Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220072

Correlation Analysis of Voting Regression and Decision Tree Algorithm to Predict House Price with Improved Accuracy Rate

G. Hanuma Reddy[a] and P.Sriramya[b,1]
[a]*Research Scholar, Dept. of CSE, Saveetha School of Engineering,*
[b]*Professor, Dept. of AI&DS, Saveetha School of Engineering,*
[a,b] *SIMATS, Chennai, India*

Abstract. The primary goal of this study is to use efficient machine learning algorithms to anticipate better house prices, typically inflated. Materials and Methods: This study will study the differences between near-accurate price prediction utilizing Novel Voting Regression (Group 2) and Decision Tree methods (Group 1). The sample size used to carry out this research was N=10 for each group studied. Clincle was used to calculate the sample size. The pre-test analysis was maintained at 80%. G-power is used to calculate the sample size. Statistical analysis yielded a significance value of 0.001. Results: The accuracy of the Novel Voting Regression Algorithm for house price prediction is 82.94%, which is greater than the Decision Tree Algorithm's 72.54%. The Independent Sample T-test has a statistical significance of 0.584. Conclusion: As a result, it can be stated that the Novel Voting Regression technique can produce results that are almost as accurate as of the Decision Tree technique.

Keywords. Novel Voting Regression, Decision Tree, House Price Prediction, Prediction analysis, Machine Learning technique, Accuracy of prediction.

1. Introduction

Almost everyone desires a home tailored to their needs and includes all the facilities they need. House prices are constantly fluctuating. Housing prices are a key economic indicator, and price ranges are a widespread issue among buyers and sellers alike. The above is the case with this research, which will anticipate housing costs based on various explanatory factors. Data plays a significant role in machine learning (B and Swathi 2019). Data is utilized for training predictive models, resulting in reasonably accurate outputs. Without data, the model cannot be trained. [1] To facilitate easy comparisons among the numerous alternatives, the decision tree makes explicit all conceivable alternatives and follows each alternative to its conclusion in a single view. One of the best features of Decision Tree is its natural transparency. Another significant benefit is the capacity to choose the most biassed feature and the type of comprehensibility. It is also simple to categorize and interpret. Both continuous and

[1] P.Sriramya, Dept. of AI&DS, Saveetha School of Engineering, SIMATS, *Chennai, India.* E-mail: sriramyap@saveetha.com

discrete data sets can be employed. In decision trees, variable screening and feature sections are sufficient. [2].

Voting ensemble [3] suggests that attendance, parental education, and other variables may influence pupils' exam success. [3] In this case, the model is in an environment where it can train itself to become more accurate through trial and error. Many research publications in the field of data mining can be found on Google Scholar, Science Direct, and IEEE. In data mining, a total of 500 journal papers were discovered, 16900 articles were identified in Google Scholar, and 2509 items were discovered in IEEE.

The study's weaknesses include that they are rarely concerned with individual model performance and ignore the less popular but advanced models. We compared and analyzed the accuracy values of the Voting Regression and Decision Tree algorithms in this study. By comprehensively validating many ways in model implementation on regression, this research will also give a favorable result for housing price prediction.

2. Materials and Methods

The research was conducted at SIMATS' "Data Analytics Lab," which is part of the Saveetha School of Engineering's Department of Computer Science and Engineering. The method was carried out using a dataset including House-related data with columns like location, number of floors, number of bedrooms, number of bathrooms, built-up space, and many others. The IBM SPSS analysis was used to assess the comparative analysis between the two groups. A Decision Tree was used in group one, while a Novel Voting Regression technique was used in group two. The dataset's sample size for this study was 5. Jupyter is a powerful tool for pre-testing. The analyses' minimum power is 0.8, while the maximum allowed error is 0.5.

The dataset "House Price prediction in Beijing" contains around 300,000 datasets with more than 26 features that indicate housing prices exchanged between 2009 and 2018. These factors, which functioned as dataset attributes, were then utilized to forecast each house's average price per square meter. [4]Out of this vast database for this research, we have considered 350 records with 14 features. For testing the groups, the dataset was split into a training set and a testing set at 80% and 20%, respectively. Both Voting Regression[5] and Decision Tree algorithms[2] were trained using the training set, and it was tested using the test set.

