15]. However, in the first step, we have to analyze the dataset in a better understandable way.

1) Categorization of Dataset: The data was properly organized in the first phase using Microsoft Excel 2016 as a perfect dataset and it was categorized as the responding variable and the exposure variables. Then correlations between the exposure variables were computed.

2) Implementation of Multiple Linear Regression Model: At the second phase, a multiple linear regression model was built from the training dataset and an ANOVA table was revealed. This model is the best-illustrated model that consists of the highly correlated independent variables [16].

By using this model, we can inspect the exposure variables which have high significance and low significance for the model. We can observe the most significant exposure variables for use in the model from the ANOVA table. Moreover, by using the ANOVA table, we can select the factors for the finalized model. To use this regression model, the responding variable should be normally distributed, should be a linear relationship among the responding variable and the exposure variables, and those variables should not be correlated, mean should be zero and variance should be constant.

In addition, to check the multicollinearity of the training dataset, we calculated the Variance Inflation Factors (VIF) values. Examining these values, we can decide the multicollinearity exists or not in the model.

3) Implementation of Ridge and LASSO Models: The Ridge and LASSO regression models were used to avoid confusion and determine the factors, which are contained in the finalized model because those methods can be used for efficient variable selection. In the Ridge and LASSO regressions, correlated exposure variables were used and there were no assumptions to use with these models. For the Ridge and LASSO, implementations have been done by the following procedures.

First, two different models were obtained for the Ridge and LASSO regression models with different λ values. Here the λ is the tuning parameter and this λ values control the strength of these models with great importance. The most affecting factor, negatively and positively affected factors and the low significant factors were obtained by the number of graphs from those Ridge and LASSO models. After that, by doing cross-validation for the models, we got the most convenient λ values. Then we selected the minimum λ values and according to that λ values obtained the most significant exposure variables to use in the finalized models of Ridge and LASSO. The most appropriate models were constructed to predict the suicide causes using the most significant factors selected by the Ridge and LASSO models.

We computed the adjacent R-squared for all the variables and therefore the most significant factors independently for examining the effectiveness of the chosen factors. We researched the medical and psychological background for each significant variable to justify the correctness. Furthermore, we checked the correctness of the models by testing new data.

Finally, we compared all the built models by use of the Residual Sum of Squares (RSS). Initially, we approximately calculated the λ values based on the training dataset by double cross-validation methods using the best λ values. Then we applied these estimated λ values in the test dataset. In addition, coefficients were estimated, and the residual sum of squares (RSS) was calculated for each model.

IV. RESULTS AND DISCUSSION

A. Variables Selection Using Regression Models

The 10 risk factors were extracted from the raw suicide dataset for 120 committed suicide persons. The responding variable of this study is Age, and the Age categorized into 4 groups, 1-15 years, 16-30 years, 31-50 years and 51-70 years. The selected nine exposure variables are,

1. *Gn* : Gender of the suicided person
2. *Lp* : Division where he/she lives
3. *El* : Education level
4. *Mod* : Which type of the suicide
5. *Jb* : Occupation
6. *Rsn* : Reason of suicide
7. *Time* : Time of suicide
8. *PrAt* : Number of previous attempts to suicide
9. *Ms* : Marital status (single/married/divorced)

We have selected 110 samples among the suicide dataset as the training data and to assess the correctness of the regression model, the last 10 samples remained as testing data.

1) Categorization of the Suicide Dataset:

Table 1 illustrates the correlations among all response and exposure variables. It shows that correlations between variables are powerful. Some factors are positively dependent on other factors and some factors are negatively correlated with others. For illustration, the responding variable *Age* shows high dependency with the exposure variables, *Jb*, and *Rsn*. In addition, *Age* negatively depends on the variables *Gn*, *Time*, and *Ms* and positively depends on all other variables. Therefore, according to the Table 1, we can examine, the higher number of factors in this dataset are multi-correlated covariance variables.

