Landslide Susceptibility Mapping using Association Rule Mining Based Apriori Algorithms and Multiple Clustering Algorithms

CN Madawala[1], BTGS Kumara[2]

[1,2] *Department of Computing & Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka*

[#1] CN Madawala; cnmadawala@std.appsc.sab.ac.lk

[#2] BTGS Kumara; btgsk2000@gmail.com

Abstract— Haphazard development activities on mountain slopes and inadequate attention to construction aspects have led to increasing landslides and sustaining damages to the lives and the infrastructure. According to the National Research Building Organization (NBRO) reports, within the study area, nearly 3275 sq.km of the area expanded over the Ratnapura District; and 2178 sq.km area is to be highly prone to landsliding. If the appropriate investigations were performed in time, most of the landslides could be predicted relatively. This study aims to discover the real extent and severity of the landslide processes and risk evaluation within the study area. Machine Learning Approach based on Association Rule Mining and multiple Clustering algorithms were combined and implemented for the final prediction. This study possesses a strong capability to predict landslides risk by considering causative factors slope, Landuse, Geology, Soil material, elevation, intensity, and triggering factor; rainfall. Apriori Association rule algorithm, K-mean and Expectation Maximization (EM) clustering algorithms are the highest-ranking prediction algorithms. While applying the EM clustering method showed accuracy over 84% of the results with high speed and time taken to build was 0.66 seconds. The K-means algorithm gained the highest accuracy over 92% was applied and time taken to build was 1.58 seconds, though it was more time-consuming than EM algorithm. While applying the Apriori algorithm to obtain the best results, therefore ten (10) efficient prediction rules have found to fulfil the ultimate goal of this research. Moreover, the results show that the EM and Apriori algorithms have the best degree to fit for the Landslide susceptibility mapping.

Keywords— **Apriori, K-mean, Expectation Maximization (EM)**

I. INTRODUCTION

Landslides are the geological incident which includes widely spread land movement which resulting meticulous damages to the people and their belongings. Fundamentally the landslide transpired when a part of a natural slope is not capable of bearing its weight. The gravity is the fundamental driving force of the landslide refuses flow relying on the slope of the area. Landslide happens when the stability of the slope changes from a stable to an unstable state. In the last decades, there was a considerable increase in landslide frequency, in accord with the climatic changes, improper land uses and the expansion of urbanized areas in the world. Thus, landslide detection is a crucial requirement in pre and post-disaster, hazard analysis processes.

The recognition of landslide susceptibility is important to get some preventive and control actions and give some early warnings to reduce or mitigate hazards impacts. The objective of landslide-hazard mapping and risk evaluation is to determine the real extent, timing, and severity of landslide processes in selected high-priority areas of the Sri Lanka, where such knowledge will provide the most significant benefit to government officials, consulting engineering firms, and the general public in avoiding the landslide hazard or in mitigating the losses.

The formation and occurrence of landslides is a complicated evolution process, which is caused by the interaction of multiple instability factors. However, most of the methods consider only the current value of the instability factors while ignoring the factors' evolution feature over time. This study evaluates the Landslides susceptibility mapping which is an integration of Association rule mining and clustering. A variation of spatial data, including landslide geology, topography, slope, soil cover, and land use, and triggering factor rainfall were identified and collected in the study areas.

Machine learning is a sophisticated fusion of applied mathematics and computational intelligence. It focuses on 'training' an algorithm to probe for and learn from data structure robust enough to make predictions; even without predecessor knowledge of the structure. During recent decades, disaster information extraction and prediction were mainly based on artificial visual interpretation *(Gariano and Guzzetti, 2016)*. Apart from being time-consuming and modelling including heuristic, strenuous, deterministic and the traditional method also has a limitation in that the measurement process in lack of

accuracy and heavily depends on experts' knowledge *(Kadavi, Lee and Lee, 2018)*. With the development of the computer vision and pattern recognition technologies, it is possible to make the hazard assessment automatic. Therefore, exploration of new hybrid machine learning methods for landslide susceptibility modelling should be further carried out.

