AJMAS
https://alqalam.utripoli.edu.ly/science/ eISSN 2707-7179

Original article

Effect of Selecting Validation Dataset on Building Random Forest and Decision Tree Models

Mona Mohammed, Reem Alsunosi*

Department of Computer Science, Faculty of Sciences, University of Omar Almukhtar, Albaida, Libya

ARTICLE INFO

Corresponding Email. *reem.abdalhadi@omu.edu.ly*

Received: *11-09-2022* ***Accepted:*** *20-09-2022* ***Published:*** *26-09-2022*

Keywords*: Training, Validation and Testing Data, Train-Test Data Split, Random Forest (RF), Decision Tree (DT).*

ABSTRACT

Background and aims. *Machine learning models are trained using appropriate learning algorithm and training data. The dataset partition into training and testing data, the training data were used by the model to learn, and the testing data used by the model to predict on unseen data which will evaluate model performance. The train-test split procedure was used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. Machine learning models in production needs a lot more than just creating and validating models, Data validation are used to check that the model can return useful predictions in a real-world. The basic aim of this paper was to take a closer and critical look at the training data split methods to build the best models, and point out its weakness and limitation, especially for evaluating and comparing the performance of random forest and decision tree models.* ***Methods****. For this purpose, the experiments were carried out with different combinations of training and validation data which explain the effect of the method of selecting validation dataset in random forest and decision tree models performance for both classification and regression problems. Moreover, the experiments were going on testing the effect of increasing the training data size.* ***Results.*** *Classification tasks 60/40 ratio for training, and validation splits optimal for big data sets and 80/20 ratio for training, and validation splits optimal for small data sets in most experiments. In regression tasks the models performance increased as fold size increased in most cross-validation experiments.* ***Conclusion.*** *Performance of Random Forest classification, Decision Trees classification, Random Forest regression and Decision Trees regression under different ratios train/validation split better than the performance using cross-validation.*

Cite this article. *Mohammed M, Alsunosi R. Effect of Selecting Validation Dataset on Building Random Forest and Decision Tree Models. Alq J Med App Sci. 2022;5(2):470-478.*https://doi.org/10.5281/zenodo.7113928

INTRODUCTION

Machine learning is the process of developing artificial intelligence in computers to build models with high prediction [1,2]. By thinking, it is meant that it adapts to new data sample and discover hidden patterns without actually programming with high ability to give reliable outcomes, and improve their performance extracted knowledge [1-6]. Regression and classifier are common statistical tools for prediction [2].

Machine learning models are training using appropriate learning algorithm and training data. learning algorithm are able to adaptively improve their performance with each new data sample and discover hidden patterns. [6]. However, the data are divided into two or more subsets, typically, with a two-part split. The first part is commonly used to fit the machine learning model and is referred to as the training dataset, and the second subset used to verify if the model works correctly, this second dataset are referred to as the test of dataset [1,7].

In most of the applications simple random sampling is used [1,4,7]. The randomized or cross-validated split of training and testing sets have been adopted as the gold standard of machine learning for decades [7]. Training and testing data on the same portion of data do not give us an accurate view of how models perform. Data validation mean checking the accuracy and quality of data source before training a new models version [8-10]. Generally, the term validation set is used to give an estimate of model skill before training a new model version. The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set [8-10].

This paper was aimed to evaluate the influence of different training and validation data combinations on the performance machine learning algorithms or models. However, Random Forest (RF) and Decision Trees (DT) which will build using various methods for division training data splitting with different size of dataset.

METHODS

This paper is organized in five sections as follow: first, random forest, decision tree and performance evaluation criteria are described in methods used section. A brief overview of data splitting introduced in section 3, the fourth section reviews the data set using in the experiment and then all experimental work is captured in section 5. Section 6 contains analysis of the discussed data splitting methods and experimental results with a short discussion. Finally, section 7 makes conclusions.

Random Forest

Random Forest(RF) is a type of ensemble learning method which widely considered to be a one of the most accurate models. RF generates many binary recursive partitioning trees by selecting subsets of the given dataset and selecting subsets of predictor variables randomly, finally aggregating the results of all models to form a powerful model [11,12]. RF has been widely introduced to solve regression and classification tasks problems. It fits a number of DT on different samples and takes their majority vote for classification which predict categorical variables and average in case of regression problems which predict continuity [13,14].

