International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Analysis and Prediction of the Quality of Biocontrol using Machine Learning Classifications

Mohamed Taybe Elhadi

Zawia, University, Az Zawiyah, Libya
Email: *mtelhadi[at]zu.edu.ly*

Abstract: ***Weed control relies mainly on integrated control methods of preventive, agricultural and chemical methods. On pasture lands, however, the chemical methods of spraying pesticides in large area is expensive, has negative consequences on ground water, on environment and on health in general. A safer and more cost effective alternative is biocontrol of weeds in which harmful and unwanted grass, weeds in general are subjected to some natural enemy to control it directly and indirectly. Leafy spurge is one common weed native to central and southern Europe that have spread across western Canada and North America. Not only does this invasive alien plant expand to overtake nearby areas; the milky liquid from its stems and flowers causes severe skin rashes or irritation in in livestock and humans. The weed has been targeted by beetles from the flea beetle genera Aphothona as biocontrol since they were introduced into Canada in the 1980s. It has been discovered that the growth of the A. n. agent and its effectiveness as a biocontrol agent is determined by the interaction of a variety of factors. However, understanding the nature of the relationships among those many factors is incomplete and unclear. A machine learning approach to the analysis of such factors and to the prediction of suitability and potential success of control sites is the subject of this paper. The methodology was used to analyse the available data taken from Regina Agriculture Station weed control project to provide scientists the ability to predict the suitability of sites and the potential success of the agent before its release. It can be also used for the evaluation of existing sites. A number of machine learning classifier algorithms have been adopted and applied to the data including Random Frost, Nearest Neighbour, Support Vector Machines (SVMs), Logistic Regression, Neural Nets and Bayes with variable degrees of accuracy. The adopted classifiers are evaluated with best ones are selected based on Matthew's correlation factor (MCC) and the overall accuracy of prediction***

Keywords: Biological Control of Weeds, Machine Learning, Random Frost, Nearest Neighbour, Support Vector Machines, Logistic Regression

1. Introduction

Losses from weeds (unwanted plants that spread accidentally or intentionally) are believed to be equal to the combined losses from insects and diseases, and rank second only to losses from soil erosion. Cultural and chemical practices constitute the main methods of weed control. These methods are very costly, bring limited relief, increase soil erosion, contaminate underground water and pollute the environment (Huffaker &Messenger, 1976). As such these methods are considered hazards, undesirable and environmentally unsafe. Alternatively, a much safer method is the use of Biological Control (Biocontrol) based on the identification and manipulation of weeds' natural enemies to influence the abundance and existence of their host plants. Earlier, Biocontrol has resulted in various levels of success in the control of a wide range of weeds around the world and all over the prairies of Canada and the USA (M. H. Julien, 9187), (Harris, 1986), (Schwarzländer, Hinz, Winston, & Day, 2018).

According to (Harris, 1986):

"over 500 biological control agents being intentionally released against nearly 200 weed species in over 90 countries. Collectively, 15 countries in Asia and 17 of the 22 countries and territories in the Pacific region have intentionally released over 80 biological control agents to help manage over 30 of their most invasive weeds. Many of these programs, have been highly successful. In fact, globally, over a third of all weed biological control programs have resulted in some form of control of the target weed, resulting in huge benefit: cost ratios of up to 4, 000: 1. In addition, there have been very few (<1%) unpredicted, sustained non - target impacts on native or economic plants by weed biological control agents. "

Still, however, biocontrol faces challenges due to many factors (Day & Witt, 2019). One critical factor is to do with a scientist's inability to make sense of the relationships and dependencies that exist among agents, weeds, and levels of success in a multi - factors environment.

This work is an attempt to help scientist use existing data and different factors to evaluate and predict the possibilities of success of new projects and sites using machine learning techniques.

Data used in this work was collected and prepared through projects conducted by Agriculture Canada dealing with the control of Leafy Spurge (Gassman, 1985) weed using, among others agents, a beetle known as Aphothona nigriscutis (A. n.) (Elhadi, 1991).

