Journal of Theoretical and Applied Information Technology
15th November 2019. Vol.97. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EFFECT OF CLUSTERING DATA IN IMPROVING MACHINE LEARNING MODEL ACCURACY

SAMIH M. MOSTAFA[1], HIROFUMI AMANO[2]
[1]Mathematics Department, Faculty of Science, South Valley University, Qena, Egypt
[2]Research Institute for Information Technology, Kyushu University, Japan
E-mail: [1] samih_montser@sci.svu.edu.eg, [2] amano@cc.kyushu-u.ac.jp

ABSTRACT

Supervised machine learning algorithms consider the relationship between dependent and independent variables rather than the relationship between the instances. Machine learning algorithms try to learn the relationship between the input and output from the historical data in order to attain precise predictions about unseen future. Conventional foretelling algorithms are usually based on a model learned and trained from historical data. The instances in the historical data may vary in its characteristics. The variation may be a result of difference in case's pertinence degree to some cases compared to others. However, the problem with such machine learning algorithms is their dealing with the whole data without considering this variation. This paper presents a novel technique to the trained model to improve the prediction accuracy. The proposed method clusters the data using K-means clustering algorithm, and then applies the prediction algorithm to every cluster. The value of K which gives the highest accuracy is selected. The authors performed comparative study of the proposed technique and popular prediction methods namely Linear Regression, Ridge, Lasso, and Elastic. On analysing on five datasets with different sizes and different number of clusters, it was observed that the accuracy of the proposed technique is better from the point of view of Root Mean Square Error (RMSE), and coefficient of determination (R^2).

Keywords: *Prediction accuracy, K-means, clustering, regression, machine learning algorithms.*

1. INTRODUCTION

Machine learning (ML) is considered as the important subfield of artificial intelligence and is being adopted for numerous of various applications [1, 2]. ML addresses the study and construction of models capable of learning from the data. Understanding *what* and *how* the ML algorithm is learning is an issue for the developers of the ML applications [3]. ML can be classified into unsupervised and supervised. *Unsupervised* learning groups the data into categories depending on the basis of the similarities between data in each group. On the other hand, *supervised* learning means that the machine learns with the assistance of the labeled training data. Estimating unknown (independent, dependent) mapping of a system using a specific number of (independent, dependent) samples is called learning [4, 5]. This process of estimating needs data collected (i.e., training data), and an algorithm that deals with this data and learns from it. Generally speaking, the learning algorithm learns pattern in the data on hand and create a set of rules to map input/output relation. Data is categorized into labeled (with outcome) or unlabeled (without outcome). Outcome variable(s) may be continuous or distinct, regression is a way of predicting for continuous outcome, and classification is a way of predicting for distinct outcome (i.e., the response to be predicted is the probability or the true of an event/class), the number of classes can be two or more. On the other hand, clustering is applied to unlabeled data using the similarities between observations to group them into clusters [6–8]. The statistical method depends on the characteristics of the data (e.g., similarities between instances in the clustering technique), in other words, the more similarities the better statistical method accuracy. Regression is one of the most common statistical processes for estimating dependent/independent relationship when the dependent to be predicted is a continuous value. The regression line is a refined outline of averages and is drawn in such a way as to reduce the error of the fitted values in relation to the actual values. Equation of the simple linear regression can be defined by the following form:

$$\mathrm{d} = \mathrm{C}_0 + \mathrm{C}_1 \mathrm{I}$$

where C_0 is the intercept and the C_1 is the slope of the regression line. In addition to explaining the relationship between dependent d and independent I variables, the model also predicts the value of dependent variable from the independent variable values from the equation:

$$\hat{d} = \widehat{C_0} + \widehat{C_1}I$$

the hat symbol $\widehat{}$ refers to the predicted value of the unknown coefficient/variable. In simple regression, for a given dependent variable there is one independent variable. However, in real cases there is more than one independent variable, so existing of multiple independent variables is called multivariate regression. The mathematical notation can take the form:

$$d = C_1 I_1 + C_2 I_2 + C_3 I_3 + \cdots + C_m I_m$$

where m is the number of independent variables.

1.1. Goodness of the Model

The performance of the model can be evaluated using following metrics:

• R-squared: how well the model fits the data.