Decision Tree

Decision Trees (DT) are one of the easiest tools to decision systems and easy to understand, it is powerful and popular tools to build Classification and regression models [3,11,15-17]. DT is a non-parametric method named according to the nature of target variable. It is named a classification tree if the target variable is categorical and a regression tree if the target variable is continuous[11,16]. DT is a simple and intuitive machine learning method uses a tree structure which provides sequential nonlinear analysis in algorithmic relationship and their possible consequences, including outcomes. DT consists of three types of nodes: root node, internal nodes, and leaf nodes [18,19]. It built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values which break down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes [16,20]. The split of a node attempts to minimize the impurity of the node. If a split is unable to achieve any improvement in terms of reducing impurity, the node is not split and is declared as a leaf node. If a split is able to reduce impurity, then the split providing the maximum reduction in impurity is selected and two branches are formed, forming two new nodes [11,21].

Performance Evaluation Criteria

For the performance assessment of classification tasks, a confusion matrixes were used to show the number of correct and incorrect which were identified by the classification model compared to the actual target value in the dataset. The matrix is NxN, where N is the number of target classes and each positive (Target) and negative (Non-Target) as following in table (1).

Table 1. Classification models evaluation (Confusion Matrix)

Confusion Matrix		**Target**			
		Positive	Negative		
Model	Positive	A	B	*Positive Predictive Value*	a/(a+b)
	Negative	C	D	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy**=(a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

For the performance assessment of regression tasks, standard statistical measures namely Mean Absolute Error (MAE) were used to compare and validate the performance of both random forest and decision tree regression. The Mean Absolute Error is a popular formula to measure the error rate of a regression model However, it can be calculated by comparing between the actual values from testing data set and predicted values from the models under different training and testing ratios.

$$MAE = \frac{\sum_{i=1}^{n} |p_i - a_i|}{n}$$

Train-Test Split

Machine learning models are trained using appropriate learning algorithm and dataset. Machine learning models should observe and learn from the training set, optimizing any of its parameters. The testing set acts as an evaluation of the final mode and compared against the previous sets of data. Here the data are divided into two parts, training data and testing data [1,22,23]. Machine Learning performance depends significantly on the quality of data and the strategy of splitting the dataset [22-25]..however, validation Data are used to check that the model can return useful predictions in a real-world, which used to give an estimate of model skill before training a new model version[8-10].

The used Data set

To perform the study, two fully available standard machines learning datasets which are the same size were used. All experimental tests investigate two datasets, the first one is Melbourne House Prices dataset [26]. The dataset involves the house price given details of the house's, it is standard machine learning dataset consisting of 13580 examples with 19 features (excluding the street address of houses) and a numerical target variable. However, this dataset will be used as a regression problem. The second one which will be used for classification tasks are part of Adult Income dataset [27] which have total of 14 features and 13580 observations. The examples in each dataset are marked with labels with two classes values.

Experiments

The experiments, have been conducted four classification tasks and four regression tasks. Then analyzed the results to show how the models performance and stability will vary with the different combinations of training and validation data with examining the effect of increasing the dataset size. However, RF and DT models will learn to perform a both classification and regression tasks.

Train- validation Split for Classification task

After the pre-analysis of the dataset, started to demonstrate how train- validation split impact in RF Classification algorithm and DT Classification algorithm on the different size of Adult Income dataset. For this purpose, the experiments were divided into two parts as following:

Experiment 1: First, examine the effect of increasing the number of training datasets in different sizes of dataset. For this purpose, the performance of RF classification and DT classification will be evaluated under different ratios that were used to divide the datasets into the training and validation datasets (90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80, and 10:90) train/ validation split. In all of these training- validation set pairs, RF and DT classification performance was fitted and then calculated the prediction performance in quarter of dataset (3395 examples), half of dataset (6790 examples), three quarters of dataset (10185 examples), and all dataset (13580 examples). Finally, the estimation of the target value on different ratio-based training and validation datasets using a confusion matrix will be evaluated.