Leafy Spurge is an herbaceous perennial of an Eastern European origin. It has dominated and excluded most other herbaceous plants on uncultivated land in the North American prairies since its introduction around 1865 with alarming increases in the last few decades (Elhadi, 1991). It has been discovered that the growth of the A. n. agent and its effectiveness as a biocontrol agent is determined by the interaction of a variety of factors. However, understanding of the nature of the relationships among those many factors is incomplete and certainly unclear. A data analysis methodology that results in a predictive classifications system, the subject of this paper, based on the application of

machine learning techniques was introduced and used. The methodology was used to automatically provide scientists the ability to predict the suitability of sites before the release of the beetle as well as evaluation of success of existing sites.

The rest of the paper contains a description of the used classifiers and datasets used, experiments performed, results and analysis followed by conclusions.

2. Machine Learning Classifiers

Machine learning algorithms have become an integral part of many data analyses especially classification and prediction. Approaches for machine learning are divided into three main categories: (1) Supervised learning which is used if the available data to be used for training has a labeled attribute and other data does not contain a label; (1) Unsupervised learning which has no labeled information, but the algorithms strive to discover any existing patterns in the data. (3) Deep learning which learns and improves using artificial neural networks with larger, sophisticated neural networks that aid in classification problems and its different applications such as language translation, and speech recognition (Mirtaheri& Shahbazian, 2022).

Supervised Learning

Supervised learning algorithms are the subject of our experiments. A number of algorithm have been developed and tested proofing themselves in the field of supervised learning. The following is a brief count of those algorithms used in our work:

- Random Forest (RF): RF algorithm is based on Decision trees which are a type of model used for both classification and regression. Decision tree models allow problems solving in an orderly and systematic way to draw logical conclusions (Biau &Scornet, 2016)]. The model behaves with "if this then that" conditions ultimately yielding a specific result. One wants to minimize bias errors as well as variance due errors. RF mitigates this problem well. A random forest is simply a collection of decision trees whose results are aggregated into one final result in order to limit overfitting without substantially increasing error levels. This normally done by training on different samples of the data.
- Naive Bayes (NB): is a simple probability model that is based on Bayes' theorem and strong (naïve) independence assumptions between attributes. It is a probabilistic machine learning model that's used for classification. Bayes' based classifiers are fast and easy to implement. They are however, based on the simplistic assumption of independence of predictors. It is mostly used in sentiment analysis, spam filtering, recommendation systems (Kaur&Oberai, 2014)
- Logistic Regression (LR): Logistic regression is used for predicting the categorical dependent variable using a given set of independent variables. It predicts a binary outcome, based on prior observations of a data sethas the ability to provide probabilities and classify new data using continuous and discrete datasets based the well - known sigmoid function (Wang, Yu, Qi, Hu, Zheng, Shi, Yao, 2019)
- K - Nearest Neighbor (KNN): The k - nearest neighbors are a supervised learning classifier, which uses proximity to make classifications or predictions. It is mostly for classification based on the assumption that similar points can be found near one another. A class label is assigned on the basis of the label that is most frequently represented around a given data point (Taunk, De, Verma, & Swetapadma, 2019).
- Support Victor Machines (SVM): SVM is used for both regression and classification but more in classification with the aim of finding a hyperplane in an N - dimensional space representing the number of features that distinctly classifies the data points. Support vectors are data points that are closer to the hyperplane which are used to maximize the margin of the classifier (Abiodun, Jantan, Omolara, Dada, Mohamed, & Arshad, 2018).
- Artificial Neural Networks (ANN): In particular, Multi - layer Perceptron classifier which relies on an underlying Neural Network to perform the task of classification (Brereton & Lloyd, 2010).

3. Datasets and Data Collection

Machine leaning technique are mostly data - driven and data - centred applications. They are trained to learn from exiting data and to analysis and classify unknown data. In this work empirical data has been collected from a number of locations (sites) where the agent beetles were released and monitored by researchers at the Regina Research Station of Agriculture Canada. Data from release sites were collected, tabulated and refined to arrive at a workable set of data that is suitable for computers.