Clustering is a significant data mining technique that the author has used to extract useful information from various dimensional datasets *(A. Mingoti and A. Matos, 2012)*. It based on a set of objects into clusters so that the objects are similar in the same cluster but dissimilar compared to objects in other clusters. Several clustering methods have proposed and defined as a mathematical technique designed for exposing classification structures in the data collected in real-world phenomes. This study represents two clustering algorithms; K-means and the Expectation Maximization (EM) algorithm for Landslide risk analysis*(Chae et al., 2017)(Korup and Stolle, 2014)*.

Association rule mining used to find the connection between different data items in a database. Therefor Apriori Association rule algorithms used to discover frequent association which is first-rate, simple and could be applied for rule mining. The Apriori leads to the best performance by reducing the size of candidate sets *(Nafie Ali and Mohamed Hamed, 2018)*.

The paper aims to suggest an approach for employ association rules mining and clustering algorithms to offering new rules from a broad set of discovered rules. Meanwhile, identify an expected location that would cause damages or disruptions to existing standards of safety in Ratnapura District, Sri Lanka. Hence, the study helps to predict the most triggering factors of a landslide and changes can be expected in the activity of massive landslides in the future under the impact of environmental changes.

II. METHODOLOGY AND EXPERIMENTAL DESIGN

A. Study Area

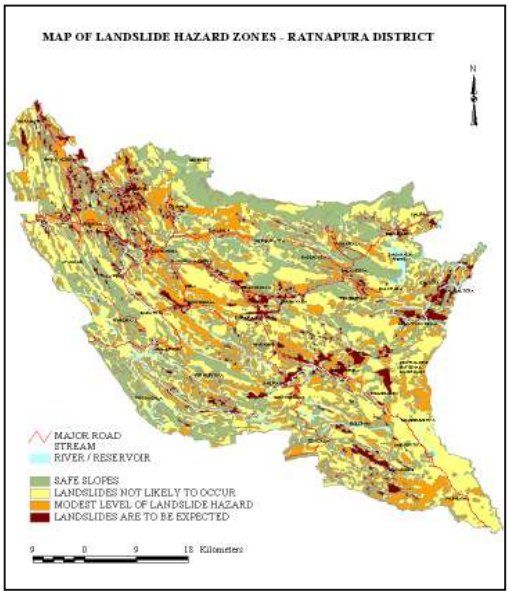


Figure.1 The study areas, the Ratnapura District, Sabaragamuwa Province.

Nearly 3275 sq.km of area expanded over the Ratnapura District, 1097 sq.km forestry areas and seems to be highly prone to land sliding of 2178 sq.km. 473 out of 575 GM divisional areas reported as Landslide-prone areas. Fig 1. shows a severe landslide occurred in Eheliyagoda, Nivithigala, Ayagama, kalawana, Dolapallehena, Kiribathgala, Alupathgala, Hortonwatta, Girapagama and approximately 14 of Landslide-prone and Embilipitiya, Godakawela, Kolonna 03 non-landslide-prone AGM divisional areas around Ratnapura District, Sabaragamuwa province in Sri Lanka.

B. Data pre-processing

1) *Data cleaning:* Data cleaning concerns the detection and removal of errors and inconsistencies in a data set due to the corruption or inaccurate entry of data in database. Moreover, identify the incomplete, incorrect or irrelevant data and then either replaced, modified or deleted.

2) *Identify Outlier:* An outlier is an observation in a set of data that is inconsistent with the majority of the data. An outlier can represent a valid score of a subject who just happens to represent an extreme case of the variable under study.

3) *Normalization:* This technique used to organize the data in the database to discover the new range of data from existing data range.

Here is the equation of Min-Max Normalization;

$$V = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Here, V_i is normalizing value between 1 and 0. x_i represent i^{th} attribute value of the data. Min(x^i) and Max(x^i) are Minimum and Maximum value of i^{th} attribute.