Experiment2:It examined cross-validation, and run RF classification and DT classification models on different subsets of the data to get multiple measures of models quality. the dividing the data into 3,4,5,6,7,8,9,10,11 pieces, then, run one experiment for each fold. First, the use of the first fold as a validation set and everything else for training the model this gives a measure of model quality. After that hold out data from the second fold to get a second estimate of model quality, and use everything except the second fold as training set. the repeat of this process, using every fold once as the validation set. Then putting this together, and end up with a measure of model quality that is based on 100% of the dataset. Typically, wanted a single measure of model quality so take the average across experiments. However, in all training-validation set pairs, both RF classification and DT classification models performance will fit and then calculated the prediction performance of performance in quarter of dataset (3395 examples), half of dataset (6790 examples), three quarters of dataset (10185 examples), and all dataset (13580 examples). lastly, the estimation of the target value by models will evaluated using a confusion matrix.

Train-Validation Split for Regression task

After the pre-analysis of the dataset, started to a demonstrate how to use the train- validation split to evaluate RF and DT regression algorithms on Melbourne House Prices dataset. For this purpose, the experiments are divided into two parts as following:

Experiment 3: examine the effect of increasing the number of training datasets. For this purpose, the performance of RF regression and DT regression will evaluated under different ratios were used to divide the datasets into the training and validation datasets (90:10, 80:20, 70:30, 60:40, 50:50, 40:60, 30:70, 20:80, and 10:90) train/ validation split. In all of these training-validation set pairs, RF and DT regression performance was fitted and then calculated the prediction performance in quarter of dataset(3395 examples), half of dataset(6790 examples), three quarters of dataset(10185 examples), and all dataset(13580 examples) . Finally, the estimation of the target value on different ratio-based training and validation datasets will evaluated. in this case, select Mean Absolute Error(MAE).

Experiment4: examine cross-validation, run RF regression and DT regression models on different subsets of the data to get multiple measures of models quality. Dividing the data into 3,4,5,6,7,8,9,10,11 pieces, then, run one experiment for each fold. First, use the first fold as a validation set and everything else for training the model this gives us a measure of model quality. After that hold out data from the second fold to get a second estimate of model quality, and use everything except the second fold as training set. Repeat this process, using every fold once as the validation set. Then putting this together, and end up with a measure of model quality that is based on 100% of the dataset. Typically, want a single measure of model quality so take the average across experiments. However, in all training-validation set pairs, both RF regression and DT regression models performance will fit and then calculated the prediction performance in quarter of dataset (3395 examples), half of dataset (6790 examples), three quarters of dataset(10185 examples), and all dataset(13580 examples). Finally, the estimation of the target value by models will evaluated using (MAE).

RESULTS AND DISSECTION

In the previous section, conducted two classification tasks and another two regression tasks. Models were trained in the training set and tested in the validation set, followed by estimating the model performance. To that end, modeling was repeated many times, with different starting datasets. In this section, the prediction results of the experiments are presented. However, all statistical analysis done using Python3.8.3.

First, the results of train-validation split for classification tasks for experiment 1 The performance of RF classification and DT classification models under different ratios train/validation split in quarter of dataset, half of dataset, three quarters of dataset, and all data set were evaluated as shown in table(2).

Table 2. The summery and comparison results of RF and DT classification model under various ratios train/validation split

		DecisionTreeClassification				**RandomForestClassification**			
		Quarter of dataset	Half of dataset	Three quarter	All of data set	Quarter of dataset	Half of dataset	Three quarter	All of dataset
Validation dataset	Training dataset	accuracy score				accuracy score			
10%	90%	0.854	0.808	0.819	0.828	0.857	0.845	0.866	0.874
20%	80%	0.859	0.817	0.829	0.823	0.852	0.842	0.882	0.87
30%	70%	0.847	0.813	0.825	0.824	0.85	0.843	0.887	0.869
40%	60%	0.835	0.82	0.826	0.823	0.85	0.851	0.888	0.874
50%	50%	0.835	0.821	0.824	0.822	0.859	0.851	0.875	0.866
60%	40%	0.84	0.818	0.826	0.823	0.874	0.863	0.871	0.865
70%	30%	0.83	0.808	0.828	0.818	0.879	0.867	0.863	0.864
80%	20%	0.822	0.8	0.821	0.819	0.866	0.853	0.859	0.861
90%	10%	0.811	0.791	0.812	0.811	0.845	0.84	0.858	0.858

