HOUSE PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES

Basit Ali CS201177, Usman Saeed Motiwala CS201157, Abdul Moiz CS201004

Department of Computer Science, Dha Suffa University

Artificial Intelligence Lab, CS-3102L

Ms Aniqa Hussain

June 20, 2023

ABSTRACT

This research paper aims to predict house prices using two machine learning algorithms, Support Vector Machine (SVM) and Random Forest Classifier (RFC). The purpose of this research is to compare the performance of SVM and RFC in predicting house prices and determine which algorithm yields accurate results. The study focuses on utilizing the number of bedrooms and bathrooms as the input features for the prediction task. The findings of this study can provide valuable insights for real estate agents and homeowners in estimating house prices more accurately.

KEYWORDS: Machine Learning, House Prices, Support Vector Machine, Random Forest Classifier, Accuracy Comparison.

I. INTRODUCTION

Predicting house prices accurately has been a significant research area in the field of machine learning. Accurate house price prediction is crucial for various purposes, such as real estate investments, mortgage approvals, and property valuation. House prices trends are not only the concerns for buyers and sellers, but they also indicate the current economic situations [1]. That is why it is important to predict accurate house prices without bias to help buyers and sellers.

The research compares the performance of Support Vector Classifier (SVC) and Random Forest Classifier (RFC) algorithms in this prediction task. There are many factors which affect house prices such as the number of bedrooms and bathrooms. This study is to compare which among these two algorithms provides more accurate prediction. The research is being conducted on houses of Pakistan and the dataset is used from Kaggle of Graana to train the machine learning model.

II. LITERATURE REVIEW

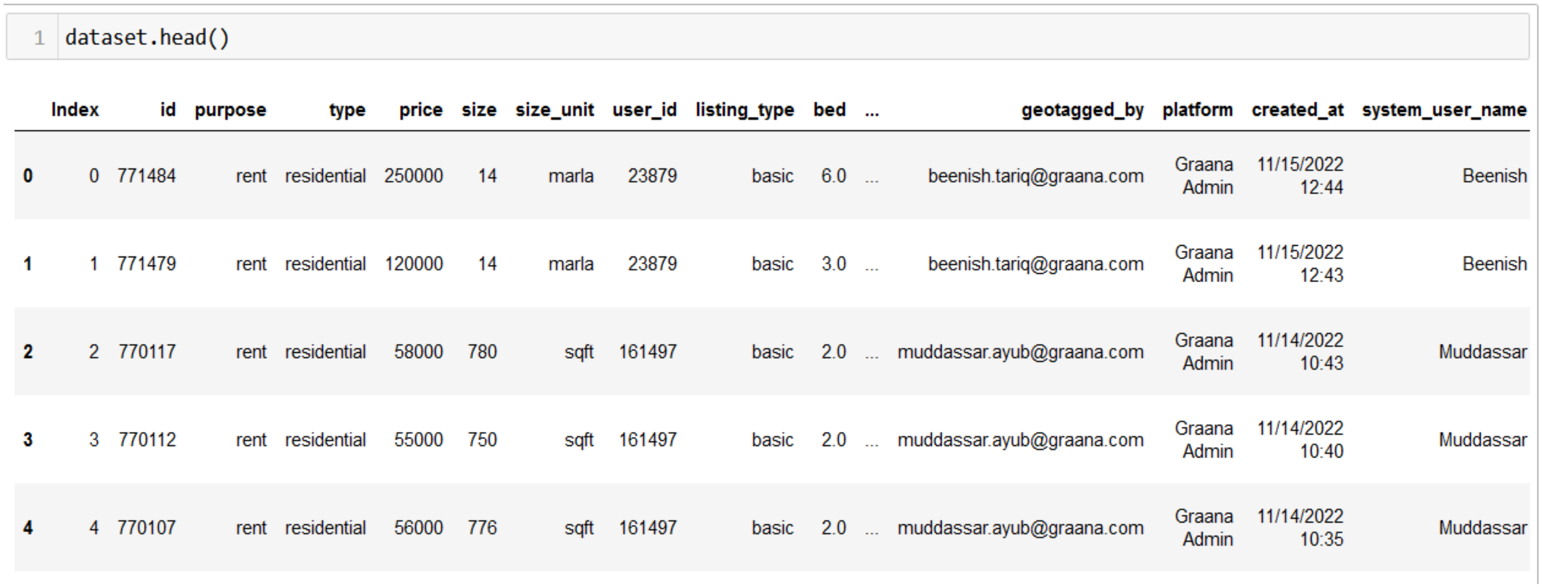
Name	Description	Pros	Cons	Reference
Support Vector Machine	SVM is a machine learning algorithm that classifies data by finding the best hyperplane to separate different classes, maximizing the margin between them	Effective for high-dimensional data. Robust against overfitting. Versatile kernel functions. Memory-efficient.	Computationally expensive. Sensitivity to parameter tuning. Lack of probabilistic outputs. Limited effectiveness with noisy data.	Auria, L., & Moro, R. A. (2008). Support vector machines (SVM) as a technique for solvency analysis. [2]
Random Forest Classifier	Random Forest is an ensemble learning algorithm that combines multiple decision trees and combines their prediction through voting or averaging to make accurate predictions.	Good performance on large datasets. Reduces overfitting. Robust to outliers and noise. Deals with irrelevant inputs.	Low prediction accuracy. High variance Computationally expensive training. Lack of interpretability. Requires careful parameter tuning.	Montillo, A. A. (2009). Random forests. Lecture in Statistical Foundations of Data Analysis. [3]