•Root Mean Square Error (RMSE): how close the estimated values are to the actual values.

1.2. Paper Contributions & Novelty

This paper gives a brief discussion of machine learning types, proposes a method for improving the prediction accuracy, and compares between the proposed method and the common methods. In contrast to other techniques, e.g., Clustering Lasso (CL) which selects groups of variables that have the same mechanism of predicting the dependent variable [9], the novelty of the proposed work boils down in benefiting from similarities between instances and applying the selected prediction algorithm for each cluster. The proposed work does not depend on collinearity among variables or the number of variables compared to instances.

1.3. Organization

The rest of this paper is organized as follows: Section 2 presents the proposed method supported with an illustrative example; dataset specification, the prediction algorithms used in the comparison, the parameters used in the comparison, results and discussion are discussed in section 3.

2. PROPOSED METHOD

In the proposed method, the authors avail from the similarity attribute of the data by clustering it into groups (clusters) before applying the statistical method. In the proposed method, the number of clusters is determined by the elbow method (heuristic method of validation and interpretation of symmetry within cluster analysis). The data is clustered using K-means algorithm in which the number of clusters resulted from the elbow method is used for the clustering. The selected prediction algorithm is applied for every cluster. For deeper clarification, the next subsection discusses an illustrative example.

Algorithm:	Proposed method
1.	Input data
2.	Select the prediction algorithm
3.	Find the value of K using elbow method
4.	Cluster the data using K-means
5.	Apply the selected prediction algorithm in each cluster

In the proposed method, clustering prepressing step is applied to the data before applying the prediction algorithm for the purpose of improving the accuracy of model generated by the user-selected prediction algorithm.

2.1 Illustrative Example

To clarify the proposed method, an artificial data with 600 observations has been generated, and linear regression algorithm is applied to the data (Figure 1). K-means algorithm is applied to clustering the data, the value of K is determined using elbow method (Figure 2), the data is clustered into three clusters (Figure 3), then linear regression algorithm is applied to the data in each cluster (Figure 4).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

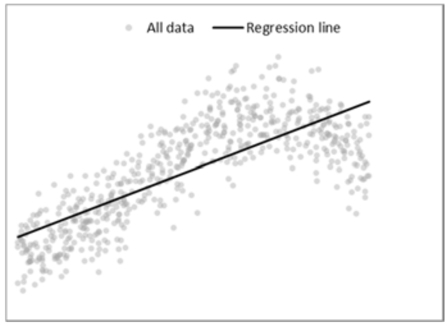

Figure 1. Regression line for all data

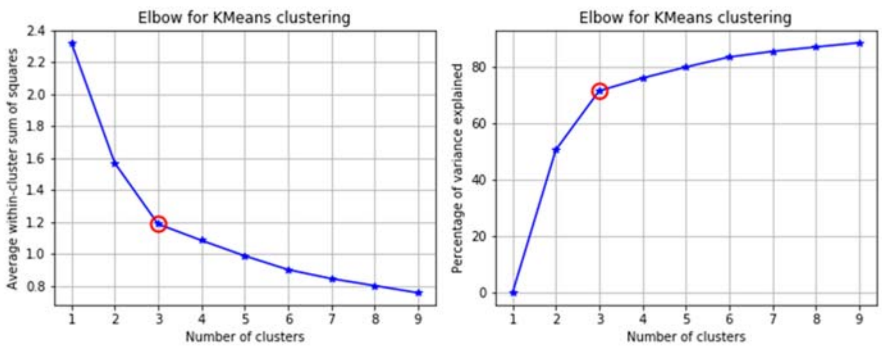

Figure 2. Determining the value of k using elbow method.