Figure1 illustrates the prediction results of RF classification model, it showed that the model performance that was changed under different combinations of training and validation data. It can be seen that, in the small size data as the number of data in the validation datasets increased, the accuracy score increased in most of ratios train/validation split until 70:30 train/validation split which are the best accuracy score. On the other hand, the accuracy score decreased in most of ratios train/validation split when number of data in the validation datasets decreased, in the big dataset size.

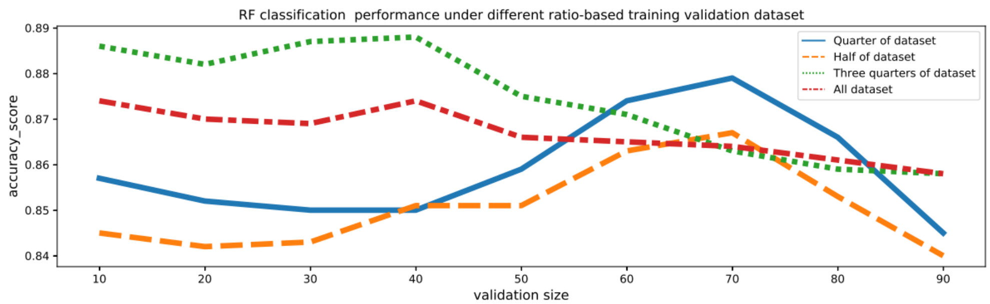

Figure 1. RF classification model under various ratios train/validation split.

As shown in figure2,DT model performance decreased in most of ratios train/validation split in the different data size when number of data in the validation datasets increased.

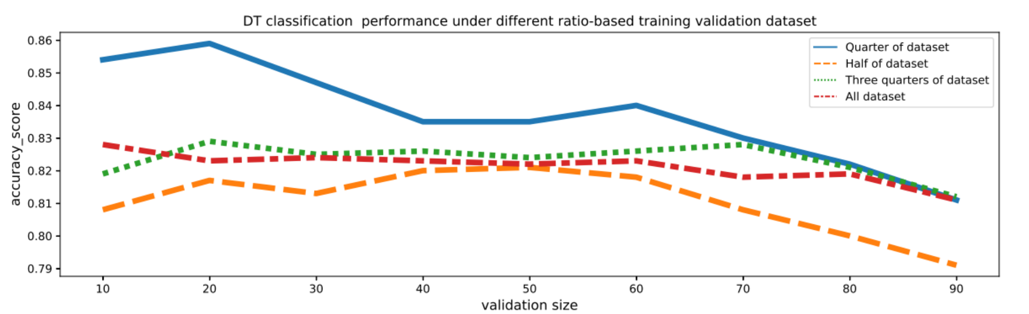

Figure 2. DT classification model under various ratios train/validation split

For experiment2, results presented in table(3) which shows a comparison results of RF classification and DT classification models under different fold size in quarter of dataset, half of dataset, three quarters of dataset, and all dataset.

Table 3: The performance of RF and DT classification model under various fold size

	DecisionTreeClassification				**RandomForestClassification**			
	Quarter of dataset	Half of dataset	Three quarter	All of data set	Quarter of dataset	Half of dataset	Three quarter	All of dataset
Fold size	accuracy score				accuracy score			
3	66.11	74.9	73.82	76.44	68.62	77	78.01	80.56
4	65.94	72.33	70.94	74.55	68.67	77.02	74.32	78.54
5	65.15	72.98	74.26	73.8	70.77	76.6	78.19	77.7
6	70.55	70.72	72.37	75.71	73.66	74.66	77.45	80.29
7	73.32	73.26	73.04	75.71	76.55	77.6	78.74	79.75
8	73.16	74.25	72.14	74.14	75.37	77.61	78.06	78.04
9	74.35	73.55	72.57	74.35	78.28	77.53	78.29	80.51
10	74.41	74.3	73.22	73.41	77.29	78.46	79.03	79.39
11	75.58	75.22	72.79	74.04	79.42	78.71	79	79.92

Figure3 shows that the prediction results of RF classification model showed that the model performance was increased as the fold size increased in most of quarter of dataset, half of dataset, three quarters of dataset. moreover, the model was oscillating performance in the full dataset.