		Computationally Scalability.	Difficult to visualize the decision-making process.	

III. CONCEPT/ANALYSIS

Data Set:

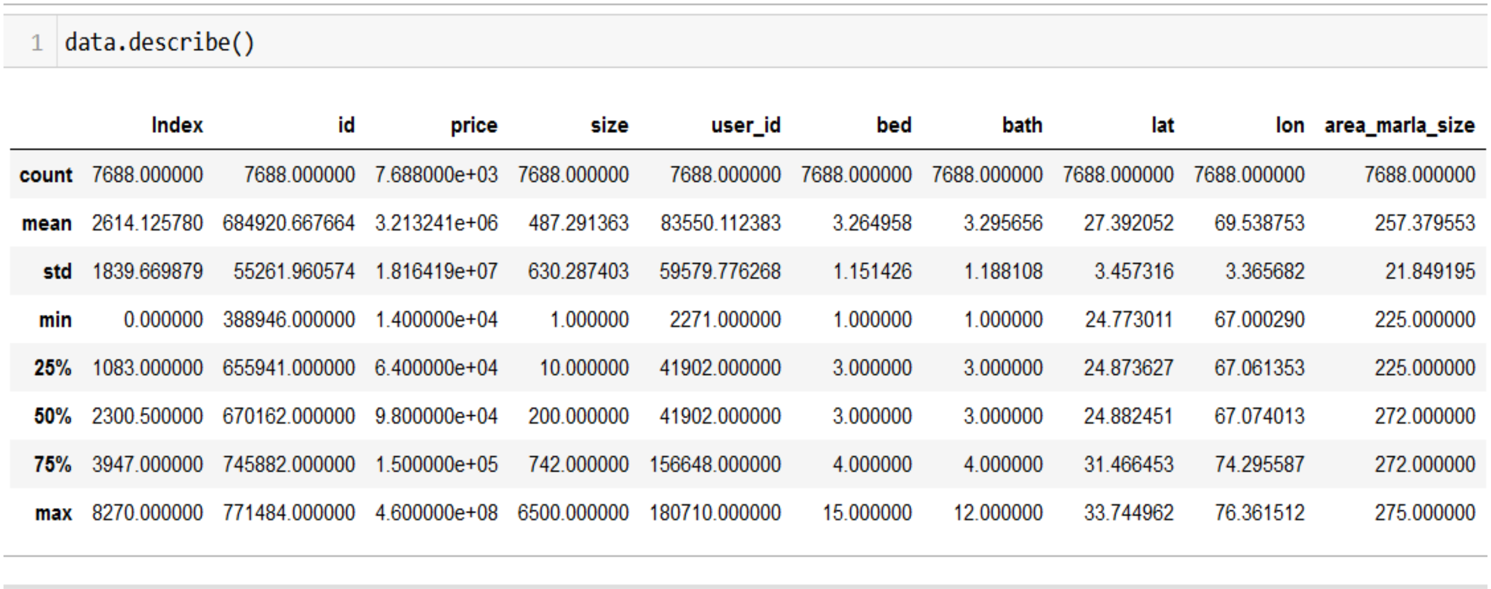
Figure 1: *Dataset first 5 records or rows.*



```
dataset.head()
```