The work was statistically analyzed using the IBM Statistical Package for Social Sciences in addition to the experimental analysis (SPSS). Mean, Standard Deviation, and Standard Error Mean were calculated as part of the analysis. To compare the two groups, the researchers used an independent sample T-test. The "IBM SPSS Independent T-test Analysis" is used for statistical analysis of two independent variables (Decision Tree method and Novel Voting Regression Algorithm). The nalyses' minimum power is set at 0.8, with a maximum allowed error of 0.5.

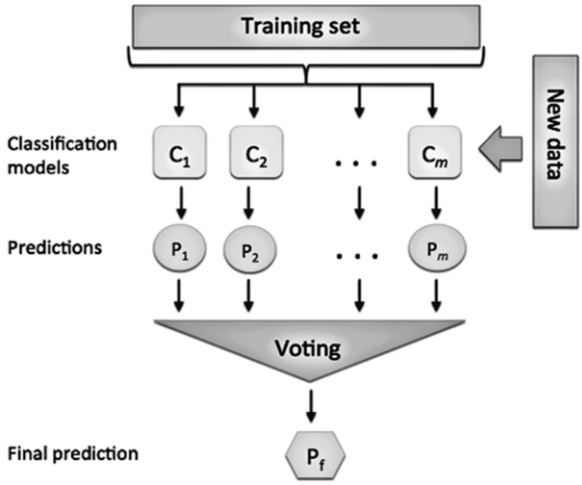


Figure 1. Working Model of Voting Regression Algorithm.

2.1 Decision Tree Algorithm

The Decision Tree Technique is a well-known supervised machine learning algorithm for classification. This method delivers an optimum output based on a tree structure containing criteria or rules. Decision Nodes, Design Links, and Decision Leaves are all part of the decision tree algorithm. Splitting, trimming, and tree selection are part of the system's operation. The decision tree may be built using both numerical and categorical data. For big datasets with low time complexity, decision tree methods are efficient. This Algorithm is primarily utilized in consumer segmentation and the execution of marketing strategies in businesses.

The working process can be explained in the following steps,

Step 1: Begin in the root node, which includes the whole dataset, according to S.

Step 2: Using the Attribute Selection Measure, find the best attribute in the dataset (ASM).

Step 3: Subdivide the S into subsets containing the best attribute's potential values.

Step 4: Make the optimal attribute decision tree node.

Step 5: Develop new decision trees iteratively using the subsets of the dataset created in step 3. Carry on in this manner until the nodes can no longer be classified, at which time the final node is known as a leaf node.

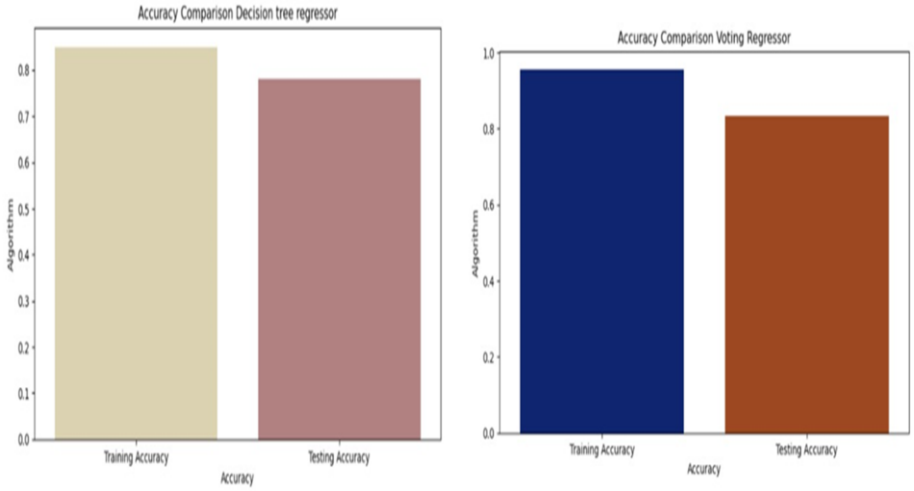
2.2 Novel Voting Regression Algorithm

A novel Voting ensemble is a model-improvement method aimed at outperforming each model in the ensemble. In a voting ensemble, the projections from numerous models are integrated. It can be used to categorize or forecast results. This implies averaging the models' predictions in the case of regression. Regarding classification, each label's predictions are combined, and the label with the most votes is predicted. Figure 1 depicts the working model for Novel Voting Regression. Regression and classification voting ensembles are the two forms of voting ensembles. In Regression Voting, predictions are the sum of contributing models. In classification voting, predictions are the votes of the majority of contributing models.