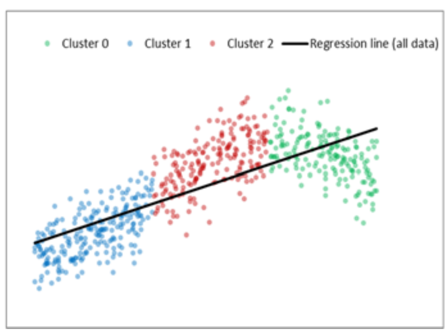

Figure 3. Three clusters of the data

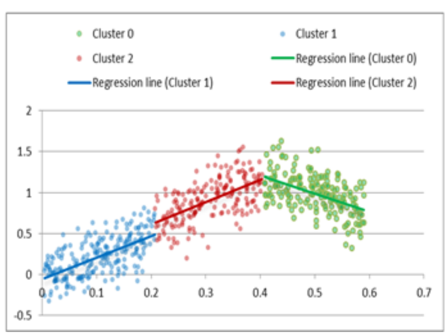

Figure 4. Three regression lines for the clusters

3. EXPERIMENTAL IMPLEMENTATION

3.1. Datasets

Six datasets that are commonly used in databases repository are used in the comparative study (Table 1).

3.2. Prediction Algorithms

The comparison is done between four common prediction algorithms namely multible linear regression (MLR), Ridge, Lasso, and ElasticNet and the proposed method. Table 2 gives short descriptions of the prediction algorithms used.

Table 1. Datasets specifications

Dataset name	#Instances	#Features	References
Diabetes	442	11	[10]
Graduate Admissions	500	8	[11]
California	20640	9	[12]
Diamonds	53940	10	[13]
Boston	506	12	[14]
Iris	150	6	[15]

Table 2. Packages and functions used

Prediction method	Short description	References
MLR	minimizes the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation.	[16]
Ridge	solves a regression model where the loss function is the linear least squares function, and imposes a penalty on the size of the coefficients.	[17]
Lasso	estimates sparse coefficients. Coefficients that add lightweight value to the model will be zero	[18]
ElasticNet	allows for learning a sparse model where few of the weights are non-zero like Lasso, while still maintaining the regularization properties of Ridge.	[19]

3.3. Performance Measure

The comparisons were done from the point of view of the following parameters:

- **Root mean squared error (RMSE):** indicates how close the forecasted values are to the actual values; hence the lower value of RMSE, the good of the model performance [20]. The mathematical notation can be written as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}}$$

where $y_i$and $\hat{y_i}$ are the actual value and forecasted value of the *i-th* observation respectively, and *n* is the number cases.

- **Coefficient of determination (R^2 score):** it is a measure of how perfectly the evaluated regression line of the model adapts the data distribution [21]. It can be written as:

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2};$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

3.4. Experimental Results and Discussion

The experiments are conducted on a computer equipped with 16 GB of RAM, Intel core i5-2400 (3.10 GHz), 1 TB of HDD, Gnu/Linux Fedora 28 of OS, and Python (version 3.7) of programming language. Table 3 summarizes the observations by comparing the improvements of the proposed approach versus common four algorithms from the point of view of R^2 score and RMSE. The notable observations are:

- Clustered MLR is better than MLR in R^2 score and RMSE for all datasets.
- Clustered Ridge is better than Ridge in R^2 score and RMSE for all datasets.
- Clustered Lasso is better than Lasso for Iris, Diamond, and Diabetes, and behaves

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

somewhat similar to Lasso for Admission, Boston, and California.

- Clustered ElasticNet is better than ElasticNet for Iris, Diamond, and Diabetes, and behaves somewhat similar to ElasticNet for Admission, Boston, and California

Table 4 shows the comparisons between the four common methods and the proposed method from the point of view of R^2 score and RMSE.

Table 3. Proposed approach versus MLR, Ridge, Lasso, and ElasticNet

	MLR		Ridge		Lasso		ElasticNet	
dataset	**R^2 score**	**RMSE**	**R^2 score**	**RMSE**	**R^2 score**	**RMSE**	**R^2 score**	**RMSE**
Admission	better	better	better	better	worst	worst	worst	same
Boston	better	better	better	better	better	better	same	same
California	better	better	better	better	worst	worst	worst	worst
Diabetes	better	better	better	better	better	better	better	better
Diamond	better	better	better	better	better	better	better	better
Iris	better	better	better	better	better	better	better	better

4. Conclusion and Future Works

This paper introduces a method that aims to improve the prediction accuracy of the model by clustering the data and applying the selected algorithm, which is a user choice, for each cluster. Unlike the traditional supervised algorithms which find the relationship between the dependent and independent variables, the proposed approach benefits from the similarities between the instances to improve the prediction accuracy. Four common algorithms are compared with the proposed method, the results showed that the proposed method achieves significant improvement from the point of view of RMSE, and coefficient of determination R^2. In the future research avenues, the proposed approach will be analysed in more dataset, other standard error metrics will be considered (e.g., P-value and T-value).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4. Comparisons between common algorithms and the proposed

