DATA MINING

TERM PROJECT: House Price Prediction using Machine Learning Algorithms

M1-ECS

Dararith KHUN
Institute of Technology of Cambodia
PO Box 86, Russian Conf. Blvd. Phnom Penh, Cambodia.

***Abstract*—In real estate, house price is growing straightly. People are carefully buying their new house with their budget and strategies in the market. Choosing the right house with suitable predicted price becomes a hot topic for this real estate era. The objective of this paper is to predict the coherent house price for non-house holder. By using machine learning algorithm to analyse the transaction of properties in Australia in order to make a good indicator of market condition and seek for a useful model that give a high efficiently prediction to buyer. This prediction covers amount of models to make a good comparison to make a better experience for buyer.**

***Index Terms*—House Price Prediction, Linear Regression, Support Vector Machine, Robust Regression, Lasso Regression, Random Forest, Machine Learning.**

I. Introduction

House price is really important in real estate market nowadays. Making a right decision of choosing house under a right budget is more likely important. Instantly, house price grows or not base many factors such as, location, house condition, and house environment etc. Therefore, predicting house not only benefit to the buyer, but it also a good thing for real estate agents to be able to arrange the market for buyer[1].

The study on house market prediction, we focus on the house price, growth trend and its related features. The motivation of the study because of the problem of buyer when they have a hard decision on choosing house, and they can't buy house base on their budget. Even more, buyer can't estimate their wanted property price range in different location.

The goal of this paper want to solve those mentioned problems for buyer and make an improvement of predicting. Base on Machine Learning algorithm, we will provide high quality of prediction buy selecting a right model among 5 models. They are **Linear Regression, Robust Regression, Lasso Regression, Support Vector Machine and Random Forest**. Furthermore, we include many feature to make sure the model test on only quality data by perform **Data Visualization, Clean Mission Value, Feature Selection and Outlier Detection**. We will select for a useful model which evaluate by using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R Squared (R2)[2].

II. Literature Review

The idea of literature review, is to make a shot summary of related work. This article makes reference to the most recent research projections that take trends into account to further plan their economies. The primary goal of the project FORECASTING VARIATIONS ON Home PRICE was to determine which algorithm was most effective at making the best predictions of house prices with the lowest possible error rates. Focuses on comparing several machine learning techniques for house price prediction analysis, including multiple linear regression, ridge regression, LASSO regression, elastic net regression, ada boosting regression, and gradient boosting[3]. One more paper they explain that forecasts are produced using a collection of algorithms that includes both conventional time series models and alternative machine and deep learning approaches. There are at least two reasons why this is interesting. A large number of ML and DL algorithms can identify nonlinear correlations and choose models using cross-validation rather than information criteria. It has to be shown empirically if such changes from conventional approaches make ML more appropriate for predicting property values. Second, we apply the concept of multi-equation models to all of our machine learning requirements, extending it from linear vector-autoregressions (VARs)[4].

Other one topic they mention that their major goal is to create a model that can estimate a customer's 1936 property cost based on that customer's preferences. In order to determine the best pricing based on the customer's needs and interests, our model analyzes a set of parameters that they have chosen. It attempts to provide an analysis of the outcomes and employs the traditional techniques of linear regression, forest regression, and boosted regression for prediction. Additionally, neural networks are employed to improve the algorithm's accuracy, which is subsequently improved even more through boosted regression. It contributes to determining the strength of the association between the dependent variable and other dynamic independent variables, referred to as label attribute and regular attribute, respectively[5]. Memristors are used as synapses to create a 2-layer feed-forward neural network in this study. The pulse voltage with BP technique allows for the live adjustment of synapses' weights. The ANN can correctly forecast the house price in the predicting mode after learning to do so while operating in the training mode. The house price samples from numerous Boston communities in the US were used to train the ANN to produce predictions, and it was discovered that the anticipated results were rather similar to the target data[6]. In this study, we conducted a thorough investigation into the usage of MTL for the problem of predicting property prices. Using a wide variety of location data sources, we design and collect a fine-grained location profile in terms of data profiling. Task definitions and technique selections are two crucial steps in the implementation of MTL-based house price prediction in terms of the prediction model. To construct tasks based on distinct house aspects, we created two kinds of strategies. To capture and make use of the relatedness between tasks, we used three generic MTL-based approaches with varied regularization terms. We initially showed, via thorough experimental assessments, that modeling based on MTL may greatly enhance the overall performance of home price prediction[7].