Table 1. Correlations "Suicide Dataset"

	Age	Gn	Lp	El	Mod	Jb	Rsn	Time	PrAt	Ms
Age	1									
Gn	-0.129	1								
Lp	0.183	-0.048	1							
El	0.164	0.056	-0.001	1						
Mod	0.201	0.076	-0.015	-0.010	1					
Jb	0.589	0.020	-0.087	0.052	-0.119	1				
Rsn	0.052	0.079	0.047	0.218	-0.024	-0.042	1			
Time	-0.102	0.004	-0.075	-0.036	-0.065	-0.030	0.213	1		
PrAt	0.107	-0.130	-0.132	0.112	-0.080	0.217	0.072	-0.125	1	
Ms	-0.341	0.005	-0.301	-0.028	-0.005	-0.246	0.091	-0.099	0.134	1

2) Implementation of Multiple Linear Regression Model: We built an ANOVA table (Table 2) for examining which are the most significant for the model and which are less significant exposure variables for the model.

Analyzing Table 2, the results show that, according to the p-values, *Job* and *Mod* have high significance. Therefore, we can predict these exposure variables will be involved in the finalized model for prediction. We cannot actually wrap up these factors are contained in the finalized model due that it is not guaranteed whether those variables are optimal or not. There is a feasibility for the not optimal, because the variables *PrAt*, *Time*, and *Rsn* are not significant factors and should not be included in the finalized model

Table 2. ANOVA "Suicide Dataset"

Responding Variable: *Age*					
	Df	**Sum Sq**	**Mean Sq**	**F Value**	**Pr (>F)**
Gn	1	1.718	1.718	3.6323	0.059541
Lp	1	3. 254	3.254	6.8785	0.010088
El	1	3. 032	3.032	6.4094	0.012909
Mod	1	4. 814	4.814	10.1779	0.001899
Jb	1	40.654	40.654	85.9428	3.932e-15
Rsn	1	0.328	0.328	0.6928	0.407182
Time	1	0.313	0.313	0.6621	0.417749
PrAt	1	0.054	0.054	0.1152	0.735060
Ms	1	1.884	1.884	3.9826	0.048692
Residuals	100	47.303	0.473		

Table 3 illustrates the computed Variance Inflation Factors (VIF) to check the multicollinearity of the training dataset. By examining Table 3, we can see some of these values are rather than high such as *VIFMs*, *VIFJob* and *VIFLp*. According to the table, we can conclude multicollinearity exists in this model.

Table 3. VIF values for "Suicide Dataset"

Variable	**VIF**
Gn	1.040016
Lp	1.182521
El	1.075680
Mod	1.032261

Jb	1.190100
Rsn	1.155741
Time	1.125189
PrAt	1.157541
Ms	1.282413

3) Ridge Regression: For obtaining the most significant exposure variables, we used the Ridge regression method with its functions to predict the responding variable *Age* by using all the exposure variables in the suicide dataset.

Fig. 1 shows a graphical representation of different values of log λ versus coefficients of each exposure variable. Each color line illustrates the exposure variables and their expansion in the model. From this graph, we can identify the entered position of each exposure variable to the model and the influenced scope of the responding variable.

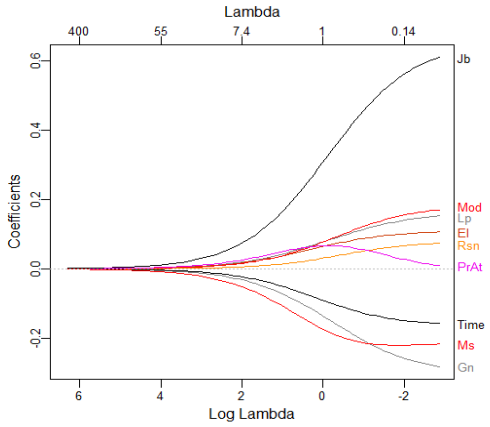

Fig. 1. Glmnet graph for exposure variables - Ridge regression

Fig. 1 reveals that the most influenced exposure variable is *Jb* in the model, it steadily and positively affected the responding variable *Age*. *Ms* is the second important variable and that negatively affected the responding variable *Age* and entered the model lately, but it also affected the trend of *Jb* after entered in the model. *Gn is* also one of the most significant exposure variables because it also enters the model later and it affects the trend of *Ms*. Furthermore, we can select *Mod* and *Lp* as other important factors by examining their trends. All other variables in the model can be discarded from the finalized model because those are less significant.