C. Materials

1) Required Data: Data sources and usage displays in Tables 1,

Table 1. Data used and Sources

Data use	Sources
Rainfall	Metrology Department
Soil Materials	Irrigation Department
Bedrock Geology	Survey Department
Land Use	LUPP Department
Intensity of landslide	NBRO (Colombo)
Slope angle/ Aspect	Survey Department
Soil Texture	Survey Department
Dip and Strike (Geology)	NBRO (Ratnapura)
Elevation	NBRO (Ratnapura)

To analyze the climate variability (rainfall); daily rainfall data in the past, recent five years (2012-2017) and monthly rainfall data (2000 -2012) have collected.
The author has collected the Hazard Zonation map and identified the most prominent landslide location. Then by the use of ArcGIS tool and Digital elevation model (DEM) tool, slope angle has been calculated.

2) Software tools and models:

- WEKA (Waikato Environment for Knowledge Analysis) Data Mining tool
- Arc GIS 10.4 version
- Digital Elevation Model (DEM)
- Clustering model

3) Programming Language

- Java

D. Association Rule mining model

Association rule was applied to find the link between data items in a transactional database and used to discover frequent association.

Let $I = \{i_1, i_2, i_3, \ldots, i_n\}$ be a set of n attributes called items and $D = \{t_1, t_2, \ldots, t_n\}$ be the set of transactions. It is called database. Every transaction, t_i in D has a unique transaction ID, and it consists of a subset of item sets in I. A rule can be defined as an implication, $A \longrightarrow B$; where A and B are subsets of I ($A, B \subseteq I$), and they have no element in common, i.e., $A \cap B$. A and B are the antecedent and the consequent of the rule, respectively *(Maitrey and Jha, 2014).*

1) *Support:* The support of an item set A, supply (A) is the proportion of transaction in the database in which the item A appears. It can be revealed as the popularity of an item set.

$$Supply\ (A) = \frac{\text{Number of transaction in which } A \text{ appears}}{\text{Total number of transactions}} \quad (2)$$

2) *Confidence:* This can be defined as follows:

$$Confidence\ (A \longrightarrow B) = \frac{supply\ (A \cup B)}{supply(A)} \quad (3)$$

3) Lift: It is defined as:

$$Lift\ (A \longrightarrow B) = \frac{supply\ (A \cup B)}{Supply\ (A) * supply\ (B)} \quad (4)$$

This signifies the likelihood of the item set B happened when item A happens while taking into account the popularity of B.

4) Conviction: This rule can be defined as:
5)

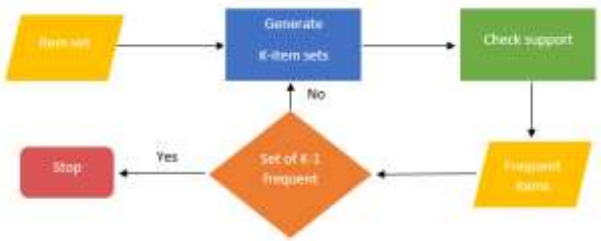
$$Conviction\ (A \longrightarrow B) = \frac{1 - supply\ (B)}{1 - confiction\ (A \longrightarrow B)} \quad (5)$$

It signifies the likelihood of item B happened when item A happens.

Various algorithms used to mine the rules therefor; the author endeavoured to find the best association rules using WEKA. Apriori is the first-rate and simple algorithm which could be applied for rule mining. The Apriori leads to the best performance by reducing the size of candidate sets.

E. The general process of Apriori Algorithm

The Apriori algorithm can be divided into two steps. The first step is to apply least support to find all the frequent sets with k items. And then use the self-join rule to find the



frequent sets with k+1 items with the help of k-item sets. Replicate this process from k=1 to the point we are unable to apply the self-join rule as shown in Fig 2.

