Global Journal of Computer Sciences: Theory and Research

Volume 06, Issue 2, (2016) 26-35

www.gjcs.eu

Comparing prediction algorithms in disorganized data

Erkut Arican *, Department of Computer Engineering, Bahcesehir University, 34349 Besiktas, Istanbul, Turkey.
Adem Karahoca, Department of Software Engineering, Bahcesehir University, 34349 Besiktas, Istanbul, Turkey.

Suggested Citation:
Arıcan, E., & Karahoca, A. (2016). Comparing prediction algorithms in disorganized data. *Global Journal of Computer Sciences: Theory and Research. 6*(2), 26-35.

Received July 11, 2016; revised September 8,2016; accepted November 28, 2016;
Selection and peer review under responsibility of Assoc. Prof. Dr. Özcan Asilkan, Akdeniz University, Turkey.

Abstract

Real estate market is very effective in today's world but finding best price for house is a big problem. This problem creates a propose of this work. In this study, we try to compare and find best prediction algorithms on disorganized house data. Dataset was collected from real estate websites and three different regions selected for this experiment. KNN, KSTAR, Simple Linear Regression, Linear Regression, RBFNetwork and Decision Stump algorithms were used. This study shows us KStar and KNN algorithms are better than the other prediction algorithms for disorganized data.

Keywords: KNN, KSTAR, simple linear regression, linear regression, RBFNetwork, disorganized data, Real Estate, BFNetwork, Decision Stump.

* ADDRESS FOR CORRESPONDENCE: **Erkut Arican,** Department of Computer Engineering, Bahcesehir University, 34349 Besiktas, Istanbul, Turkey. *E-mail address*: erkut.arican@eng.bau.com / Tel.: +090-212-381-5664

1. Introduction

This study was performed on the house for sale data which are collect from websites. The KNN, KSTAR, Simple Linear Regression, Linear Regression, RBFNetwork and Decision Stump comparison algorithms used in this study. Each of these was run on the data set and compare each of results. As a result of these comparisons, KStar and KNN algorithm are better than each other.

Today, the real estate market is very effective but there is a problem in finding the house price. Therefore, people who want to sell their homes search similar house ads. Another option is a companies which is gives expertise information about home or real estate agents.

This study is doing pre-process the data for using Weka application for trying various algorithms.

3 different regions have been selected for this experiment. These regions are Besiktas and Bahcelievler in Istanbul and Cankaya in Ankara.

We researched the internet for the best and capable application for data collection from the real estate websites. You can found more detailed information in other chapter.

In literature search, many people used different algorithms in this area. In our data we use K Nearest Neighbor [1], KStar [2], Simple Linear Regression, Linear Regression simple algorithms moreover RBFNetwork [3] and Decision Stump [4] algorithms used. WEKA application was used for calculation using the data. We can give more information about this calculation in other chapter. Nowadays these algorithms used for clustering and estimation and literature search are support this.

In literature search, Arto Harra and Annika Kangas [5] 's study are similar to this one. In this similar research, they compare KNN algorithm and Linear Regression. They examine the data and problem in 3 different ways.

1. Increased Nonlinear Effect
2. Modelling and Test Data Effect to Balance and Last One
3. Model Assumption

On the simulated data set have been used, and using simple modelling problems. Both algorithm compared by the square root of the mean squared error and give a good result. When compared algorithms using this results, KNN algorithm less prejudiced than Linear algorithm. In this study, we compare the relative absolute error with our algorithms to using 3 datasets.

In the next chapter, there is a detail explanation of the data, algorithms and results of the Besiktas dataset.

2. Material and Methods

2.1. Problem Definition

In this study, we try to compare and find the best algorithm on the house data which has been collected over the internet.

2.2. Data

Data is taken from a real estate website Sahibinden.com [6] on the Internet Neighborhoods information is important so we take this information from real estate website Hurriyet Emlak [7] and we replaced per square meter the average prices to the neighborhood.

We use Visual Web Ripper [8] for data collection process program. Other similar programs examined and Visual Web Ripper is selected to ease of use and more functional.

In Besiktas data, corrupt data is minimal to others and we know too many neighborhood price per square meter.

In Bahcelievler data, corrupt data is more than Besiktas data we know many neighborhood price per square meter.

In Cankaya data, corrupt data is too much for other data and neighborhood price is average price for Cankaya.

You can see all properties of the Besiktas data in Figure 1. There is a specific range of values for each property.