Fig. 2 shows the plot that the green lines illustrate more significant factors that negatively influence the responding variable, while red lines illustrate the ones that influence it positively.

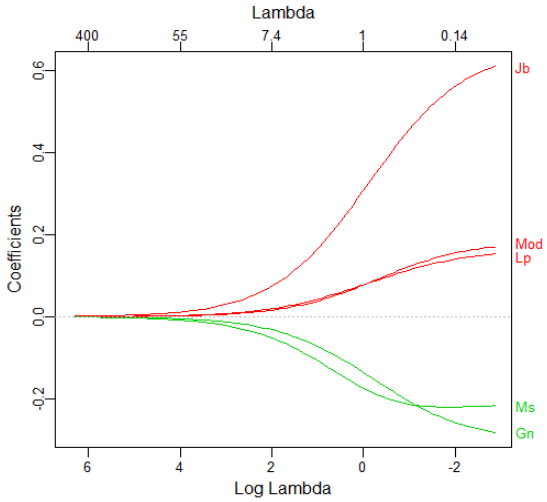

Fig. 2. Glmnet graph for significant variables - Ridge regression

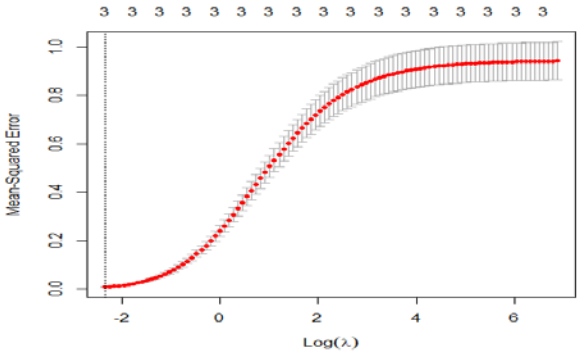

Fig. 3. Cross-validation – Ridge regression (nfold = 5)

As the next step, we used cross-validation to find the most suitable value for the λ. It is useful for controlling the strength of the penalty. For that, we considered λ_{min} that gives minimum mean cross-validated error and λ_{1se}, the standard error of the minimum. In this study, these two values are the same. Therefore, both two values are represented by the vertical dotted line in Fig. 2. Moreover, we can select the value for tuning parameter λ that better fits to the situation. In this analysis, λ_{min} is not noticeable due to that Fig. 3 plot illustrates an exponential trend. According to the plot, we can select factors as likely, the most significant variables from both the values of λ from the variables shown in Fig. 2. Thus, we have for λ_{min} and λ_{1se} as 0.09693229 and the most significant variables that we can contain in the finalized model are *Jb*, *Lp*, *Mod*, *Ms*, and *Gn*.

4) LASSO Regression: We obtained the most significant factors which are contained in the finalized model using LASSO regression.

Fig. 4 depicts the graphical representation of the different values of log λ versus coefficients of each exposure variable. Colored lines illustrate the exposure variables and their expansion into the model like Ridge regression.