Figure.2 Apriori general process

F. Method of Clustering model

The descriptive model includes the following tasks:

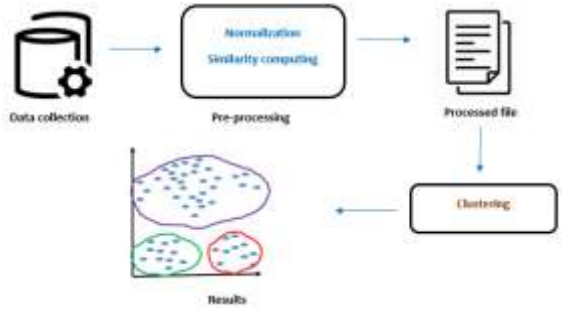


Figure.3 Descriptive model tasks

Data Mining is involved in predictive and descriptive models that are applied in many different tasks.

G. Clustering Algorithm

Clustering is significant data mining technique for extracting useful information from various dimensional datasets *(A. Mingoti and A. Matos, 2012)*. It based on a set of objects into clusters so that the objects are similar in the same cluster but dissimilar compared to objects in other clusters. Multiple clustering methods have proposed and defined as a mathematical technique designed for exposing classification structures in the data collected in real-world phenomes. This represent two clustering algorithms; K-means and the Expectation Maximization (EM) algorithm. In this study, the clustering approach is applied to EM clusters and the K-means clustering method and, it proposed for ensuring the accuracy of the Landslide risk analysis *(Chae et al., 2017)(Korup and Stolle, 2014)*.

1) K-mean clustering: The k-means clustering technique begins with a description of the basic algorithm. Choosing k initial is the first step, where k is a user-specified parameter *(Dabbura, 2018)*;

$$T_{min} = \Sigma^{k}_{i=1} \Sigma v_j \in d_i \, || \, v_j - \mu_i \, || \qquad (6)$$

Where; clustering n data points into k disjoint subsets (D_i) containing data points minimizes the sum-of-squares criterion, such that where v_j is a vector representing the j^{th} data point and μ_i is the geometric centroid of the data points in D_i.

A more common measure in K-mean clustering is **Euclidean distance**. It computed by finding the square of the distance between each variable. Thus, if a and b are two points on the real line, then the Euclidean distance between them is computed as:

$$\sqrt{(a-b)^2} = |a-b| \qquad (7)$$

As in Cartesian coordinates, if $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ are two points in Euclidean n space, then the distance from x to y or from y to x is given by *(Ng, 2012)*

$$d(x,y) = d(y,x) = \sqrt{\Sigma^{n}_{i=1}(y_i - x_i)^2} \qquad (8)$$

The second step is to specify each point to the closest centroid, and every collection of points assigned to a centroid is a cluster. The centroid of every cluster is updated based on the points assigned. These task steps are repeated and updated until there are no point change clusters *(Al-Augby et al., 2015).*

2) *Expectation–Maximization (EM):* EM is an iterative approach to discover maximum likelihood or minimum likelihood estimates of parameters in statistical models. The EM iteration substitutes between making an Expectation (E) step, which produces a function for the expectation of the log-likelihood estimated using the contemporary estimate for the parameters, and a Maximization (M) step, which estimates parameters were maximizing the expected log-likelihood found on the E step *(Eldin et al., 2017)*.

3) *EM algorithm for Gaussian Mixture Models (GMM)*: A Gaussian Mixture Model (GMM) is a probability distribution. GMMs can model distributions with many peaks. This is achieved by adding several Gaussian together. So mathematically we can define Gaussian mixture model as mixture of K Gaussian distribution that means it's a weighted average of K Gaussian distribution. So we can write data distribution as (Daniel Foley, 2019);

$$p(x) = \sum_{k=1}^{k} \pi_k \, N(x \mid \mu_k, \Sigma_k) \qquad (9)$$

where x is a D dimensional π_k vector, is the weight of the k^{th} Gaussian component, μ_k is the dimensional vector of means for k^{th} the Gaussian component and Σ_k is the D by D covariance matrix for the k^{th} Gaussian component. N is a D dimensional Gaussian of the form;

$$N(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{1/2} \, |\Sigma|^{1/2} \exp(-1/2 \, (x-\mu)^T) \, \Sigma^{-1}(x-\mu))} \qquad (10)$$

where $|\Sigma|$ is the determinant of Σ.