	Index	id	purpose	type	price	size	size_unit	user_id	listing_type	bed	...	geotagged_by	platform	created_at	system_user_name
0	0	771484	rent	residential	250000	14	marla	23879	basic	6.0	...	beenish.tariq@graana.com	Graana Admin	11/15/2022 12:44	Beenish
1	1	771479	rent	residential	120000	14	marla	23879	basic	3.0	...	beenish.tariq@graana.com	Graana Admin	11/15/2022 12:43	Beenish
2	2	770117	rent	residential	58000	780	sqft	161497	basic	2.0	...	muddassar.ayub@graana.com	Graana Admin	11/14/2022 10:43	Muddassar
3	3	770112	rent	residential	55000	750	sqft	161497	basic	2.0	...	muddassar.ayub@graana.com	Graana Admin	11/14/2022 10:40	Muddassar
4	4	770107	rent	residential	56000	776	sqft	161497	basic	2.0	...	muddassar.ayub@graana.com	Graana Admin	11/14/2022 10:35	Muddassar

Figure 2: *Dataset description.*



```
data.describe()
```

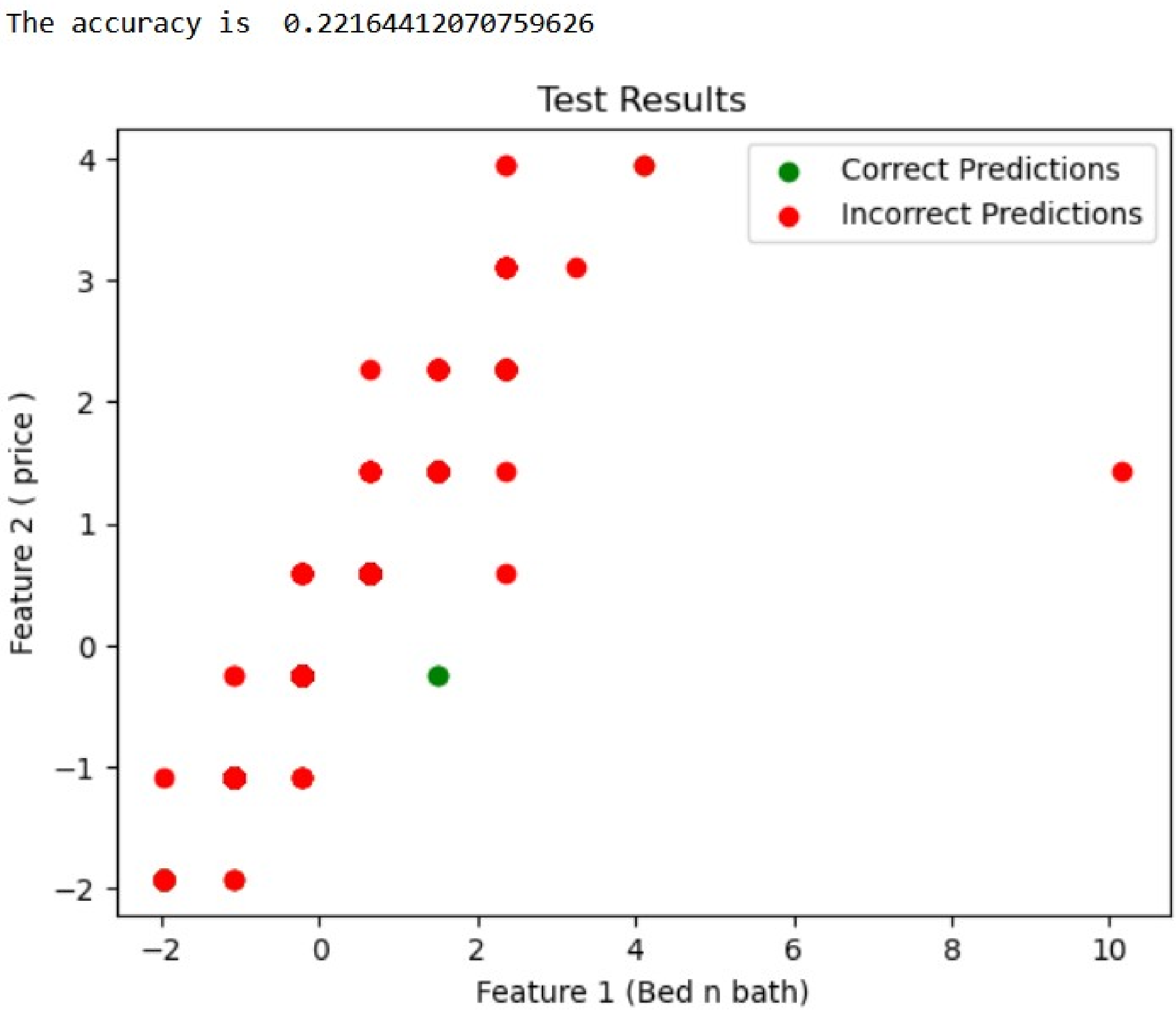
	Index	id	price	size	user_id	bed	bath	lat	lon	area_marla_size
count	7688.000000	7688.000000	7.688000e+03	7688.000000	7688.000000	7688.000000	7688.000000	7688.000000	7688.000000	7688.000000
mean	2614.125780	684920.667664	3.213241e+06	487.291363	83550.112383	3.264958	3.295656	27.392052	69.538753	257.379553
std	1839.669879	55261.960574	1.816419e+07	630.287403	59579.776268	1.151426	1.188108	3.457316	3.365682	21.849195
min	0.000000	388946.000000	1.400000e+04	1.000000	2271.000000	1.000000	1.000000	24.773011	67.000290	225.000000
25%	1083.000000	655941.000000	6.400000e+04	10.000000	41902.000000	3.000000	3.000000	24.873627	67.061353	225.000000
50%	2300.500000	670162.000000	9.800000e+04	200.000000	41902.000000	3.000000	3.000000	24.882451	67.074013	272.000000
75%	3947.000000	745882.000000	1.500000e+05	742.000000	156648.000000	4.000000	4.000000	31.466453	74.295587	272.000000
max	8270.000000	771484.000000	4.600000e+08	6500.000000	180710.000000	15.000000	12.000000	33.744962	76.361512	275.000000