			Clustered MLR		MLR		Clustered Ridge		Ridge	
dataset	**Cluster ID**	**# obs. per cluster**	**R^2 score**	**RMSE**	**R^2 score**	**RMSE**	**R^2 score**	**RMSE**	**R^2 score**	**RMSE**
Admission (500 obs.)	0	80	0.900587	0.042692	0.903082	0.042153	0.896687	0.043522	0.90232	0.042319
	1	85	0.764728	0.066934	0.7314	0.071518	0.772732	0.065785	0.734592	0.071091
	2	84	0.900393	0.032677	0.772022	0.049436	0.90909	0.031218	0.771618	0.04948
	3	83	0.87173	0.035889	0.876285	0.035246	0.869495	0.0362	0.875802	0.035314
	4	82	0.914561	0.040976	0.917142	0.040353	0.906431	0.042882	0.915482	0.040755
	5	86	0.88446	0.063127	0.853827	0.071004	0.888193	0.062099	0.853866	0.070994
	Average		**0.872743286**	**0.047049204**	**0.842293064**	**0.051618251**	**0.873771396**	**0.046950917**	**0.842280112**	**0.051658971**
	Improvement				**0.036151576**	**0.088516127**			**0.037388136**	**0.091137201**
Boston (506 obs.)	0	185	0.815499	3.809681	0.759997	4.345084	0.804651	3.920084	0.755214	4.388164
	1	137	0.60816	4.940371	0.329145	6.464268	0.599721	4.993285	0.307886	6.565894
	2	184	0.798881	4.067129	0.776842	4.284174	0.807025	3.983925	0.774684	4.304843
	Average		**0.740846565**	**4.272393928**	**0.621994708**	**5.03117545**	**0.737132497**	**4.299097932**	**0.612594663**	**5.086300222**
	Improvement				**0.19108178**	**0.150815953**			**0.203295656**	**0.154769136**
California (20640 obs.)	0	7208	0.665958	0.601686	0.627722	0.635189	0.666001	0.601648	0.627689	0.635217
	1	11339	0.642995	0.730971	0.60065	0.773106	0.642993	0.730973	0.600663	0.773095
	2	234	0.753532	0.448398	0.705991	0.489738	0.72469	0.473908	0.705947	0.489775
	3	1859	0.649944	0.581809	0.630402	0.597829	0.649204	0.582424	0.630355	0.597867
	Average		**0.678107166**	**0.590716165**	**0.641191351**	**0.623965652**	**0.67072195**	**0.597238209**	**0.641163447**	**0.623988367**
	Improvement				**0.057573789**	**0.053287367**			**0.046101354**	**0.042869642**
Diabetes (442 obs.)	0	180	0.244554	38.00049	-0.58888	55.11027	0.179101	39.6125	-1.15089	64.1204
	1	262	0.092901	34.33915	-1.30065	54.68739	0.10635	34.08364	-1.21911	53.70962
	Average		**0.168727508**	**36.16982113**	**-0.9447616**	**54.89883005**	**0.142725619**	**36.84807093**	**-1.18500069**	**58.9150139**
	Improvement				**1.178592683**	**0.341154974**			**1.12044349**	**0.374555508**
Diamond (53940 obs.)	0	32963	0.878649	302.7003	0.330429	711.032	0.878381	303.0345	0.329419	711.5678
	1	5654	0.404659	1986.213	-0.51074	3164.012	0.404285	1986.838	-0.5104	3163.653
	2	15323	0.592287	1007.742	0.291297	1328.63	0.592706	1007.224	0.291769	1328.188
	Average		**0.625198361**	**1098.885094**	**0.036994739**	**1734.558121**	**0.625123609**	**1099.032344**	**0.036929956**	**1734.469421**