III. RESULTS

Results of this research based on two categories. First one is Applying Apriori Algorithm for finding the corresponding mining rules which can be used to landslide susceptibility mapping. And the other one relies on comparative analysis between K-mean and EM clustering algorithms for evaluate the landslide risk within the study area.

A) Applying Apriori Association Rule mining

Table 2 shows the best rules found in the Apriori algorithm. The results depend on the comparison of confidence, leverage, and convince. All rules in this table support the rules presented in the above Table.

After the Apriori algorithm executed, the author obtained many results. Table 2 explains the results, **Rule 1** if the main category of land use equal to Build-up areas and sub category equal to Homesteads landslide could occur in the hill country. And the **Rule 2** sub_cat of land use was Homesteads and Slope angle equal to very high angle (Normalized value range 0.81-0.9) also caused for the landslide. Most landslides could occur when the Rainfall is very very high (0.91-1.0), and Land use was Main_cat equal to Build-up areas and Sub_cat equal to Homesteads according to **Rule 3**. And **Rule 4** explains the high intensity could be obtained with Land use main_cat equal to Rocky areas and sub_cat equal to Areas exposed with rocks. If the slope is very very high (0.91-1.0) and rainfall is high (0.71-0.8) at Build-up areas is has the highest possibility to occur landslide (**Rule 10**). High elevation (0.91-1.0) expanded over the study area with the land use of Build-up areas (**Rule 8**). **Rule 9** the rock type, charnockite gneiss was spread over the rocky areas within the study area. **Rule 5,6** displays the results reached with Soil type is Sandy Clay Residual is most prominent in the Build-up areas, and with the high elevation, it could occur landslide situations within the Ratnapura District, Sri Lanka.

Results were indicated the Performance measures of Apriori algorithm. The measure of Minimum Support performed over Apriori was 0.4 (378 instances), and Minimum metric <Confidence> was 0.9. And the Lift and Conviction of every rule was mentioned in below Table 2.

Table 2. Best rules using Apriori Algorithm

List of attributes	Best rules found
Land cover (main_Cat)	1. Main_cat=Built - Up areas 44 ==> Sub_cat=Homesteads 44 <conf:(2)> lift:(4.34) lev:(0.24) [11] conv:(22.88)
Land cover (sub category)	2. Sub_cat=Homesteads 44 ==> Slope=Very high angle 38 <conf:(1)> lift:(2.94) lev:(0.15) [14] conv:(14.52)
Rainfall	3. Rainfall= Very High rainfall 26 Main_cat=Built - Up areas ==> Sub_cat=Homesteads 39 <conf:(1)> lift:(4.34) lev:(0.14) [7] conv:(14.02)
Intensity	4. Intensity=High intensity 24 Sub_cat=Areas exposed with rocks ==> Main_cat=Rocky areas 48 <conf:(1)> lift:(5.82) lev:(0.16) [7] conv:(14.92)
Slope	10. Slope=Very very high angle 60 Rainfall=High ==> Main_cat=Built - Up areas 20 <conf:(1)> lift:(14.55) lev:(1.16) [7] conv:(14.68)
Elevation	8. Elevation=very high elevation Sub_cat=Homesteads ==> Main_cat=Built-up areas 42 <conf:(1)> lift:(6.94) lev:(0.21) [7] conv:(14.52)
Rock Type	9. Rock Type=charnockite gneiss 88 Main_cat=Rocky areas ==> Sub_cat=Areas exposed with rocks 88 <conf:(1)> lift:(2.17) lev:(0.05) [5] conv:(5.4)
Soil Type	6. Main_cat=Built - Up Areas Soil=Sandy Clay Residual 33 ==> Sub_cat=Homesteads 33 <conf:(1)> lift:(2.17) lev:(0.06) [5] conv:(5.94) 5. Soil=Sandy Clay Residual Elevation=Very very low elevation 12 ==> Sub_cat=Homesteads 12 <conf:(1)> lift:(2.17) lev:(0.06) [6] conv:(6.48)

B) Applying Clustering Algorithms

As a first step, k is specified as value 3. Therefor the contents were classified into three classes High Risk (0), Moderate Risk (1) and Low Risk (2) based on an area for a visible and straightforward explanation of Landslides are most likely to occur.