One of intersting paper the mention that, They looked at ciphertext search in the context of cloud storage in this article andlook at the issue of keeping the semantic relationship between various plain texts over the connected encrypted documents and provide a design strategy to improve the functionality of semantic search. In order to meet the demands of the data explosion, online information retrieval, and semantic search, we also suggest the MRSE-HCI design. A verifiable technique is also suggested at the same time to ensure the accuracy and completeness of search results[8]. The proposed research methodology entails four stages, namely Data Collection, Pre Processing the gathered data then remodeling it in accordance with the good format, increasing clever models using desktop learning algorithms, training, testing, then validating the model of house prices over the housing need in the capital, Coimbatore. Online religious retailers, who offer a reasonable account of the city housing market, are asking for the data that has been accumulated since model validation and checking out. The calculating model significantly aids in the estimation of upcoming housing expenses in Coimbatore. The regression results are positive and provide useful guidance for further estimation tasks involving the gathered dataset[9]. One of a person's fundamental necessities is a home, and costs vary from location to location based on features like parking and neighborhood. One of a family's greatest and most significant decisions is to purchase a home since they invest all of their money and gradually pay it off with loans. When machines learn from data and utilize it to produce new data, modeling is done using machine learning algorithms. based on XGBoost, Machine Learning, Lasso, Ridge, and Random Forest Regression, as well as linear regression [10].

III. Data Preparation and Visualization

A. Data Preparation

The data for implement on this paper is from Kaggle data. It describes about the real estate market in Sydney and Melbourne. This data was collected since 2014 that consist of 4600 rows and 18 columns. A sizable collection of records for property sales that are kept in an unknown format and have unidentified data quality concerns. It consists of types variable:

- Transactions variables are price, date.
- Some related for prediction are condition, street, city, statezip, and country
- Other features are bedrooms, bathrooms sqftliving, sqftlot, floors, water front, views, sqftabove, sqftbasement, yrBuilt, yrrenovated.

Models are ready to apply on this data but before we apply, we need to do some pre-process on data in order to check for some missing value to make sure our data are cleaned.

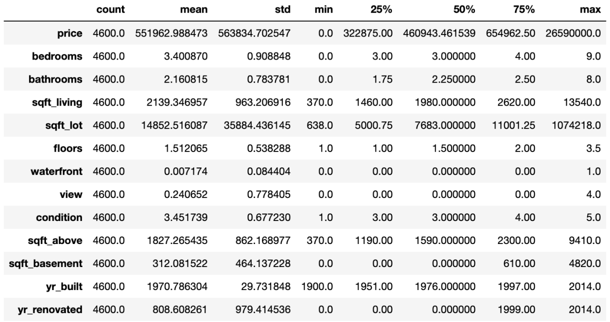
TABELA I
Table of features

Feature Name	type	Null Value
date	object	0
price	float64	0
bedrooms	float64	0
bathrooms	float64	0
sqftliving	int64	0
sqftlot	int64	0
floors	float64	0
waterfront	int64	0
view	int64	0
condition	int64	0
sqftabove	int64	0
sqftbasement	int64	0
yrbuilt	int64	0
yrrenovated	int64	0
street	object	0
city	object	0
statezip	object	0
country	object	0

After checking for isnull() data, we found that our data are already cleaned(no null value found). That means our data are ready to work with model for prediction.

B. Data Visualization

To be easy on understanding of data, we study on data by describe it in term of **counted data, mean value, minimum value, maximum value and standard deviation** of each features.



	count	mean	std	min	25%	50%	75%	max
price	4600.0	551962.988473	563834.702547	0.0	322875.00	460943.461539	654962.50	26590000.0
bedrooms	4600.0	3.400870	0.908848	0.0	3.00	3.000000	4.00	9.0
bathrooms	4600.0	2.160815	0.783781	0.0	1.75	2.250000	2.50	8.0
sqft_living	4600.0	2139.346957	963.206916	370.0	1460.00	1980.000000	2620.00	13540.0
sqft_lot	4600.0	14852.516087	35884.436145	638.0	5000.75	7683.000000	11001.25	1074218.0
floors	4600.0	1.512065	0.538288	1.0	1.00	1.500000	2.00	3.5
waterfront	4600.0	0.007174	0.084404	0.0	0.00	0.000000	0.00	1.0
view	4600.0	0.240652	0.778405	0.0	0.00	0.000000	0.00	4.0
condition	4600.0	3.451739	0.677230	1.0	3.00	3.000000	4.00	5.0
sqft_above	4600.0	1827.265435	862.168977	370.0	1190.00	1590.000000	2300.00	9410.0
sqft_basement	4600.0	312.081522	464.137228	0.0	0.00	0.000000	610.00	4820.0
yr_built	4600.0	1970.786304	29.731848	1900.0	1951.00	1976.000000	1997.00	2014.0
yr_renovated	4600.0	808.608261	979.414536	0.0	0.00	0.000000	1999.00	2014.0

Fig. 1. Figure of Described data.