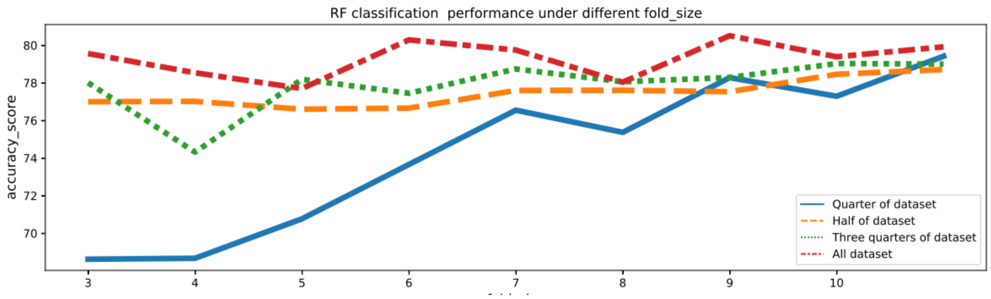

Figure 3. RF classification model under various fold size

Depending on figure4 which showed the prediction results of DT classification, model performance increased as the fold size increased in quarter of dataset which present small size dataset. In addition to that, the best model performance was when fold size7each pieces 14.28% of the full dataset.

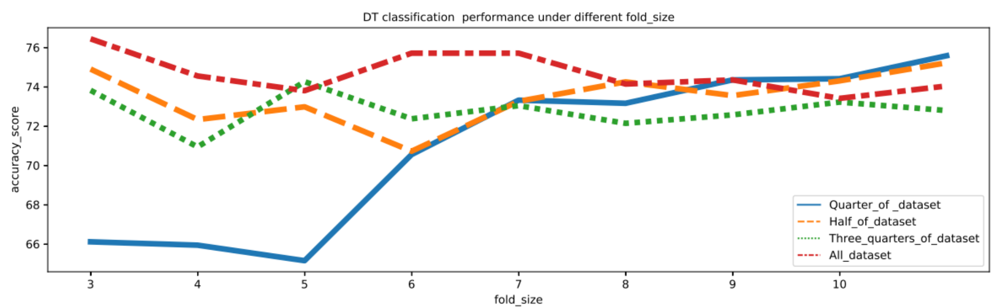

Figure 4. DT classification model under various fold size

Second, the results of train-validation split for regression tasks in table (4) show the summarize and compare the results of experiment3 which presented the performance of RF regression and DT regression models under different ratios train/validation split in quarter of dataset, half of dataset, three quarters of dataset, and all dataset the errors in contrast.

Table 4. The summery and comparison results of RF and DT regression model under various ratios train/validation split

		DecisionTreeRegressor				**RandomForestRegressor**			
		Quarter of dataset	Half of Dataset	Three quarter	All of data set	Quarter of dataset	Half of dataset	Three quarter	All of dataset
Validation dataset	Training dataset	mean_absolute_error(MAE)				mean_absolute_error(MAE)			
10%	90%	292,654.61	248,623.17	281,197.93	285,104.11	203,082	191,256	194,617	207,656
20%	80%	295,380.48	278,136.21	260,092.68	250,714.38	204,021	200,539	185,744	193,585
30%	70%	291,143.54	273,922.50	252,313.92	263,708.27	205,878	199,217	193,282	190,066
40%	60%	303,734.19	296,324.74	255,439.75	249,821.68	212,406	203,405	193,069	183,082
50%	50%	310,438.24	295,621.97	261,990.40	263,386.38	219,207	207,953	193,390	189,364
60%	40%	307,022.46	276,825.50	277,074.57	262,788.58	226,854	206,215	196,108	190,966
70%	30%	321,567.84	279,324.18	288,660.95	265,207.44	237,070	212,539	198,834	191,531
80%	20%	324,135.75	300,592.57	282,992.95	279,935.65	248,116	218,738	204,743	200,054
90%	10%	343,930.90	325,513.72	307,647.31	290,731.83	250,048	235,571	226,527	220,540

From figure 5, it has been noticed that as the number of data in the validation datasets increased, the errors (MAE) of RF regression increased in all ratios train/validation split. In addition, the model performance decreased as the number of data in the validation datasets increased in10 to 40%validation dataset ratio in the big data size.