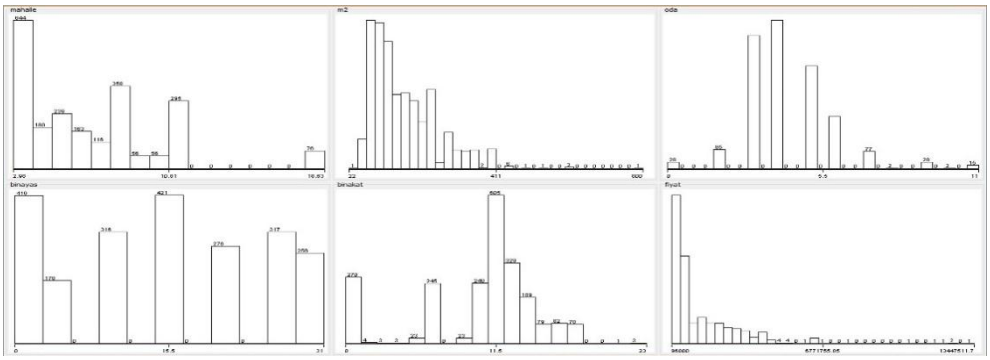
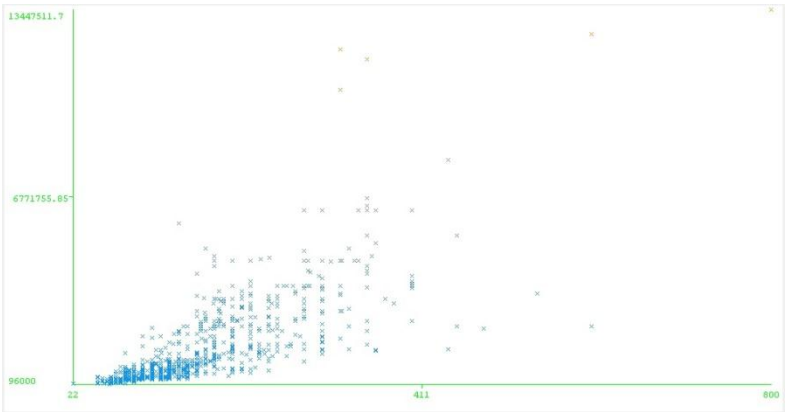


Figure 1. Besiktas Dataset Before Normalization

You can found neighborhood information, square meters of the house, number of rooms, age of the building, floor and price information in the data. As mentioned before, with our aim to find our



price column minimum number errors. After algorithm runs we found price and compare the knowing price. In Figure 2, there is almost a linear relationship between square meter and price without the normalization.

Figure 2. Square meter – Price Relation Before Normalization

Use the data in the columns, respectively, the neighborhood, the square meter, building age, floor and price. Detailed information on each described in Table 1.

Table 1. Describing Data

District	The average number of square meters of neighborhood information price
M2	Square meter of house
Room	Number of rooms
Age	Age of house
Floor	Floor of house
Price	Sale price of the house

Some algorithms do not work properly before the data normalization. For this reason, we were made a data normalization. In Figure 3, all the attributes of after the data normalization in Besiktas data and you can see the square meters and price relationship.

We can examine more detailed in experiment chapter.

2.3. Methods

2.3.1. K Nearest Neighbor & KStar

IB1 is a simple learning algorithm. Given the data set to finding the nearest neighbor using the Euclidean distance. If more than one finds the nearest first found is the selected. Euclidean distance is still looking IBk k neighbor to decide. [9]