To predict house prices using Support Vector Classifier and Random Forest Classifier, we first removed any data with null values, then we used two features from the dataset to predict house price: number of bedrooms and bathrooms. The dataset was divided into training and testing sets (80% training set and 20% testing set) to evaluate the performance of the algorithms. SVC was implemented using a radial basis function kernel, while RFC was trained with 100 decision trees. The house price of test data is predicted using bedrooms and bathrooms and is then compared with its actual value of test data to calculate Accuracy Score. The models were evaluated on bases of Accuracy score to assess their predictive accuracy.

IV. RESULTS

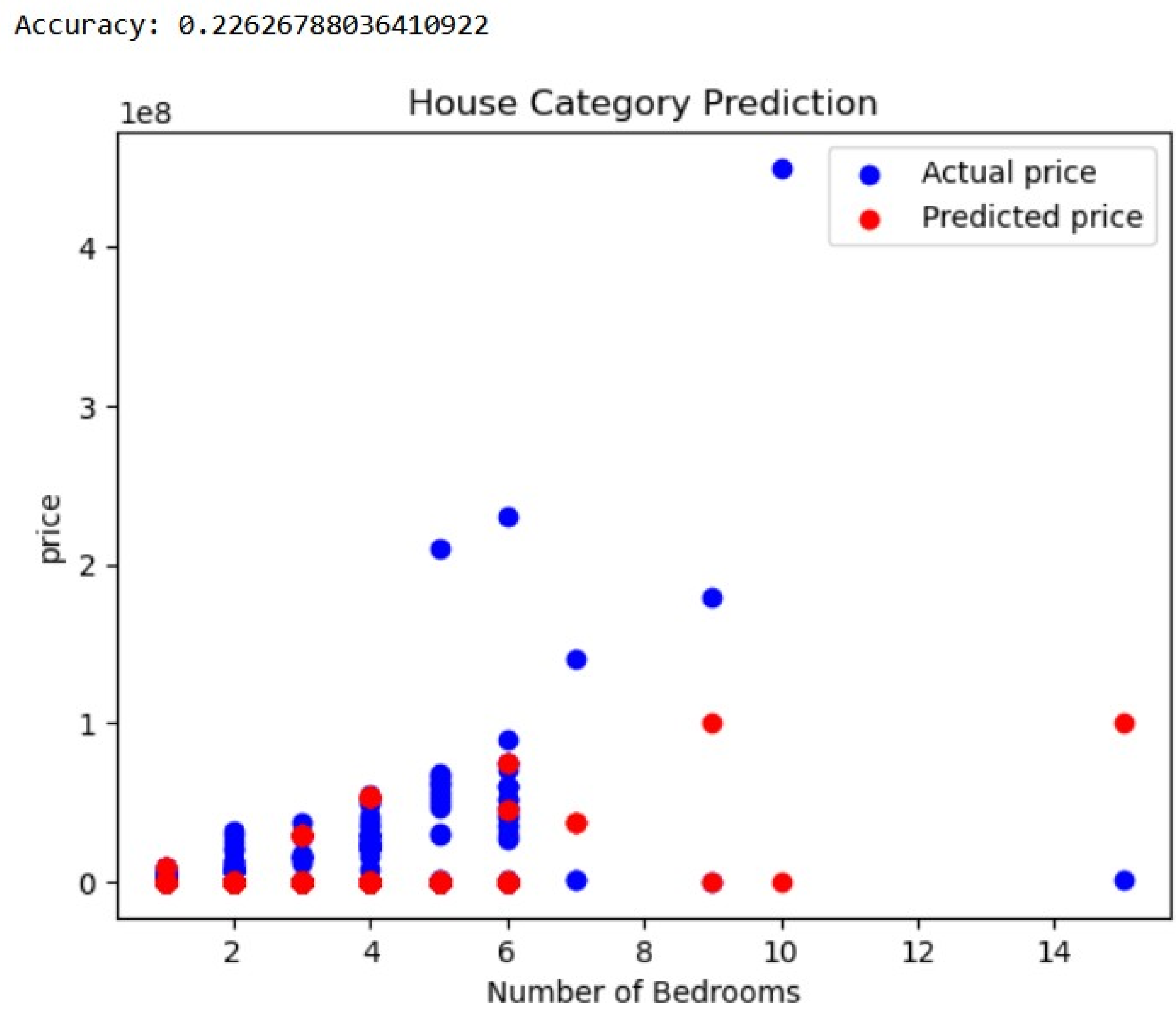
Screenshot:

Figure 3: *SVM Graph of number of correct and incorrect predictions.*



```
The accuracy is  0.22164412070759626
```

Figure 4: *Random Forest Classifier Prediction Graph.*



```
Accuracy: 0.22626788036410922
```

The results of our analysis indicate that both Support Vector Classification and RFC performed well in predicting house prices. However, RFC exhibited slightly better performance compared to SVC. The Accuracy Score for RFC were higher than those of SVC, indicating that SVC produced more accurate predictions. The feature importance analysis revealed that the number of bedrooms and bathrooms were the most significant factors influencing house prices, as captured by both algorithms. From the result of the Accuracy Score we can see that for this particular dataset RFC is better than SVC.

V. CONCLUSION

In this research paper, we compared the performance of two machine learning algorithms, Support Vector Classifier and Random Forest Classifier, in predicting house prices. The results demonstrated that RFC outperformed SVC in terms of predictive accuracy, as indicated by their accuracy score. These findings suggest that both algorithms have similar results for house price prediction, with RFC slightly more accurate and better suited for this dataset than SVC. However, further research can explore other machine learning algorithms and features to improve the accuracy of house price predictions.

VI. REFERENCE

[1] D. Kumar, S. and K. , "HOUSE PRICE PREDICTION USING RANDOM FOREST AND CNN," *International Research Journal of Modernization in Engineering Technology and Science,* vol. 4, no. 8, p. 5, 2022.

[2] A. Laura and R. A. Moro, "Support vector machines (SVM) as a technique for solvency analysis.," *Discussion Papers,* p. 18, 2008.

[3] M. Albert A, "Random forests," *Lecture in Statistical Foundations of Data Analysis,* p. 28, 2009.

[4] W. and J. Yang, "Housing price prediction using support vector regression," *Master's Projects,* p. 58, 2017.