The minimal hardware requirements are as follows: Processor: Pentium IV, 8 GB RAM, 2.4 GHz processor, Main memory: 8 GB RAM, 600 MHz processor, 1TB hard disc drive Software specifications deal with the resources and requirements that must be installed on a device for a program to run. Before the software may be installed, some criteria must be met. The following are the software requirements at a bare minimum: The front end is written in the Python programming language. IDE–Jupiter notebook, operating system -7/8/10.

Table 1. Experiment Results of the accuracy achieved for Novel Voting Regression and Decision Tree.

Iterations	Accuracy (%)	
	Decision Tree	Voting Regression
1	69.23	79.67
2	71.16	79.02
3	71.69	80.26
4	72.03	81.25
5	72.96	82.97
6	73.06	82.96
7	73.62	83.62
8	73.96	83.96
9	74.08	84.08
10	74.76	84.76



(a) (b)

Figure 2a and 2b. Accuracy comparison graph of Training and Testing dataset for Decision Tree Algorithm and Voting Regression Algorithm

3 Results

Numerous iterations are performed to tune and find the optimal solution for each model. Decision Tree and Novel Voting Regression are the two group accuracy values achieved for ten different iterations are shown in Table 1. The T-Test sample has

calculated the mean accuracy, standard deviation, and standard error mean. The sample size of each algorithm is 10. Table 2 gives the SPSS results from the mean accuracy of the Voting Regression algorithm (82.94%) and Decision Tree Algorithm (72.54%). The Voting Regression's mean accuracy is significantly higher when compared with the Decision Tree.

Fig. 2a gives the comparative analysis of Test and training data for the performance evaluation parameters accuracy for the Decision Tree Algorithm. Fig. 2b shows the comparative graph of Test and training data for the performance evaluation parameters accuracy for the Novel Voting Regression Algorithm. From Table 3, it is observed that there is a slight significant increase in accuracy values in both the algorithms in the case of testing data. Since testing data is considered for the results, we can prove that Voting Regression can accurately predict the price. In Fig. 3, the accuracy gained is statistically calculated, and the results are given as a bar graph of two algorithms. They can confirm that the Novel Voting Regression algorithm has high accuracy compared to the Decision Tree algorithm.

Table 2. Statistics for the average accuracy, std deviation, standard error mean for Voting regression and Decision tree

Algorithm	N	Mean	Std.Deviation	Std.Error Mean
Multiple linear	10	72.54	7.552	2.388
RF	10	82.94	1.798	0.568

Table 3. Independent Sample T test for the two groups Decision Tree and Novel Voting Regression algorithms[significance is 0.001 (accuracy)]

	Levene's Test for Equality of variances		t-test for Equality of Means					95% Confidence interval of the Difference	
		F	Sig.	T	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
r-square	Equal variances assumed	14.008	.001	-4.238	10	.000	-10.404	2.455	-15.562	-5.246
	Equal variances not assumed			-4.238	10.016	.002	-10.404	2.455	-15.873	-4.935

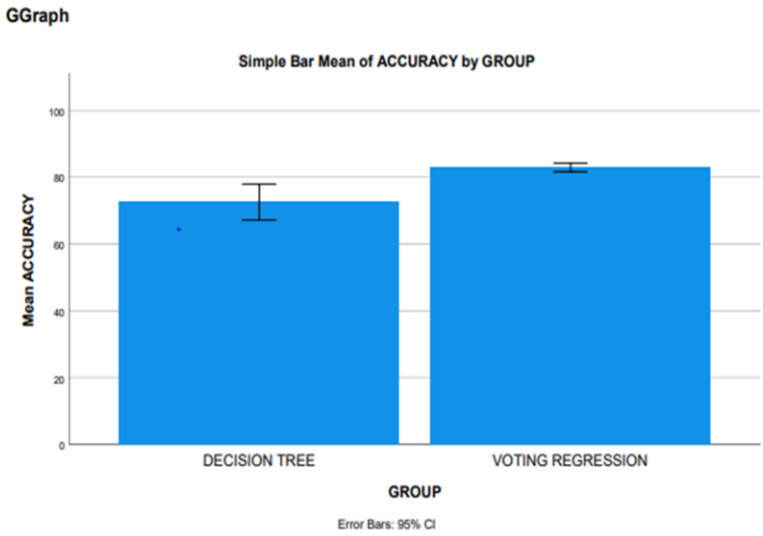


Figure 3. X-Axis: Novel Voting Regression vs Decision Tree and Y-Axis: Mean accuracy of detection ± 1SD.