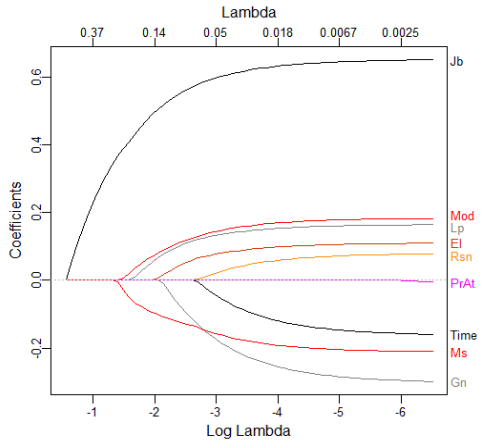

Fig. 4. Glmnet graph for exposure variables - LASSO

By examining Fig. 4, the most influenced exposure variable is revealed, which is *Jb*, and it steadily and positively affected the responding variable *Age*. We decided *Ms*, *Gn*, *Mod*, and *Lp* as the other important variables as in Ridge regression by looking at their trends and all the other variables in the model discarded from the finalized model because those are less significant.

Fig. 5 plot shows that the green lines determine more significant attributes that negatively influence the responding variable, while red lines determine the ones that influence it positively. Using the cross-validation into the LASSO, we found the most suitable value for the λ, like the Ridge regression. The λ_{min}, minimum mean cross-validated error and λ_{1se}, the standard error of the minimum is the same for the obtained LASSO values. Therefore, we represented both two values as a vertical dotted line in Fig. 5. In addition, we selected the value for the tuning parameter λ for the better fit problem.

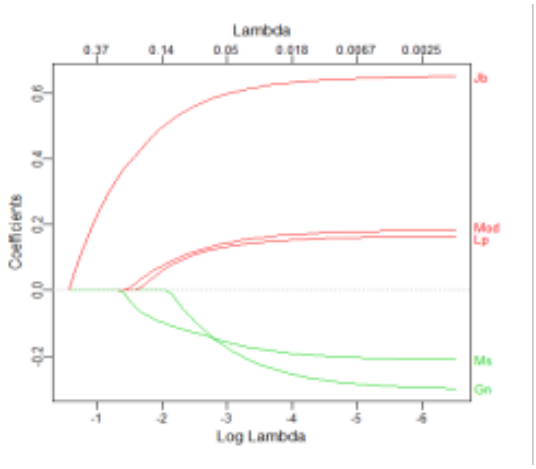

Fig. 5. Glmnet graph for most significant variables – LASSO

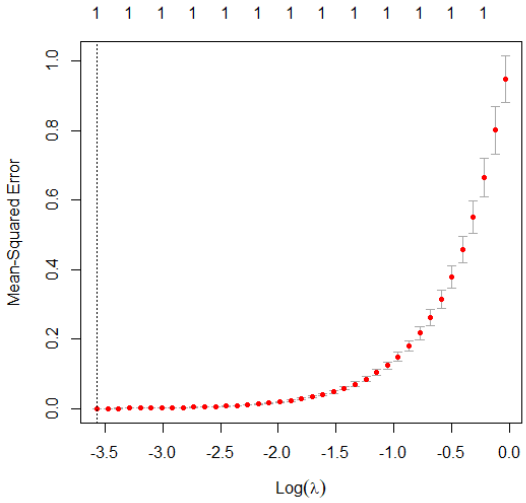

Fig. 6. Cross-validation (nfold=5)

In this analysis, λ_{min} is not visible because Fig. 6 plot illustrates an exponential trend. Moreover, the factors were selected as likely, which are the most significant exposure variables according to both the values of λ. From that, we have the value for λ_{min} and λ_{1se} as 0.02825628. Thus, the most significant factors included in the finalized model are *Jb*, *Lp*, *Mod*, *Ms*, and *Gn* and those variables were selected according to the optimized λ_{min} and λ_{1se} values.

B. Adjusted R Squared

The calculated adjusted R squared value is 0.5011 for multiple linear regression with all the nine exposure variables. After the selection of the most significant exposure variables *Jb*, *Lp*, *Mod*, *Gn*, and *Ms*, the adjusted R squared value was obtained as 0.4883. Thus, we proved that the factors we have selected using the Ridge and LASSO for the finalized model are highly significant risk factors for the responding variable *Age*.