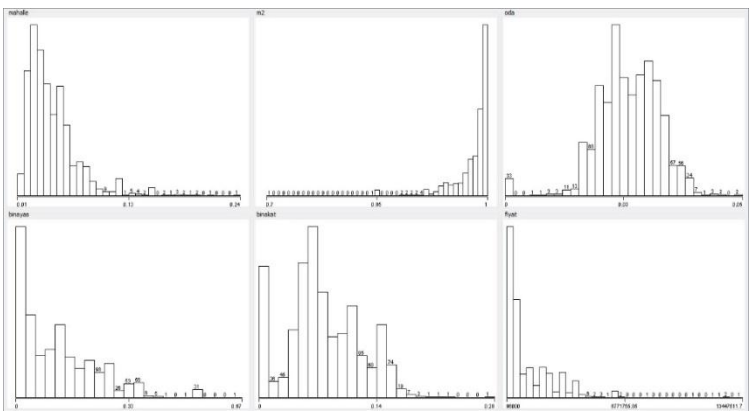


Figure 3. Besiktas Dataset After Normalization

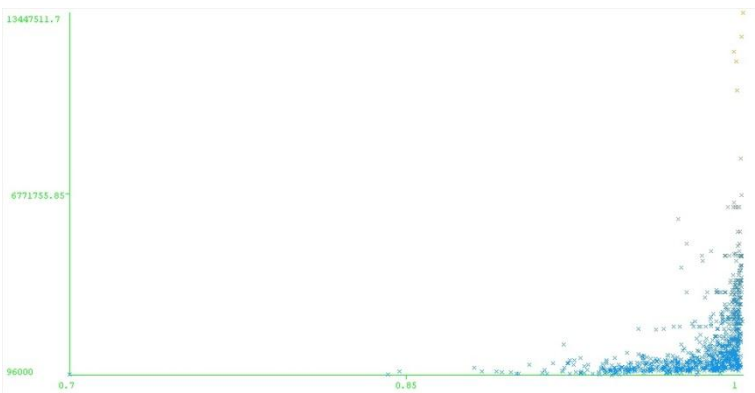


Figure 4. Square meter – Price After Normalization

2.3.2. Simple Linear Regression & Linear Regression

Linear regression analysis of the function of the parameters from the weight and distance is determined by the weight coefficient [9].

2.3.3. RBFNetwork

RBFNetwork, based on Gaussian radial basis function network. Widths of hidden units from the center, and use the k-means. If the data is hidden using a nominal logistic regression combining the outputs layers is obtained, if the linear regression that uses numerical [9].

2.3.4. Decision Stump

If using a simple one-level decision trees in two data problems for finding the result is referred to as Decision Stump [9].

3. Findings

The test results performed on all data sets. In this article, the best results Besiktas results indicated that the data set. After the normalization data set, each data – price relations were examined in WEKA. Between Figure 5 and in Figure 9 graphs can be seen in these relationships.