4 Discussion

Based on the significant values obtained in statistical analysis, we found that the Novel Voting Regression technique appears to be more accurate than the Decision Tree approach. The average accuracy for Voting Regression (82.94%) and Decision Tree (72.54%). Table 2 shows the values of Mean Std. Deviation, Std. Error Mean for Decision Tree and Voting Regression algorithms. This research resulted in the prediction of house prices using the Voting Regression algorithm with an accuracy of 82.94% is more accurate than the values obtained using the Decision Tree algorithm with an accuracy of 72.54%. This article presents that the prediction of house prices [6] is better accurate with Novel Voting Regression compared to the values obtained with the Decision Tree algorithm.

Similar findings related to this research work are [7] XG Boost Algorithm and giving an accuracy of 85.06%. The research explores the differences between different advanced models using both classic and sophisticated [8] machine learning methodologies[9, 10] [11], [12]. The existing system deals with the price index of the highly stochastic or the temporary index values based on the feature of the product [13]. Using Principal Component Analysis to generate new indicators, the existing system has a more accurate 97% prediction of house prices [14]. It is the most practical forecasting method for predicting the price. The existing work has more accuracy of 95.5% in price prediction by using a multilayer neural network model [4]

Even though the study's results are superior in both experimental and statistical analysis, it does have certain drawbacks. They are providing a reliable and practical house price prediction method. The future work of this research study is to give more accurate values in predicting house prices based on different features.

5 Conclusion

We observed that in the SPSS results, the mean frequency of the Voting Regression algorithm (82.94%) and Decision Tree Algorithm (72.54%). In **Figure. 3**, the bar graph of the two algorithm's accuracy values can confirm that the Voting Regression algorithm has high accuracy compared to the Decision Tree Algorithm, which can be confirmed with these results that the Novel Voting Regression technique can give more accurate values than Decision Tree technique.

References

[1] Kaplan J. Defining Artificial Intelligence. *Artificial Intelligence*. Epub ahead of print 2016. DOI: 10.1093/wentk/9780190602383.003.0001.
[2] Maimon OZ, Lior R. *Data Mining With Decision Trees: Theory And Applications (2nd Edition)*. World Scientific, 2014.
[3] Qamar U, Niza R, Bashir S, et al. A Majority Vote Based Classifier Ensemble for Web Service Classification. *Business & Information Systems Engineering* 2016; 58: 249–259.
[4] Truong Q, Nguyen M, Dang H, et al. Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science* 2020; 174: 433–442.
[5] Pal M, Bharati P. Prediction of Voting Pattern. *Applications of Regression Techniques* 2019; 85–103.
[6] Chen W-T. *Overview of Machine Learning Methods in Predicting House Prices and Its Application in R*. 2017.
[7] B S, Swathi B. House Price Prediction Analysis using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology* 2019; 7: 1484–1493.
[8] Schapire RE. The Boosting Approach to Machine Learning: An Overview. *Nonlinear Estimation and Classification* 2003; 149–171.
[9] Pandiri VS. *Machine Learning Models to Predict House Prices Based on Home Features*. 2017.
[10] Rana S, Luo W, Tran T, et al. Application of Machine Learning Techniques to Identify Data Reliability and Factors Affecting Outcome After Stroke Using Electronic Administrative Records. *Front Neurol* 2021; 12: 670379.
[11] Sengar MSASAGVBAN, Shahi M, Singh A, et al. Machine Learning House Price Prediction. *International Journal for Modern Trends in Science and Technology* 2020; 6: 186–189.
[12] Li Z. Prediction of House Price Index Based on Machine Learning Methods. *2021 2nd International Conference on Computing and Data Science (CDS)*. Epub ahead of print 2021. DOI: 10.1109/cds52072.2021.00087.
[13] Xu Y, Cohen SB. Stock Movement Prediction from Tweets and Historical Prices. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Epub ahead of print 2018. DOI: 10.18653/v1/p18-1183.
[14] Xiao L, Yan T. Prediction of house price based on RBF neural network algorithms of principal component analysis. In: *2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*. IEEE. Epub ahead of print November 2019. DOI: 10.1109/iciibms46890.2019.8991474.