Table 1: List of Factors Used

Name	Description	Name	Description
Size	Number of beetles released in the site	Aspect	Direction of slopes
Date	Part of the summer when site created	Relief	Site's relief: concave, convex
Span	Time from site creation till evaluation	Shade	Presence of shade
Organic C	Level of Organic Carbon in soil	Cover	Presence of bare ground
pH	Ph level of the soil.	Shrubs	Presence of shrubs
Clay	Percentage level of Clay in the soil	S. c.	Presence of Stipa comata
Silt	Percentage level of Silt in the soil	S. v.	Presence of Stipa viridulav
Sand	Percentage level of Sand in the soil	P. p.	Presence of Poa Pratensis
Texture	Soil texture in terms of Sandy, loamy	B. i.	Presence of Bromus inermis
Eco Region	Ecological region type where the site is	A. f.	Presence of Artemisia frigida
Slope	Indication of the existence of slopes	E. A.	Presence of Equisetum arvensis
Evaluation	Effectiveness of the control in the site	–	–

A data set made of 128 cases representing non - uniform cases from releases made in the Canadian provinces of Saskatchewan, Manitoba and Alberta was compiled.

Table 2: Sample Data & Logical Relationships

Combined
Siteis**GOOD**IFSize=LargeandSilt=MediumandTexture=SandyandShrubs=Yes Size=Large and Organic Carbon = Low and Texture = Non - Sandy and Relief = Flat and Shrubs = NoTexture=SandyandAspect=SoutherlyandRelief=ConvexandB. I=No Siteis**FAIL**IFSpan=twoyearsandEcologicalRegion=AspenandAspect=NorthP. p=No &SiltContent=LowandEcologicalRegion=AspenandAspect=NorthandP. p. =NoandB. i. =No
VegetationFactorsonly
Siteis**GOOD**IFShrubs=YesandP. P. =YesandA. f. =Yes Siteis**FAIL**IFShrubs=NoandS. c. =YesandS. v. =NoandA. f. =No
EcologicalFactorsonly
Siteis**GOOD**IFEcologicalRegion=Aspen&Aspect=Southerly& Relief=ConvexandSurface=Fullycovered Siteis**FAIL**IFEcologicalRegion=MixedGrassandShade=NoandSurface=Bare
PhysicalFactorsonly
Siteis**GOOD**IFPh=HighandSilt=MediumandTexture=Sandy Siteis**FAIL**IFOrganicCarbon=HighandPh=LowandClay=LowandTexture=non - sandy
Releasefactorsonly
Siteis**GOOD**IFSize=LargeandDate=Early Siteis**FAIL**IFSize=Small

As shown in Table 1 and 2, a list of features and sample data and relationships of the set of 81 cases were complete and non - redundant which are used for training and testing the system. The data collected included the amount of weed depression introduced by the control agent (expressed as the diameter of the control area) and density of the beetles' presence (expressed by the number of beetles in five sweeps). These two measurements were combined according to a formula:

$q = (diameter/2)^2 (number - of - beetles)$

The resulting q values were then mapped into one of GOOD or FAIL decision labels using predefined ranges given and reviewed by domain expert [10].

The values of the used attributes were re - modeled and made coarser by the merging of values resulting in a coarser value for the attributes Span, Size, Date, Organic - carbon, pH, Clay, and Sand. The collected data was grouped into the following:

- Release Factors: (date of release, span and size of the colony. ie Size, Date, and Span).
- Physical Factors: (Organic carbon, pH, Clay, Sand, Slit and Texture)
- Ecological Factors: (Ecological region, Relief, Aspect, Slope, Surface cover and Shade)
- Vegetation Factors: (This set included the occurrence of Shrubs, Bromus inermis (Bis), Stipa comata (S. c), Artemisia fridida (A. f), Poa pratensis (P. p) and Equisetum arvense (E. a).

The set of combined factors was also used. Table 2 shows the some of the values and rules used.