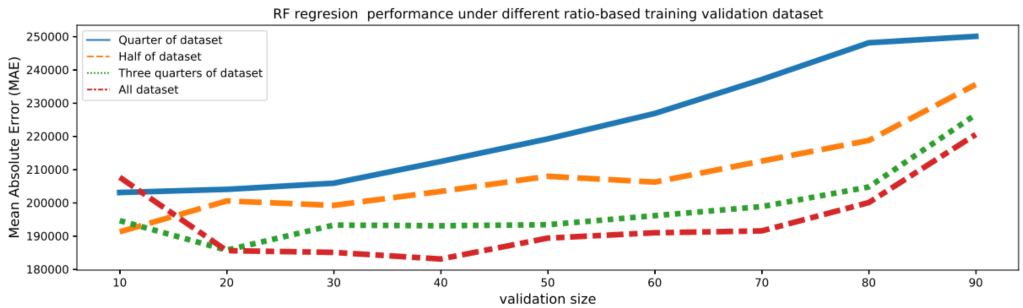

Figure 5. RF regression model under various ratios train/validation split

Furthermore, figure 6 shows DT classification performance decreased as the number of data in the training datasets increased in almost of ratios train/validation split. In addition, model performance decreased as the number of data in the training datasets increased in 90:10 train/validation split in the big data size. Lastly, in this case both RF regression and DT regression models gave the nearest results.

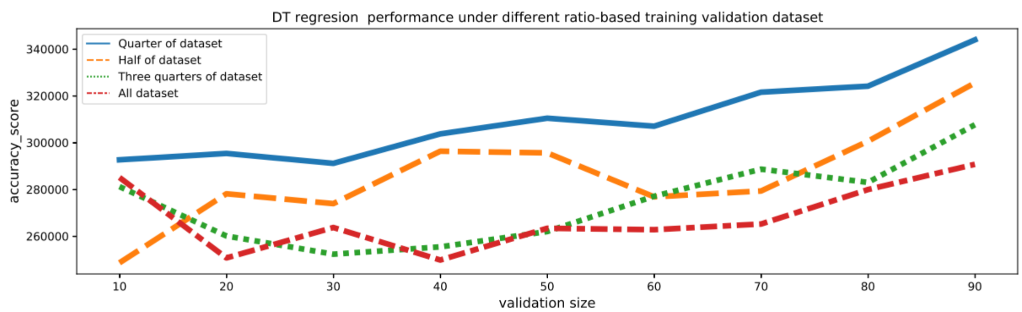

Figure 6. DT regression mode under various ratios train/validation split

Experiment4 results are shown in table (5) which summarized the performance of RF regression and DT regression models under various fold size in quarter of dataset, half of dataset, three quarters of dataset, and all dataset.

Table 5. The performance of RF and DT regression model under various fold size

	DecisionTreeRegressor				**RandomForestRegressor**			
	Quarter of dataset	Half of dataset	Three quarter	All of data set	Quarter of dataset	Half of dataset	Three quarter	All of dataset
Fold size	mean_absolute_error(MAE)				mean_absolute_error(MAE)			
3	405,042	314,590	277,907	271,930	297,231	227,434	206,918	201,012
4	381,263	301,084	283,360	263,597	279,259	213,114	206,545	190,369
5	359,484	299,824	277,245	248,721	267,972	214,010	204,653	187,959
6	356,563	303,551	274,744	252,705	267,608	218,266	202,856	188,369
7	341,101	294,822	275,107	250,798	255,543	214,399	202,380	185,348
8	338,567	297,127	271,767	255,955	252,020	213,298	198,704	184,642
9	349,813	297,242	271,854	251,274	259,784	215,337	198,442	185,905
10	348,289	288,383	272,035	252,648	249,740	213,589	197,950	185,943
11	344,329	291,170	272,871	252,553	258,098	212,446	197,607	189,538

As shown in figure 7, the errors (MAE) of RF regression decreased when the fold size increased in most of quarter of dataset however it showed slight decrease in half of dataset, three quarters of dataset and full dataset.