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	Improvement				**15.89965597**	**0.366475484**			**15.92727726**	**0.366358189**
Iris (150 obs.)	0	53	0.585892	0.159824	0.538373	0.168745	0.609486	0.155204	0.580346	0.160891
	1	97	0.505393	0.246851	0.519073	0.243414	0.509306	0.245873	0.516603	0.244038
	Average		**0.545642513**	**0.203337645**	**0.528722806**	**0.206079331**	**0.559396339**	**0.200538486**	**0.548474215**	**0.202464357**
	Improvement				**0.032001092**	**0.013304032**			**0.019913651**	**0.009512148**

dataset	Cluster ID	# obs. per cluster	Clustered Lasso R² score	Clustered Lasso RMSE	Lasso R² score	Lasso RMSE	Clustered ElasticNet R² score	Clustered ElasticNet RMSE	ElasticNet R² score	ElasticNet RMSE
Admission (500 obs.)	0	80	0.397248	0.105123	0.103192	0.128227	0.624298825	0.082995	0.47687026	0.097934
	1	85	0.418975	0.105186	0.218946	0.121955	0.558337418	0.091708	0.500034743	0.097573
	2	84	-0.01393	0.104257	0.247015	0.089845	0.590989869	0.066217	0.636564741	0.062419
	3	83	-0.04289	0.102333	0.284816	0.084743	0.383689678	0.078667	0.595595155	0.063724
	4	82	0.217802	0.123983	0.295194	0.11769	0.65317612	0.082558	0.64568584	0.083445
	5	86	0.259622	0.159799	0.29872	0.155523	0.649727827	0.109914	0.66747207	0.107093
	Average		**0.206137182**	**0.116780231**	**0.241313579**	**0.116330496**	**0.57670329**	**0.085342992**	**0.587037135**	**0.085364642**
	Improvement				**-0.14577048**	**-0.00386601**			**-0.017603393**	**0.000253615**
Boston (506 obs.)	0	185	0.504211	6.245079	0.631276	5.385678	0.555916205	5.910471	0.635021837	5.358251
	1	137	0.491277	5.629189	0.470952	5.740537	0.489438342	5.639353	0.473984885	5.724061
	2	184	0.788513	4.170639	0.642887	5.419558	0.705213571	4.923962	0.656423965	5.315845
	Average		**0.594667172**	**5.348302053**	**0.581704998**	**5.515257787**	**0.583522706**	**5.491261589**	**0.588476895**	**5.466052459**
	Improvement				**0.022283071**	**0.03027161**			**-0.008418664**	**-0.00461194**
California (20640 obs.)	0	7208	0.194511	0.934328	0.284208	0.880771	0.398249064	0.807566	0.422555231	0.791088
	1	11339	0.347412	0.988285	0.296732	1.025943	0.450214811	0.907108	0.436138572	0.918647
	2	234	-0.00836	0.906966	0.315368	0.747329	0.320412819	0.744571	0.486811721	0.647026
	3	1859	0.063351	0.951701	0.284806	0.831619	0.325355782	0.807699	0.420265792	0.748732
	Average		**0.149228941**	**0.945319905**	**0.29527854**	**0.871415265**	**0.373558119**	**0.816735984**	**0.441442829**	**0.77637343**
	Improvement				**-0.49461637**	**-0.0848099**			**-0.153779165**	**-0.05198858**

Diabetes (442 obs.)	0	180	0.016449	43.35967	-1.50228	69.16003	0.000281698	43.71458	-3.415714253	91.87302
	1	262	-0.02383	36.48171	-1.62935	58.46368	-0.023699919	36.47947	-2.644023327	68.82604
	Average		**-0.00368829**	**39.92068967**	**-1.56581185**	**63.81185592**	**-0.011709111**	**40.09702716**	**-3.02986879**	**80.34953014**
	Improvement				**0.997644487**	**0.374400116**			**0.99613544**	**0.500967497**
Diamond (53940 obs.)	0	32963	0.875739	306.3083	0.325622	713.5795	0.743478863	440.1009	-0.320155658	998.397
	1	5654	0.404575	1986.353	-0.50916	3162.36	0.253176236	2224.601	-2.206191315	4609.33
	2	15323	0.593488	1006.257	0.293033	1327.002	0.402987813	1219.449	0.416165349	1205.916
	Average		**0.624600687**	**1099.639291**	**0.036496857**	**1734.313961**	**0.466547637**	**1294.716959**	**-0.703393874**	**2271.214188**
	Improvement				**16.1138214**	**0.365951428**			**1.663280779**	**0.42994502**
Iris (150 obs.)	0	53	-0.03211	0.252318	-6.74515	0.691194	-0.032111237	0.252318	-1.973883919	0.428299
	1	97	-0.19781	0.384149	-0.18902	0.382737	-0.197813953	0.384149	0.367240443	0.279206
	Average		**-0.1149626**	**0.318233686**	**-3.46708467**	**0.53696541**	**-0.114962595**	**0.318233686**	**-0.803321738**	**0.353752519**
	Improvement				**0.966841711**	**0.407347886**			**0.85689097**	**0.10040588**