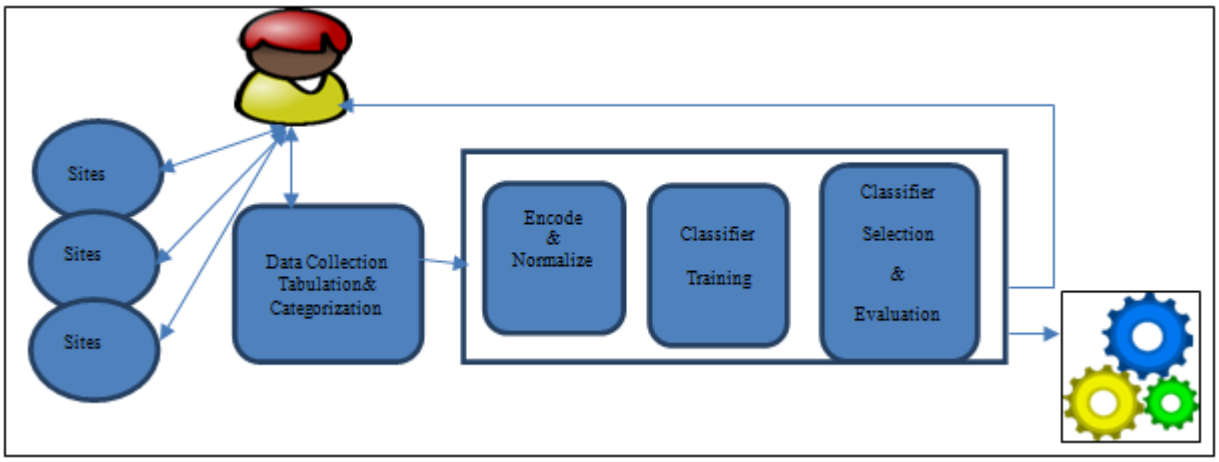

Figure 1: Overall Procedure Used

4. Suggested Procedure

A machine learning approach to the analysis of factors effecting the efficiency of weed reduction as well as those effecting the survival of the agent is introduced.

It helps in the prediction of suitability and potential of new sites as well as the success of ongoing sites. The methodology was used to analyse the available data as depicted in Figure 1, the adopted procedure was as follows:

Data Collection and Manual Preparations:
The data collection in terms of both features and values was done by scientists in Regina Agricultural Station on going bases. The collected data was then modelled into a table containing complete and non - redundant cases. Each case is

represented as a row of values representing a different factors of a single with last column being the final evaluation of success or failure. Of 128 reported data sites, 81 complete and non - redundant cases with a total of 23 attributes (features) including the evaluation feature were used.

Data Refinements:
Further manual refinement of the data into four smaller tables representing the different grouping of factors was performed as is illustrated in table 2 and 3 above.

Classifiers Set Up and Applications:
A number of machine learning classifier algorithms have been applied to the data as a whole as well as on different sub groups of factors. The used algorithms including Random Frost (RF), Nearest Neighbour (KNN), Support Vector Machines (SVM), Logistic Regression (LR), Neural Nets (ANN) and Bayes' Classifier (NB) with variable degrees of accuracy. Mathews correlation coefficient (MCC) was calculated and used along with overall accuracy of prediction to analysis the results and select the best classifier.

For each group of factors including the complete data set and the set of classifiers adopted, a cross validation was performed. A number of important evaluation indicators were collected for the training and for the validation, including:

a) Training:

- Mean Accuracy (TA)
- Mean F1 Score (TF1)
- Mean Precision (TP)
- 'Mean Recall (TR)
- Mean Matthew's Correlation Coefficient (TMCC)

b) Validation:

- Mean Accuracy (VA)
- Mean F1 Score (VF1)
- Mean Precision (VP)
- Mean Recall (VR)
- Mean Matthew's Correlation Coefficient (VMCC)

For better and more reliable the Matthew's correlation coefficient (MCC) is used. It is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset.

5. Experiments, Results and Discussions

As is shown in results Table 3 and 4. The following set of experiments involving 6 different machine learning techniques were performed: (1) Release Aspects: includes the three factors of Size, Date, and Span; (2) Physical Aspects: includes Organic carbon, pH, Clay, Sand, Slit and Texture. (3) Ecological Aspects: which includes Ecological region, Relief, Aspect, Slope, Surface cover and Shade;