C. Comparing coefficients and RSS

The linear, Ridge, and LASSO models were compared by using the Residual Sum of Squares (RSS). Initially, we approximated the λ v8alues with respect to the training dataset by using two cross-validation methods of λ_{min} and λ_{1se}. After that, we applied these estimated λ values for training and testing datasets separately to calculate the estimated coefficients and residual sum of squares for each model.

For those two cross-validation methods, we have got the same values as for LASSO the λ_{min} as 0.02825628 and λ_{1se} as 0.02825628 and for Ridge Regression the λ_{min} as 0.09693229 and λ_{1se} as 0.09693229. Therefore, we selected λ_{min} from both to estimate the coefficients and RSS.

According to the results, we proved that the coefficient estimates get closer to zero. In LASSO, some of the non-zero coefficients changed to zero. Then, we applied the estimated λ values for the testing dataset to get new estimated coefficient values. From those results, most of the coefficient values in the linear model changed as negative values and most values are close to zero. And also in LASSO, the number of zero coefficient values in the test dataset is greater than the number of coefficient values in the training dataset.

Table 4. RSS coefficient values for train and test data

RSS	Linear model	Ridgeλ_{best}	Lassoλ_{best}
RSS_{TRAIN}	47.34059	47.8663	48.08071
RSS_{TEST}	148.577	138.3825	138.4715

The RSS values for the training and testing datasets are shown in Table 4. According to that, the linear regression model predictably achieves the smallest RSS value on the training dataset. The RSS for Ridge and LASSO are again typically greater when λ is chosen using the "one-standard error" rule. Interestingly, the errors on the test dataset are ordered unexpectedly; the minimum is achieved by Ridge at λ = ridge.bestlam, and the second is LASSO at λ = lasso.bestlam while linear regression takes just the third position. Here, Ridge and LASSO are acting better than linear regression on the testing dataset.

D. Most Risky Age Group

As a sub-finding from the dataset, we have selected the most vulnerable age group for suicide. In this study, we used four categories of age groups, 1-15 years, 16-30 years, 31-50 years, and 51-70 years. For the test dataset, we have randomly selected sets of two exposure variables from our finalized model variables and used the Analysis of Covariance (ANCOVA) method for analyzing.

Table 5. ANCOVA for age groups

Responding variable is Gn				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	1.380912	0.121496	11.366	<2e-16
Lp	-0.003278	0.041676	-0.079	0.9374
Age1	0.091522	0.069983	1.308	0.1936
Age2	-0.146364	0.064107	-2.283	0.0243
Age3	0.087537	0.082822	1.057	0.2928

Table 5 shows the ANCOVA table for the variables *Gn* and *Lp*. We can see that group *Age2* (16-30 years) is the significant age group to attempt suicide because the *p*-value of *Age2* is less than a significant value of 0.05.

According to the ANCOVA values, the other sets of random significant variables also provided the most significant age category as *Age2*.

V. CONCLUSION

In this study, comprehensive regularization approaches have been presented to select significant risk factors for suicide in Sri Lanka and an important part of the machine learning process was accomplished. Ridge and LASSO regressions for the machine learning approach are explained with the regularization task. To demonstrate the regularization with those two approaches, we utilized the Sri Lankan real time suicide dataset. We have concluded that the most related factors with regularization to predict the suicide ages by Ridge and LASSO are suicided person's gender, lived place, mode of suicide, occupation, and marital status. We proved the effectiveness of these most significant factors by computing the adjusted R-squared values. In addition, use of the coefficients of each and the residuals sum of squares (RSS), we have proven that Ridge is most suitable, and then LASSO for reach for the regularization. We have obtained that the riskiest age category for committing suicides is the young adults age between 16 - 30 years. The results of this study showed that Ridge and LASSO can use to build a unique model for predicting the suicide ages more accurately in machine learning approaches.