REFERENCES:

[1] T. Segaran, Programming Collective Intelligence: Building Smart Web 2.0 Applications. 2007.

[2] D. Vallejo-Huanga, P. Morillo, and C. Ferri, “Semi-Supervised Clustering Algorithms for Grouping Scientific Articles,” Procedia Comput. Sci., vol. 108, pp. 325–34, 2017.

[3] V. K. Ayyadevara, Pro Machine Learning Algorithms. 2018.

[4] V. Cherkassky and F. M. Mulier, Learning from Data : Concepts, Theory, and Methods / V. Cherkassky, F. Mulier. 1998.

[5] W. A. Yousef and S. Kundu, “Learning algorithms may perform worse with increasing training set size: Algorithm-data incompatibility,” Comput. Stat. Data Anal., vol. 74, pp. 181–97, 2014.

[6] K. Chen and L. Liu, “The ‘Best K’ for Entropy-based Categorical Data Clustering.,” 2005, pp. 253–62.

[7] M. N. Murty, P. J. Flynn, and A. K. Jain, “Data Clustering: A Review,” ACM Comput. Surv., vol. 31, p. 43, 1999.

[8] A. K. Jain and R. C. Dubes, “Algorithms for Clustering data.” 1988.

[9] Q. Yu and B. Li, “Regularization and Estimation in Regression with Cluster Variables,” Open J. Stat., vol. 04, no. 10, pp. 814–25, 2014.

[10] N. D. of Statistics, “Diabetes Data-1-5-2019,” *NCSU Department of Statistics*. [Online]. Available: https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html. [Accessed: 01-May-2019].

[11] M. S. Acharya, “Graduate Admissions-1-5-2019,” *Kaggle Inc*, 2018. [Online]. Available: https://www.kaggle.com/mohansacharya/graduate-admissions. [Accessed: 01-May-2019].

[12] C. Nugent, “California Housing Prices-3-5-2019,” *Kaggle Inc*, 2017. [Online]. Available: https://www.kaggle.com/camnugent/california-housing-prices. [Accessed: 03-May-2019].

[13] M. Shiva, “Diamonds-20-4-2019,” *Kaggle Inc*, 2017. [Online]. Available:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

https://www.kaggle.com/shivam2503/diamonds. [Accessed: 20-Apr-2019].

[14] I. Kaggle, "Boston Housing-25-4-2019." [Online]. Available: https://www.kaggle.com/c/boston-housing. [Accessed: 25-Apr-2019].

[15] I. Kaggle, "Iris Dataset-25-4-2019." [Online]. Available: https://www.kaggle.com/jchen2186/machine-learning-with-iris-dataset. [Accessed: 25-Apr-2019].

[16] D. W. Gareth James Trevor Hastie, Robert Tibshirani, An introduction to statistical learning : with applications in R. New York : Springer, [2013] ©2013.

[17] Scikit-learn, "Ridgre Regression-26-4-2019." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html. [Accessed: 26-Apr-2019].

[18] Scikit-learn, "Lasso Regression-27-4-2019." [Online]. Available: https://scikit-learn.org/stable/modules/linear_model.html. [Accessed: 27-Apr-2019].

[19] Scikit-learn, "Elastic Net Regression-28-4-2019." [Online]. Available: https://scikit-learn.org/stable/modules/linear_model.html#elastic-net. [Accessed: 28-Apr-2019].

[20] I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," Inf. Sci. (Ny)., vol. 233, pp. 25–35, 2013.

[21] T. Beysolow II, Introduction to Deep Learning Using R. Apress, 2017.