Table 3: Results of the different subsets of factors

Factors	Rates	GNB	LR	SVC	KN3	RF	NN	Factors	Rates Algs	GNB	LR	SVC	KN3	RF	NN
Release	**TA**	57.19	60.94	60.94	60.62	**60.94**	**60.94**	**Ecological**	**TA**	66.25	64.38	71.25	76.56	**87.19**	**87.19**
	TF1	0.599	0.702	0.702	0.738	0.702	0.702		**TF1**	0.732	0.720	0.789	0.813	0.893	**0.894**
	TMCC	0.157	**0.212**	**0.212**	**0.170**	**0.212**	**0.212**		**TMCC**	0.294	0.252	0.385	0.518	0.741	**0.742**
	VA	48.75	50	50	**55.00**	48.75	48.75		**VA**	38.75	41.25	37.5	43.75	36.25	**46.25**
	VF1	0.555	0.635	0.635	**0.709**	0.625	0.622		**VF1**	0.486	0.509	0.501	**0.564**	0.464	0.544
	VMCC	- 0.216	- 0.166	- 0.166	**- 0.091**	- 0.206	- 0.206		**VMCC**	- 0.305	- 0.260	- 0.374	- 0.206	- 0.358	**0.523**
Physical	**TA**	60	73.13	75.31	75.63	**80.31**	**80.31**	**Vegetation**	**TA**	58.44	61.88	66.88	66.56	**70.63**	70.62
	TF1	0.742	0.786	0.797	0.811	**0.831**	**0.831**		**TF1**	0.528	0.673	0.716	0.699	**0.734**	0.726
	TMCC	0.184	0.446	0.498	0.510	**0.601**	**0.601**		**TMCC**	0.250	0.205	0.360	0.335	0.440	**0.448**
	VA	60	**70**	63.75	66.25	67.5	70		**VA**	**55.00**	48.75	38.75	42.5	35	36.25
	VF1	0.742	**0.736**	0.642	0.710	0.678	0.721		**VF1**	0.409	**0.499**	0.410	0.387	0.372	0.377
	VMCC	0.117	**0.436**	0.325	0.337	0.397	0.431		**VMCC**	0.155	**- 0.09**	- 0.351	- 0.192	- 0.434	- 0.400

(4) Vegetation which includes occurrence of Shrubs, Bromus inermis (Bis), Stipa comata (S. c), Artemisia fridida (A. f), Poa pratensis (P. p) and Equisetum arvense (E. a); and the (5) Combined set.

Table 4: Results of the combined set of factors

	Rates	GNB	LR	SVC	KN3	RF	NN
Combined Factors	TA	64.062	80.937	80.312	81.875	100	100
	TF1	0.6937	0.8378	0.8468	0.8448	1	1
	TMCC	0.3117	0.6082	0.6080	0.6297	1	1
	VA	53.75	42.5	42.5	40	38.75	38.75
	VF1	0.6463	0.4124	0.4817	0.4942	0.4023	0.4219
	VMCC	- 0.061	- 0.178	- 0.197	- 0.245	- 0.270	- 0.258

Table 5 summarizes the results. Best results are obtained using the combined set with 22 features.

Table 5: Summary of best results

Factors	Rates	Accuracy		MCC	
		Classifier	Rate	Classifier	Rate
Release	Training	RF	61	RF	.29
	Validation	KN3	55	KN3	.70
Physical	Training	RF/NN	80	RF/NN	.60
	Validation	LR	70	LR	.43
Ecological	Training	RF/NN	87	RF/NN	.74
	Validation	NN	46	KN3	.56
Vegetation	Training	RF/NN	71	NN	.45
	Validation	GNB	55	GNB	.16
Combined	Training	RF/NN	100	RF/NN	1
	Validation	GNB	54	GNB	- .661

A 100% training accuracy obtained using Random Frost or Artificial Neural Network and 100% MCC. In the validation as is usual is lower with 54% accuracy obtained by GNB. Obtained MCC is poor however. For the subsets of factors, best results are obtained by ecological factors of 87% training accuracy obtained by Random Forest and Artificial Neural Network with very good MCC of.74. Validation accuracy was 46% with.56 MCC. Other sets of factors have shown reasonable results reaching 80% training accuracy with.6 MCC and validation accuracy of 55% with.7 MCC.

6. Conclusion and Recommendation

Use of chemical and cultural method for the weed control on pasture lands is expensive. It has negative consequences on ground water, environment and health in general. A safer and more cost effective alternative is biocontrol of weeds in which harmful and unwanted grass, weeds in general. Weeds affect crops directly and indirectly.

Leafy spurge is an invasive alien weed native to central and southern Europe that have spread across western Canada and North America with milky liquid that causes severe skin rashes or irritation in in livestock and humans. The weed has been targeted by beetles from the flea beetle genera Aphothona as biocontrol since they were introduced into Canada in the 1980s.
The A. n. agent and its effectiveness is determined by the interaction of a variety of factors that merit understanding of their relationships and effects on weed and agent.