We will target on *'price'* column. Let's see how distribution of target column look like.

We don't see a direct linearity of features with our target price. Although we can see that there are outliers. Outliers

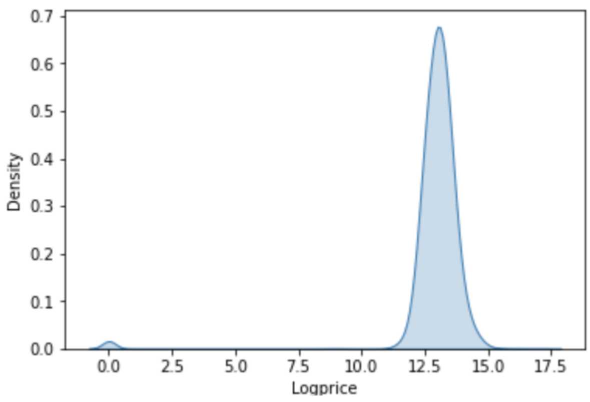


Fig. 2. Figure of log(price).

can either be a mistake or just variance. We will perform and handle on outlier later in methodology.

IV. Methodology

There are some methods that use to filter on data set before we apply model for prediction.

Apply Feature Selection To strengthen the capacity for interpretation and raise the effectiveness of prediction models, selecting qualities features is really need. To handle on this feature selection, we will use mutual information regression from **Scikit Learn**. It uses to estimate mutual information for a discrete target variable. The function of mutual information uses nonparametric techniques based on entropy estimate from distances between k-nearest neighbors.

Back to implement, first we import *mutual info regression from sklearn* then we find the discrete value of each features. After we found that feature *street* has a highest discrete value 1.149118, otherwise feature *country* has zero discrete value.

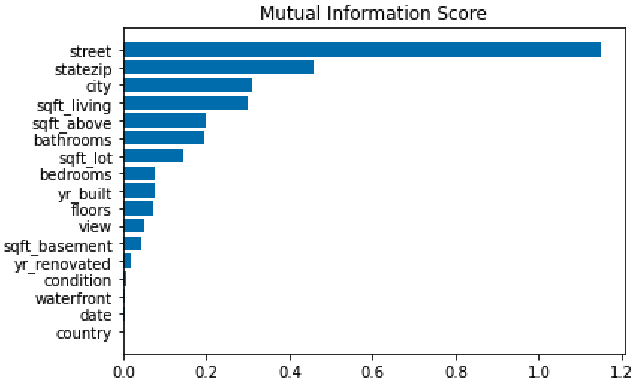


Fig. 3. Mutual information score.

Notice: no feature price in the plot because price is out target column

Base on mutual information score, we decide to select top 9 out of 17 feature manually that show in the plot and in total 10(include price feature) out of 18 features. That means after we perform feature selection, there are 8 features are dropped out. Out new shape of data set will be (4600 x 10).

Handle on Outlier As we mention in visualizing data, we cannot find a direct linearity of features with our target price that cause of outlier. To overcome this problem we will use **Z-score**.

Formula

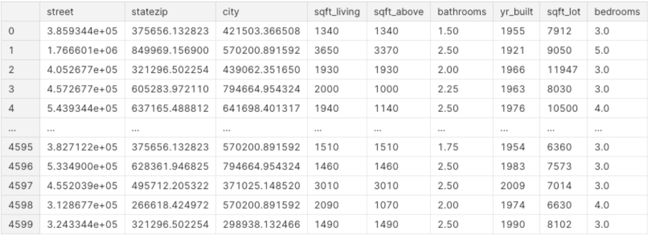
$$Z = \frac{x - \mu}{\sigma}$$

where

- x is the raw score
- μ is the population mean
- σ is the population standard deviation

After applying Z-Score, we define threshold 3 or -3 to check if z-score value greater than 3 or smaller than -3, it is the outlier. As a result, 273 rows data were found as outlier and our new data shape is (4435 x 10).