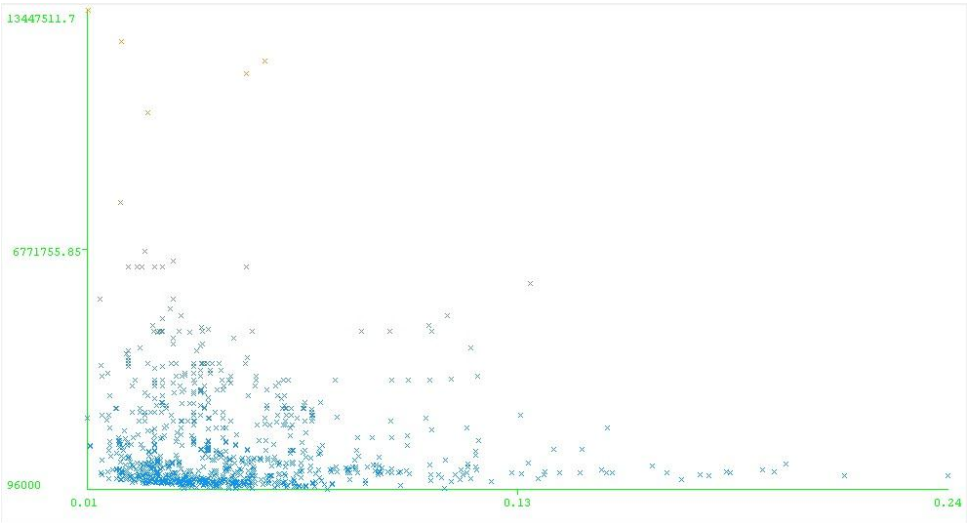


Figure 5. District - Price Relation

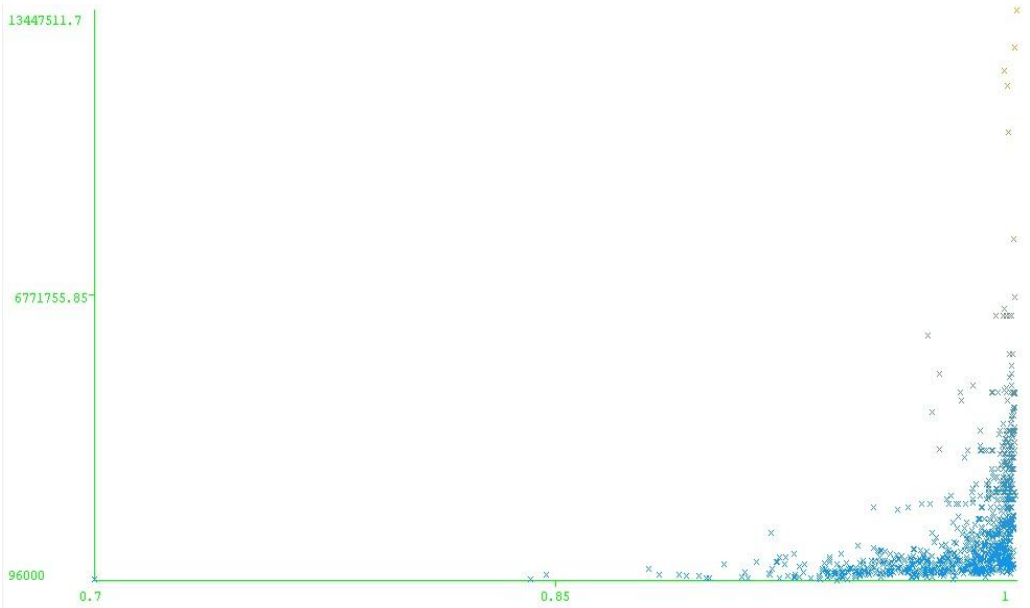


Figure 6. Square meter - Price Relation

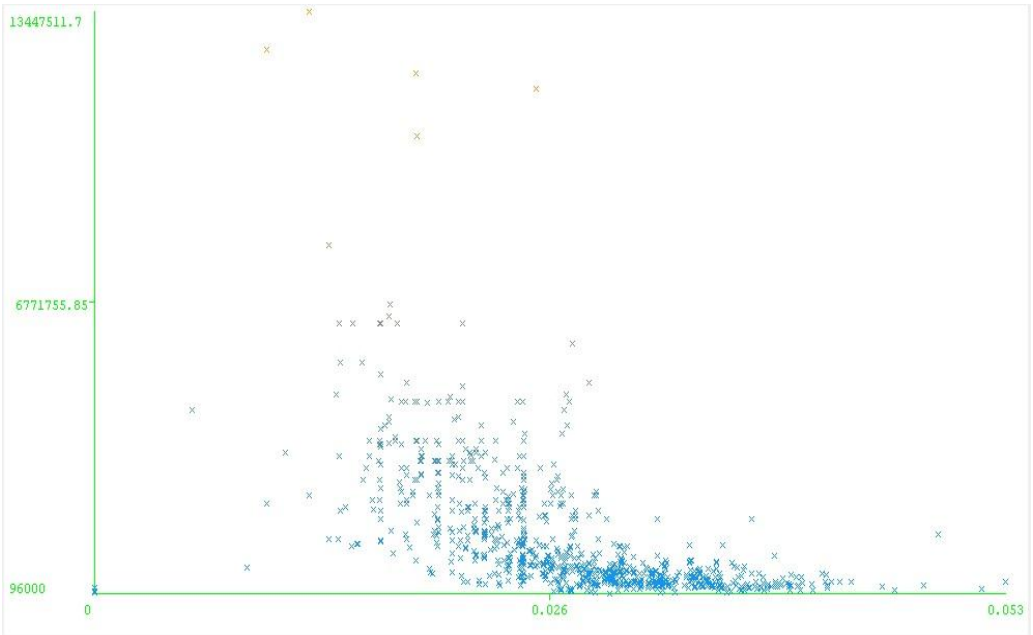


Figure 7. Room - Price Relation

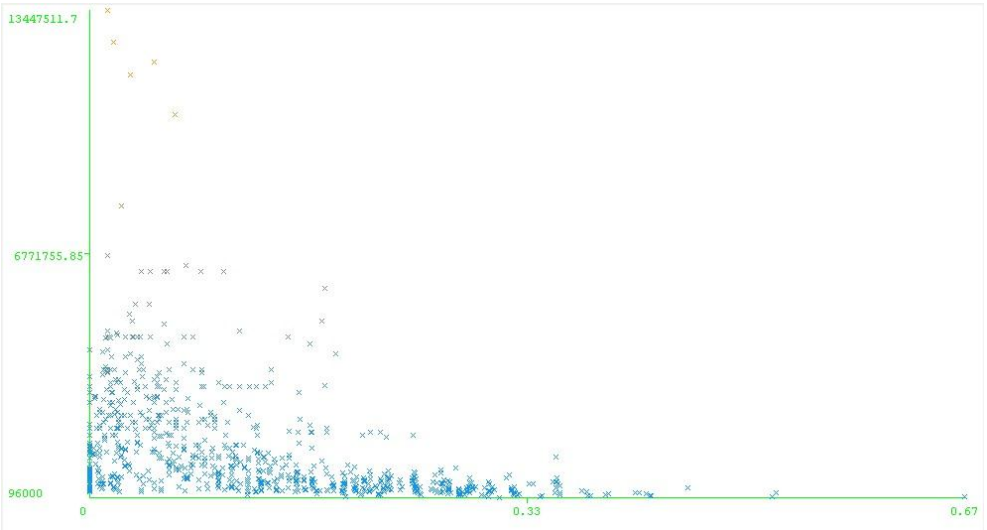


Figure 8. Age - Price Relation

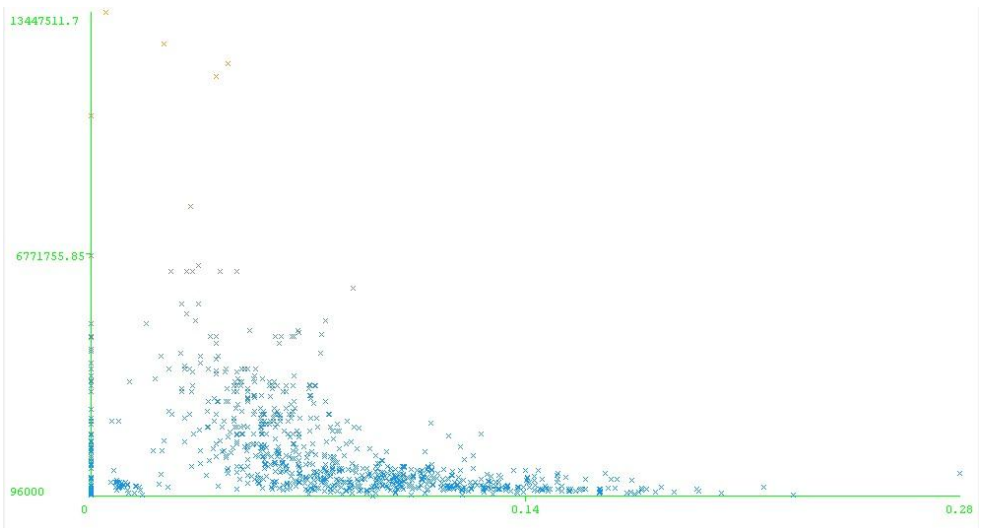


Figure 9. Floor - Price Relation

We are tested separately for each algorithm. When comparing the results between each other Relative Absolute Error (Relative Absolute Error) values were compared.

Each experiment 66% Training Set and the entire data (~ 100% training set) to use each of two experiments conducted and the results for the algorithm is specified separately.

Table 2. Simple Linear Regression

	%66 Training Set	~%100 Training Set
Correlation coefficient	0.5356	0.5265
Mean absolute error	814677.4996	781943.8101
Root mean squared error	1252716.7563	1192581.9984
Relative absolute error	74.7948 %	73.8725 %
Root relative squared error	84.5825 %	85.0182 %
Total Number of Instances	813	2392

In Table 2 and Table 3, we show the Simple Linear Regression and Linear Regression algorithms. There is not any difference between two columns and data size is not important. So in both algorithms errors are too much and these algorithms are failed.

Table 3. Linear Regression

	%66 Training Set	~%100 Training Set
Correlation coefficient	0.6713	0.6677
Mean absolute error	751067.6957	730101.6543
Root mean squared error	1100693.6992	1044276.1548
Relative absolute error	68.9548 %	68.9748 %
Root relative squared error	74.318 %	74.4456 %
Total Number of Instances	813	2392

In Table 4 and Table 5, we show the K Nearest Neighbor and KStar algorithms. These two table represent the both algorithms are better disorganized data. Especially KNN is more usable algorithm in disorganized data.

Table 4. KNN

	%66 Training Set	~%100 Training Set
Correlation coefficient	0.8365	0.9995
Mean absolute error	230799.2148	4663.5811
Root mean squared error	886002.0525	46290.4215
Relative absolute error	21.1895 %	0.4406 %
Root relative squared error	59.8222 %	3.3 %
Total Number of Instances	813	2392

Continue from the Table 4 and Table 5, data numbers are too important for these two algorithms. In 66% training set have got 813 data and in 100% training set have got a 2392 data. Algorithms which is looking the neighborhood, data number is important.