A machine learning approach to the analysis and prediction potential of such factors was used to analyse the available data taken from Regina Agriculture Station in order to provide scientists the ability to predict the suitability of sites and the success of the agent before the release of the beetle. A number of machine learning classifier algorithms have been tested and applied to the data including Random Frost, Nearest Neighbour, Support Vector Machines (SVMs), Artificial Neural Networks and Logistic Regression with variable degrees of accuracy. Based on Matthews correlation coefficient (MCC) and the overall accuracy of prediction, the best classifier were Random Forest and Artificial Neural Networks. It is worth note, however that even though the data was very limited, the results were encouraging especially when considering some of the subsets have reasonable accuracies. The importance of such an observation stems from the fact that such subsets such as Ecological factors are much easier to observe and collect data on.

References

[1] Huffaker, C. B., P. S. Messenger, Theory and practice of biological control, 481 - 496: Academic Press Inc, 1976.
[2] Julien, M. H., Griffiths, M. W. (eds.): Biological Control of Weeds: A World Catalogue of Agents and Their Target Weeds, 4th edn. CABI Publishing and the Australian Centre for International Agricultural Research, Antony Rowe, Chippenham, United Kingdom (1999).
[3] Harris, P.: Biological control of weeds. In: Franz, J. M. (ed.) Biological Plant and Health Protection. Fortschr. Zool., vol.32, pp.123–138 (1986).
[4] Schwarzländer M, Hinz HL, Winston RL, Day MD. Biological control of weeds: an analysis of introductions, rates of establishment and estimates of success, worldwide. BioControl.2018 Jun; 63 (3): 319 - 31.
[5] Day M, Witt AB. Weed biological control: Challenges and opportunities. Weeds - Journal of the Asian - Pacific Weed Science Society.2019 Dec 31; 1 (2): 34 - 44.
[6] Gassman, A.: Aphthona nigriscutis Foudras (Coleoptera: Chrysomelidae): a candidate for the biological control of cypress spurge and leafy spurge in North America. Final screening report. International Institute of Biological Control, Delmont, Switzerland (1985).
[7] Elhadi M. T., An experimental analysis of rough sets theory as applied to biological control of weeds. M. Sc. Thesis, Department of Computer Science, University of Regina.1991
[8] Mirtaheri SL, Shahbazian R. Machine Learning: Theory to Applications. CRC Press; 2022 Sep 29.
[9] Biau G, Scornet E. A random forest guided tour. Test.2016 Jun; 25 (2): 197 - 227.
[10] Kaur G, Oberai EN. A review article on Naive Bayes classifier with various smoothing techniques. International Journal of Computer Science and Mobile Computing.2014 Oct; 3 (10): 864 - 8.
[11] Wang QQ, Yu SC, Qi X, Hu YH, Zheng WJ, Shi JX, Yao HY. Overview of logistic regression model analysis and application. Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine].2019 Sep 1; 53 (9): 955 - 60.
[12] Taunk K, De S, Verma S, Swetapadma A. A brief review of nearest neighbor algorithm for learning and classification. In2019 International Conference on Intelligent Computing and Control Systems (ICCS) 2019 May 15 (pp.1255 - 1260). IEEE.
[13] Brereton RG, Lloyd GR. Support vector machines for classification and regression. Analyst.2010; 135 (2): 230 - 67.
[14] Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H. State - of - the - art in artificial neural network applications: A survey. Heliyon.2018 Nov 1; 4 (11): e00938.
[15] Amer, K. (2005). Manager - physician relationships: An organizational theory perspective. The Health Care Manager, 24 (2), 165 - 181.
[16] Amaratunga, D., Baldry, D., Sarshar, M., & Newton, R. (2002). Quantitative and qualitative research in the built environment: application of mixed research approach. Work study, 51 (1), 17 - 31.
[17] Asika, N. (2005). Research methodology in behavioral sciences: Lagos: Longman Publishing.
[18] Bright, D. (2011). Leadership for quality improvement in primary care groups (doctoral thesis), The Heller School for Social Policy and Management, Brandeis University, USA.
[19] Creswell, J. W., & Junior, M. (2012). Qualitative inquiry & research design: choosing among five approaches (4th ed.). Thousand Oaks, CA: Sage.