Split Data The data were split into 80 percent as training set and 20 percent as testing set. By the way, some feature such as [street, statezip, city] are type of non-numerical. To make a better performance on prediction we will encode those feature to numeric.



	street	statezip	city	sqft_living	sqft_above	bathrooms	yr_built	sqft_lot	bedrooms
0	3.859344e+05	375656.132823	421503.366508	1340	1340	1.50	1955	7912	3.0
1	1.766601e+06	849969.156900	570200.891592	3650	3370	2.50	1921	9050	5.0
2	4.052677e+05	321296.502254	439062.351650	1930	1930	2.00	1966	11947	3.0
3	4.572677e+05	605283.972110	794664.954324	2000	1000	2.25	1963	8030	3.0
4	5.439344e+05	637165.488812	641698.401317	1940	1140	2.50	1976	10500	4.0
...	...	...	...	...	...	...	...	...	...
4595	3.827122e+05	375656.132823	570200.891592	1510	1510	1.75	1954	6360	3.0
4596	5.334900e+05	628361.946825	794664.954324	1460	1460	2.50	1983	7573	3.0
4597	4.552039e+05	495712.205322	371025.148520	3010	3010	2.50	2009	7014	3.0
4598	3.128677e+05	266618.424972	570200.891592	2090	1070	2.00	1974	6630	4.0
4599	3.243344e+05	321296.502254	298938.132466	1490	1490	2.50	1990	8102	3.0

Fig. 4. Data encoded.

Evaluation There many ways to evaluate machine learning model. In this paper, we use Mean Squared Error, Mean absolute error, Root Mean Squared Error, and R-Squared evaluate the performance of the model in regression analysis.

Mean Absolute Error: use to represent the standard of absolute different between actual and predicted value in dataset.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y}|$$

where

- $\hat{y}$ is predicted value of y

Mean Squared Error: use to represent the standard of the squared difference between the original and predicted values in the data set.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y})^2$$

Root Mean Squared Error: is the squared root of MSE.

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

R-Squared: use to represent standard of the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score.

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

Model Selection There are five models that are selected to perform the prediction.

Linear Regression is a method in supervised learning that uses to predict quantitative variable base on linear connection with one or more independent feature. It can be represented by:

$$y = mx + b$$

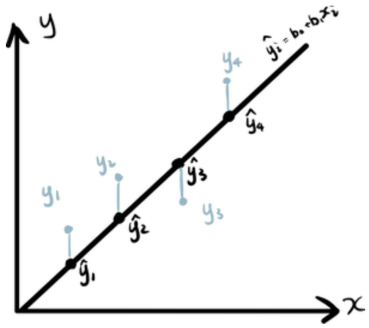


Fig. 5. Simple linear regression.

Robust Regression: comes from a form of regression technique. It is used to overcome the problem in Linear Regression. Robust Regression comes directly with outlier detection and cleaning. It is represented by

$$Y = f(X_i, \beta) + e_i$$

where

- Y_i is dependent variable
- f is function
- X_i is independent variable
- β is unknown parameter
- e_i is error terms

LASSO Regression: is a type of regression. It is used to overcome(Least Absolute Shrinkage and Selection Operator), and It can pick relevant features that will be useful for modelling. Moreover, LASSO such the best subset selection and perform feature selection by it own.

Support Vector Machine: A potent method for supervised learning is the Support Vector Machine (SVM). The SVM technique raises the original data's dimension in order to look for a hyperplane for data separation. SVM tends to deliver better accuracy due to its ability of fitting nonlinear boundary.

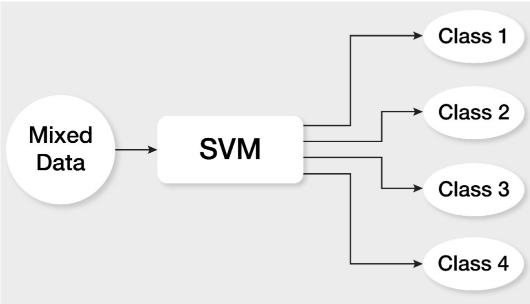


Fig. 6. Simple SVM.

Random Forest: is a method for classification and also regression. It conduct the many decision tree during training. The output of it when it is a classification, is the selected class by the most of the tree. When it perform task as regression, it return as the mean prediction of each trees.