1) *Attribute Selection using Ranking:* Rank attributes can eliminate irrelevant attributes; where there are many ways of scoring the features, which are called attributes. In this study author has used *InforGainattributeEval* algorithm which is available as subclasses of WEKA and found four causative factors as attributes for clustering approach which mentioned above Table 3.

Table 3. Ranked Attributes

Attribute	*Ranked*
0	1 Rainfall
0.165	2 Elevation
0.537	3 Slope
0.619	4 Intensity
0.856	5 Strike (Bedrock)
0.751	6 Dip (Bedrock)
0.36	7 Rock Type
0.654	8 Main_cat (Land Use)
0.35	9 Sub_cat (Land Use)
0.158	10 Soil Type

Clustering Algorithm	Attribute	Mean value (#clusters)			Standard deviation			True positive	True negative
		0	1	2	0	1	2		
EM	Rainfall	0.2754	0.2026	0.1824	0.2227	0.2053	0.1347	83.485	16.515
	Slope	0.4698	0.344	0.2364	0.1611	0.2154	0.1901		
	Intensity	0.5624	0.3021	0.0954	0.144	0.1821	0.1005		
	Elevation	0.2231	0.8369	0.1008	0.1653	0.2247	0.0832		

A Clustering algorithms defines the predictive value based on Accuracy. For this instance, the above equation used to calculate the accuracy.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

1) Experimental results applying K-mean Clustering: The accuracy of the data analysis based on the correctly classified and incorrectly classified is 91.132%, and the inaccuracy 8.868%, respectively (Table 4).

Table 4. Final centroid of K-mean

2) Experimental results applying EM Clustering:

The accuracy of the data analysis was based on the mean value and the standard deviation. The accuracy of correctly and incorrectly classified was 83.485% and 16.515%, respectively below Table 5.

The achieved results explained that the processing speed of K-mean was slower than EM clustering, but the classification accuracy of the data (Table 4) were greater than that achieved by EM.

Through the results obtained from current study showed that in K–mean cluster algorithm time taken to build model was (1.58 s), EM cluster algorithm (0.66 s), and the

Table 5. Applying EM Clustering

probability of identifying a correct group of data elements Log likelihood was (more than 85%). The K-means algorithm achieved the highest accuracy over 92%, but it was more time-consuming than EM algorithm.

IV. DISCUSSION AND CONCLUSION

Discussion and Conclusions of this study; Landslide risk evaluation was done at Ratnapura District, Sri Lanka using Association rule mining and clustering algorithms. In order to comparative analysis between K-means and EM clustering methods were compared in terms of accuracy of clustering results and processing speed. And Apriori mining rule was used to investigate the best association rules which performed landslide prediction. All methods examined showed better accuracy to experimental data with WEKA program.

The EM clustering method showed accuracy over 84% of the results with high speed. The highest accuracy over 92% was achieved by the K-means algorithm was applied, but it was more time-consuming than EM. Moreover, comparing both clustering algorithm performance measurements; EM cluster algorithm more effectively better than K-mean algorithm for Landslide prediction. Therefor EM-based Landslide risk evaluation model can help implement a guide to reduce the disruptive impacts of a natural disaster on surrounding communities.

While Applying Apriori algorithm in this research, the author has obtained for the 10 best results because

Clustering Algorithm	Attribute	Mean value (clusters(k))		
		0	1	2
K-mean	Rainfall	0.1424	0.164	0.4777
	Slope	0.5324	0.1324	0.5122
	Intensity	0.5624	0.3021	0.0954
	Elevation	0.5857	0.0946	0.3641

efficient rules have found through all frequent itemsets to fulfil the ultimate goal of this research, Landslide susceptibility mapping. Moreover, the results show that the EM cluster algorithm and Apriori algorithm have the best degree to fit for the Landslide susceptibility mapping.