Table 5. KStar

	%66 Training Set	~%100 Training Set
Correlation coefficient	0.9004	0.989
Mean absolute error	197339.9616	59073.6541
Root mean squared error	644424.5592	210872.6515
Relative absolute error	18.1176 %	5.5809 %
Root relative squared error	43.5111 %	15.0329 %
Total Number of Instances	813	2392

In Table 6, RBFNetwork's errors are shown and these algorithm is not good enough for disorganized data.

Table 6. RBFNetwork

	%66 Training Set	~%100 Training Set
Correlation coefficient	0.4568	0.4699
Mean absolute error	858387.4507	818279.8421
Root mean squared error	1318251.4006	1238188.3816
Relative absolute error	78.8078 %	77.3053 %
Root relative squared error	89.0073 %	88.2694 %
Total Number of Instances	813	2392

Last algorithm shows in Table 7, this algorithm is better than RBFNetwork but much worse than KNN and KStar so Decision Stump is also not good for disorganized data.

Table 7. Decision Stump

	%66 Training Set	~%100 Training Set
Correlation coefficient	0.6412	0.6181
Mean absolute error	761457.8657	734558.2964
Root mean squared error	1142125.3307	1102699.3307
Relative absolute error	69.9087 %	69.3959 %
Root relative squared error	77.1154 %	78.6105 %
Total Number of Instances	813	2392

As we shown KNN algorithm is much better than the other algorithms. Before the normalization we do these test but result is not show and KNN is also have got a better error rates before normalization. In our opinion, KNN is better because in disorganized data it will look the neighborhood and decide it.

Also in this experiment, we learned the data must be fulfilled and regular. Data number is also important but regular data is much important than data size.

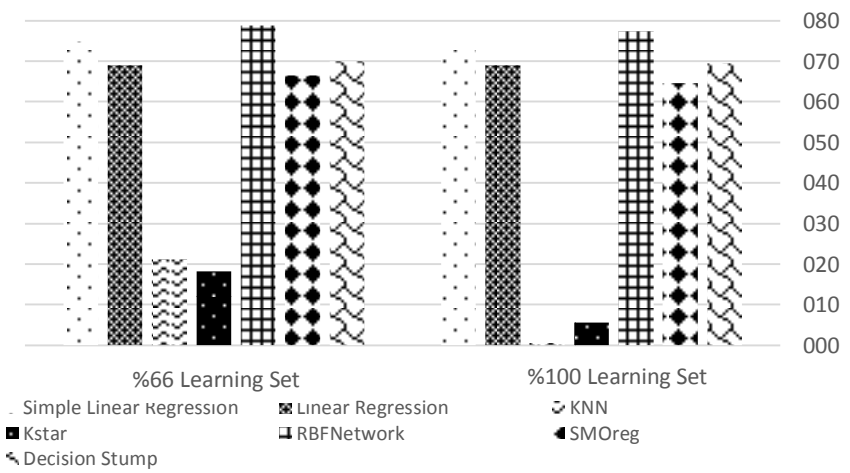


Figure 10. Comparison Results

As we seen in the results, K Nearest Neighbor algorithm comes first in disorganized data and in Figure 10 shows the comparison between all algorithms using these experiments. KNN errors rate much smaller than the others.

4. Conclusion

In this study, we compared some algorithms such as Simple Linear Regression, Linear Regression, KNN, KStar, RBFNetwork and Decision Stump. Dataset collected from the real estate websites. All findings show us KNN and KStar algorithms are better than the other algorithms on disorganized data.

References

[1] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician,* 1992, pp. 175-185.

[2] J. G. Cleary and L. E. Trigg. K*: An instance-based learner using an entropic distance measure. *Proceedings of the 12th International Conference on Machine learning,* 1995, pp. 108-114.

[3] D. S. Broomhead and D. Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks,1988.

[4] W. Iba and P. and Langley. Induction of One-Level Decision Trees. *Proceedings of the Ninth International Conference on Machine Learning,* 1992, p. 233–240.

[5] Haara, A. and A. S. Kangas. Comparing K nearest neighbours methods and linear regression – Is there reason to select one over the other? *MCFNS,* 2012, 4(1), pp. 50-65

[6] Sahibinden. Sahibinden.com. Available from: www.sahibinden.com.

[7] H. Emlak. Hurriyet Emlak. Available from: www.hurriyetemlak.com.

[8] V. W. Ripper. Visual Web Ripper. Available from: http://www.visualwebripper.com/.

[9] H. Witten and E. Frank. *Data Mining Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.