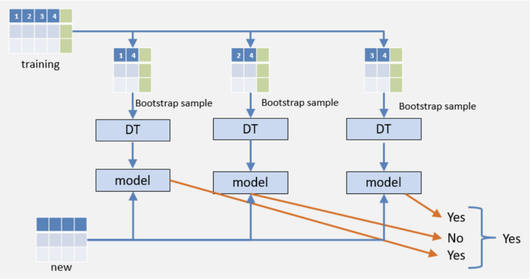


Fig. 7. Simple of random forest.

V. Implementation and Result

A. Linear Regression

Applying model linear regression base on library of Scikit Learn. When the data is already clean and select only important feature. The model fit well with the data which provid 99 percent accuracy.

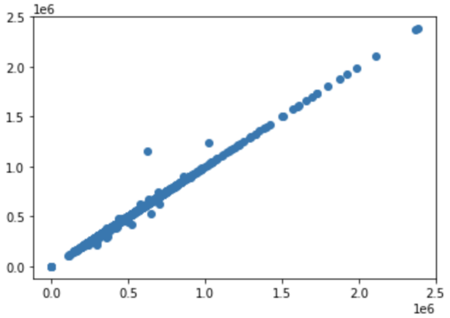


Fig. 8. Data fitting linear regression.

B. Robust Regression

Unlikely from Linear Regression, we import it from Sklearn with data randomly selected. The model fit really well and provide 99 percent accuracy.

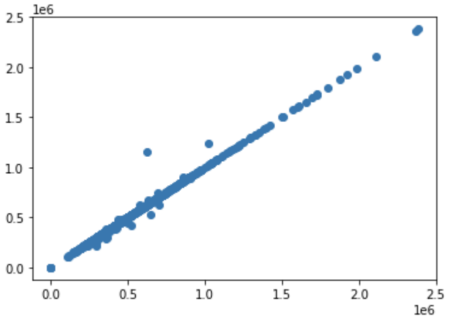


Fig. 9. Data fitting robust regression.

C. LASSO Regression

This regression model also provide a really good result. As we implement, we apply feature selection and in this model also include feature selection. We see that this model also fit very well with 99 percent.

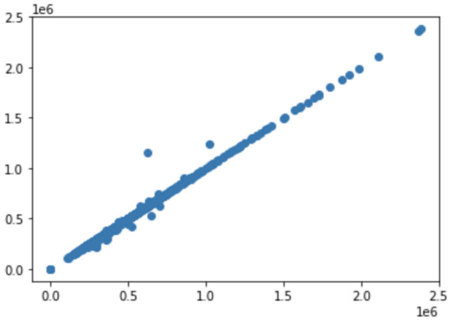


Fig. 10. Data fitting LASSO regression.

D. Support Vector Machine

SVM can perform in task of both classification and also regression. It still work well on predicting with 99 percent.

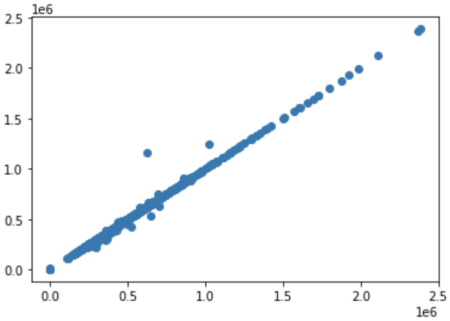


Fig. 11. Data fitting SVM.

E. Random Forest

This last model, still keep good performance with 99 percent accuracy. But this model seem has a bit noise base on fitting plot data.

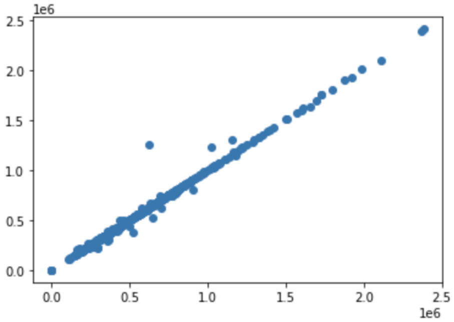
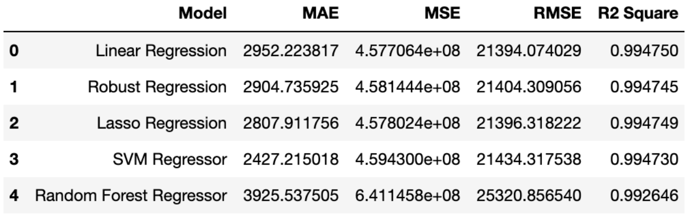


Fig. 12. Data fitting Random Forest.