Apriori utilizes a level-wise approach where it will generate patterns containing item sets. On the other hand, **FPGrowth** utilizes a depth-first search instead of a breadth-first search and uses a pattern-growth approach, unlike Apriori, it only considers patterns existing in the database. As a **recommendation**, the author has suggested using FPGrowth for pattern recognition for Landslide Prediction. In this **future research** researcher intend to concentrate on Deep Learning Approach which is a part of a broader family of machine learning methods based on artificial neural networks. Moreover, soon, a landslide warning system may be established by forecasting landslides induced by rainfall.

ACKNOWLEDGEMENT

The access to Hazard Zonation map of Ratnapura District from the Sri Lanka National Research Building Organization (NBRO) Distractor Dr. Mr. R.M.S. Bandara and Scientist Mr. Laksiri Indrathilaka, Landslide and non-landslide data have been provided by Mr. Abhitha Wanasundara, District officer; NBRO Ratnapura, Director General, Department of Metrology, has been acknowledged, Mr. A.L.K Wijemanna, Director (Computer, Research, Climate change, International Affairs) was supported to collect Rainfall Data and also Director General, Department of Land Use & Policy Planning, Mrs. A.S. Illangamge helped to collect Land Use data, and my sincere gratitude to the Sabaragamuwa University of Sri Lanka for encouragement.

REFERENCES

A. Mingoti, S. and A. Matos, R. (2012) 'Clustering Algorithms for Categorical Data: A Monte Carlo Study', *International Journal of Statistics and Applications*, 2(4), pp. 24–32. doi: 10.5923/j.statistics.20120204.01.

Al-Augby, S. *et al.* (2015) 'A Comparison Of K-Means And Fuzzy C-Means Clustering Methods For A Sample Of Gulf Cooperation Council Stock Markets', *Folia Oeconomica Stetinensia*, 14(2), pp. 19–36. doi: 10.1515/foli-2015-0001.

Chae, B. G. *et al.* (2017) 'Landslide prediction, monitoring and early warning: a concise review of state-of-the-art', *Geosciences Journal*, 21(6), pp. 1033–1070. doi: 10.1007/s12303-017-0034-4.

Dabbura, I. (2018) *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*, *Towards Data Science*.

Daniel Foley (2019) *Gaussian Mixture Modelling (GMM) – Towards Data Science*, *Data Scientist, Economics and Machine Learning*. Available at: https://towardsdatascience.com/gaussian-mixture-modelling-gmm-833c88587c7f (Accessed: 27 May 2019).

Eldin, Y. N. *et al.* (2017) 'A Comparative Analysis between K – Means & EM Clustering Algorithms', pp. 12809–12816. doi: 10.15680/IJIRSET.2017.0607048.

Gariano, S. L. and Guzzetti, F. (2016) 'Landslides in a changing climate', *Earth-Science Reviews*. The Authors, 162(August 2016), pp. 227–252. doi: 10.1016/j.earscirev.2016.08.011.

Kadavi, P. R., Lee, C. W. and Lee, S. (2018) 'Application of ensemble-based machine learning models to landslide susceptibility mapping', *Remote Sensing*, 10(8), pp. 1–18. doi: 10.3390/rs10081252.

Korup, O. and Stolle, A. (2014) 'Landslide prediction from machine learning', *Geology Today*, 30(1), pp. 26–33. doi: 10.1111/gto.12034.

Maitrey, S. and Jha, C. K. (2014) 'Association Rule Mining : A Technique for Revolution in Requirement Analysis', 4(8), pp. 4–9.

Nafie Ali, F. M. and Mohamed Hamed, A. A. (2018) 'Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents', *Journal of Information and Telecommunication*. Taylor & Francis, 2(3), pp. 231–245. doi: 10.1080/24751839.2018.1448205.

Ng, A. (2012) 'CS2999 Lecture Notes: The EM Algorithm', *Machine Learning*, 1(X), pp. 139–172. doi: 10.1007/978-3-642-21551-3_6.