This study provides not only a platform for the machine learning approaches to further investigation of risk factors of suicide in Sri Lanka but also it will help for the activities of suicide prevention in Sri Lanka. Moreover, in the medical field researchers, counsellors, and doctors give advice or treatment to the most risk category of people.

REFERENCES

[1] World Health Organization, *Suicide in the World: Global Health Estimates*, 2019.
[2] Lakshmi Vijayakumar Sujit John, Jane Pirkis, Harvey Whiteford Suicide in Developing Countries (2): Risk Factors, Crisis. 2005; 26(3):112-9.
[3] Lakshmi Vijayakumar1 Suicide prevention: the urgent need in developing countries, World Psychiatry. 2004 Oct; 3(3): 158–159.
[4] Murad M Khan Suicide Prevention and Developing Countries, Journal of the Royal Society of Medicine, Volume: 98 issue: 10, page(s): 459-463, 2005.
[5] Michael RPhillips Gonghuan Yang Yanping Zhang et al., Risk factors for suicide in China: a national case-control psychological autopsy study, The Lancet, Volume 360, Issue 9347, 30 November 2002, Pages 1728-1736.
[6] Wim Van Der Hoek, Flemming Konradsen Risk factors for acute pesticide poisoning in Sri Lanka, Tropical medicine and international health, Volume10,issue 6 June 2005, Pages 589-596
[7] E B R Desapriya, P Joshi, G Han, F Rajabali Demographic risk factors in pesticide related suicides in Sri Lanka, Injury prevention, Volume 10, Issue 2, 2004.
[8] Jeanne Marecek Young Women's Suicide in Sri Lanka: Cultural, Ecological, and Psychological Factors, Asian Journal of Counselling, 2006, Vol. 13 No. 1, 63–92

[9] Lakmali, S. M. M., & Nawarathna, L. S. (2019). Identifying and Predicting Major Factors Affecting the Suicides in Sri Lanka. Asian Journal of Probability and Statistics, 2(3), 1-7. https://doi.org/10.9734/ajpas/2018/v2i328785

[10] V. Fonti., "Feature Selection Using Lasso", Vrije Universiteit Amsterdam,2017

[11] M. M. Najafabadi1, F. Villanustre, T. M. Khoshgoftaar1, N. Seliya1, R. Wald1 and E. Muharemagic, "Deep learning applications and challenges in big data analytics", Najafabadi et al. Journal of Big Data, 2:1, DOI 10.1186/s40537-014-0007-7, 2015.

[12] L.E. Melkumova, S.Ya. Shatskikh," Comparin Ridge and Lasso Estimators for Data Analysis", Elsevier 2017: P. 746-755.

[13] Gregory D. S. "Regularization Techniques for Multicollinierity: Lasso, Ridge, and Elastic Nets," Henry M Jackson Foundation, 2018.

[14] (2018) Linear and Ridge Regression with R, [Online]. Available: https://www.pluralsight.com/

[15] Imran Amin, "Prediction of Suicide Causes in India Using MachineLearning", Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, 2017.

[16] Prabha M. Kumarage, B. Yogarajah, and Nagulan Ratnarajah. "Efficient Feature Selection for Prediction of Diabetic Using LASSO", *19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2019..

[17] (2016)Adolescent suicide behaviours in 32 low- and middle – income countries homepage on World Health Organization. [Online]. Available: https://www.who.int/bulletin/volumes/94/5/15-163295/en/

[18] Jana Vorlickova, "Least Absolute Shrinkage and Selection Operator Method", Faculty of Social Science, Charles University, 2017.

[19] Thevaraja, Mayooran & Gabriel, Mathew, "Comparing Linear, Ridge and LASSO Regressions". 10.13140/RG.2.2.30269.77282, 2018.

[20] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani," An Introduction to Statistical Learning with Applications in R", 2013.

[21] Comparing Multiple Means in R homepage on DATANOVIA. [Online]. Available:https://www.datanovia.com/en/lessons/ancovain .

[22] Tibshirani, R., Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B (Methodological), 58, 267-288, 1996.