Result

As a result, we see that each model work really well. That Linear Regression has the best performance while Random Forest is the lowest. But it is not a problem since they are all 99.99 percent.



	Model	MAE	MSE	RMSE	R2 Square
0	Linear Regression	2952.223817	4.577064e+08	21394.074029	0.994750
1	Robust Regression	2904.735925	4.581444e+08	21404.309056	0.994745
2	Lasso Regression	2807.911756	4.578024e+08	21396.318222	0.994749
3	SVM Regressor	2427.215018	4.594300e+08	21434.317538	0.994730
4	Random Forest Regressor	3925.537505	6.411458e+08	25320.856540	0.992646

Fig. 13. Result of each model.

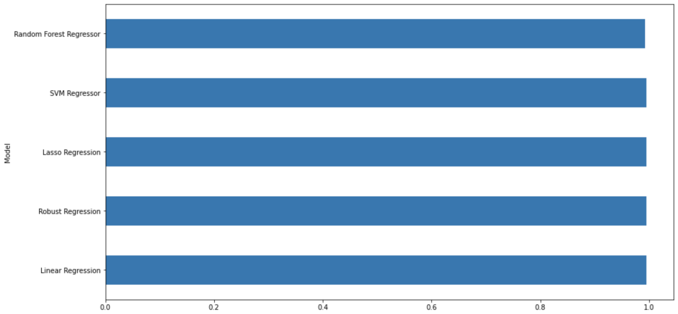


Fig. 14. Bar graph of each model performance.

VI. Conclusion

In conclusion, house price prediction is really important for market trend nowadays. In this project is really a good practice of data analyse especially in real estate. We can use machine learning algorithm to predict house price base on market. The result of our model are nearly the same (99.99 percent). Maybe our data are really good and some case maybe our data are over fitting which lead to this high performance.

In future work, i would like to improve more on this project make a good use of this project to apply in real life condition and deploy for general uses. Especially, i want to apply on the real data of Cambodia in order to preserve the national uses.

Referências

[1] Natalie Campisi. *Housing market predictions 2022: When will prices drop?* Jul. de 2022. URL: https://www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/.

[2] Davide Chicco, Matthijs J Warrens e Giuseppe Jurman. “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation”. Em: *PeerJ Computer Science* 7 (2021), e623.

[3] CH Raga Madhuri, G Anuradha e M Vani Pujitha. “House price prediction using regression techniques: a comparative study”. Em: *2019 International conference on smart structures and systems (ICSSS)*. IEEE. 2019, pp. 1–5.

[4] George Milunovich. “Forecasting Australia’s real house price index: A comparison of time series and machine learning methods”. Em: *Journal of Forecasting* 39.7 (2020), pp. 1098–1118.

[5] Ayush Varma, Abhijit Sarma, Sagar Doshi e Rohini Nair. “House price prediction using machine learning and neural networks”. Em: *2018 second international conference on inventive communication and computational technologies (ICICCT)*. IEEE. 2018, pp. 1936–1939.

[6] JJ Wang, SG Hu, XT Zhan, Q Luo, Qi Yu, Zhen Liu, Tu Pei Chen, Y Yin, Sumio Hosaka e Yang Liu. “Predicting house price with a memristor-based artificial neural network”. Em: *IEEE Access* 6 (2018), pp. 16523–16528.

[7] Guangliang Gao, Zhifeng Bao, Jie Cao, A Kai Qin e Timos Sellis. “Location-centered house price prediction: A multi-task learning approach”. Em: *ACM Transactions on Intelligent Systems and Technology (TIST)* 13.2 (2022), pp. 1–25.

[8] Zhongyun Jiang e Guoxin Shen. “Prediction of house price based on the back propagation neural network in the keras deep learning framework”. Em: *2019 6th International Conference on Systems and Informatics (ICSAI)*. IEEE. 2019, pp. 1408–1412.

[9] T Suganya, M Gowtham e Hari Shanmuga T Raja. “Machine Learning Based Prediction of House Price”. Em: *International Journal of Health Sciences* III (), pp. 9554–9567.

[10] Kumari Sandhya e Sarwar Siddiqui. “House Price Prediction Using Machine Learning”. Em: ().