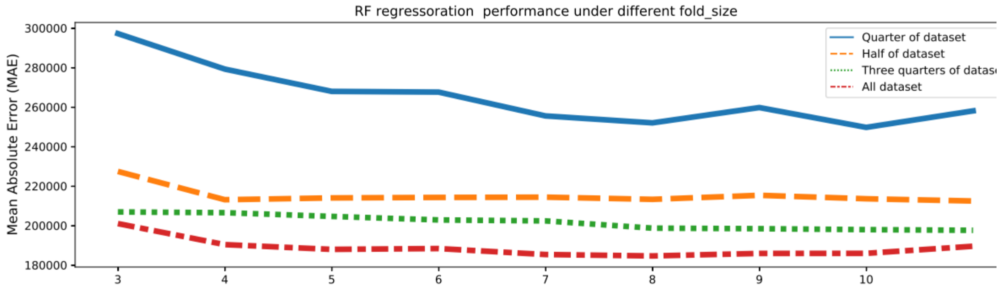

Figure 7. RF regression model under various fold size

As shown in figure 8, the errors (MAE) of DT regression decreased when the fold size increased in most of quarter of dataset. However, it showed slight decrease in half of dataset, three quarters of dataset. Moreover, both RF regression and DT regression models gave the nearest results in this case.

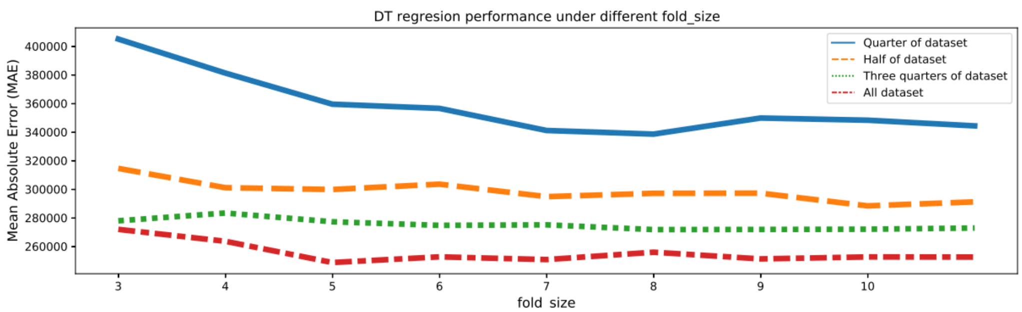

Figure 8. DT regression model under various fold size

CONCLUSIONS

In general, the results demonstrated that RF classification, DT classification, RF regression, and DT regression performance under different ratios train/validation split better than the performance using cross-validation. However, in classification tasks 60/40 ratio for training, and validation splits optimal for big data sets and 80/20 ratio for training, and validation splits optimal for small data sets in most experiments. In regression tasks the models performance increased as fold size increased in most cross-validation experiments.

Disclaimer

The article has not been previously presented or published, and is not part of a thesis project.

Conflict of Interest

There are no financial, personal, or professional conflicts of interest to declare.

REFERENCES

1. Nguyen QH, Hai-Bang Ly, Lanh Si Ho, Al-Ansari N, Hiep Van Le, Van Quan Tran, et al. Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. Mathematical Problems in Engineering.2021; 2021(6): 1-15.

2. Smith PF, Ganesh S, Liu PA. comparison of random forest regression and multiple linear regression for prediction in neuroscience. Journal of Neuroscience Methods. 2013; 220(1): 85-91.
3. RakhraM,SoniyaP,TanwarD, Singh P,BordoloiD,AgarwalP,VermaN. Crop Price Prediction Using Random Forest and Decision Tree Regression: A Review. SciencedirectJournals& Books, Materials Today: Proceedings, 2021.
4. ReitermanovaZ. Data Splitting. Proceedings of Contributed Papers. In: WDS. Prague: Matfyzpress. 2010;10: 31–36.
5. Basştanlar Y, Özuysal M. Introduction to Machine Learning. MicroRNA Biology and Computational Analysis, Methods in Molecular Biology. 2014; 1107: 105-128.
6. Mitchell TM. The Discipline of Machine Learning. Machine Learning Department technical report CMU-ML-06-108. Carnegie Mellon University 2006: 1-7.
7. Tan J, Yang J, Wu S, Chen G, Zhao J. critical look at the current train/test split in machine learning. arXiv preprint arXiv:2106.04525. 2021; 1-21
8. Vabalas A, Gowen E, Poliakoff E, Casson A. J. Machine learning algorithm validation with a limited sample size. PLoS ONE. 2019;14(11): 1-20
9. Schelter S, Lange D, Schmidt P, Celikel M, Viessmann F. and Grafberger A. Automating Large-Scale Data Quality Verification. Proceedings of the VLDB endowment. 2018; 11(12):1781–1794.
10. Breck E, Zinkevich M, Polyzotis N, Whang S. and Roy S. Data validation for machine learning. In Proceedings of the 2nd SysML Conference, Palo Alto, CA, USA, 2019.
11. Sazonau V. Implementation and Evaluation of a Random Forest Machine Learning Algorithm.University of Manchester, Oxford Road, Manchester, M13 9PL, UK, 2012.
12. Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 2012; 9(5): 272-278.
13. Smith PF, Ganesh S, Liu PA. comparison of random forest regression and multiple linear regression for prediction in neuroscience. Journal of Neuroscience Methods. 2013; 220(1): 85-91.
14. Esmaily H, Tayefi M, Doosti H, Ghayour-Mobarhan M, Nezami H, Amirabadizadeh A. A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes ", Journal of Research in Health Sciences. 2018;18(2): 412.
15. Prajwala T. A Comparative Study on Decision Tree and Random Forest Using R Tool. International Journal of Advanced Research in Computer and Communication Engineering. 2015; 4(1): 196-199.
16. Caffrey R. Using the Decision Tree (DT) to Help Scientists Navigate the Access to Space (ATS) Options. In: 2022 IEEE Aerospace Conference Proceedings. 2022.
17. Pranckevičius T, Marcinkevičius V. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic Journal of Modern Computing. 2017; 5(2): 221-232.
18. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27(8): 861-874.
19. Tayefi M, Tajfard M, Saffar S, Hanachi P, Amirabadizadeh AR, Esmaeily H et al. hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. Computer methods and programs in biomedicine. 2017;141:105–109.
20. Elhazmi A, Al-Omari A, Sallam H, Mufti H, Rabie A, Alshahrani M, Arabi Y. M.Machine learning decision tree algorithm role for predicting mortality in critically ill adult COVID-19 patients admitted to the ICU. Journal of Infection and Public Health. 2022; 15(7): 826-834.
21. Brijain M, Patel R Kushik, MR Kushik, Rana K A. Survey on Decision Tree Algorithm for Classification. International Journal of Engineering Development and Research, 2014; 2(1): 1-5.
22. May R, Maier H, Dandy G. Data splitting for artificial neural networks using SOM-based stratified sampling. Neural Netw. 2010;23(2): 283–294.
23. Xu Y, Goodacre R. A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Journal of Analysis and Testing. 2018;2(3): 249–262.
24. Ly H, Pham B, Le L, Le T, Le V,Asteris G, Estimation of axial load-carrying capacity of concrete-filled steel tubes using surrogate models. Neural Computing and Applications. 2021; 33(8): 3437–3458.
25. Asteris P, Apostolopoulou M, Armaghani D, Cavaleri L, Chountalas A, Guney D, et al. On the umetaheuristic models for the prediction of cement-metakaolin mortars compressive strength. Metaheuristic Computing and Applications. 2020; 1(1):63-99
26. The adult Income dataset. https://www.kaggle.com/datasets/johnolafenwa/us-census-data
27. The Melbourne House Prices dataset https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot.