MANIPAL
ACADEMY of HIGHER EDUCATION
(Deemed to be University under Section 3 of the UGC Act, 1956)

MANIPAL SCHOOL OF INFORMATION SCIENCES

(A Constituent unit of MAHE, Manipal)

Housing Prices Prediction

Reg. Number	Name	Branch
201046003	KAVYA UMESH NAIK	BDA
201046017	DIVYA G B	BDA
201046033	SHREYA RAMDAS NAYAK	BDA

Under the guidance of

AROCKIARAJ S

Assistant Professor- Selection Grade,

Manipal School of Information Sciences,

MAHE, MANIPAL

DECLARATION

We declare that this mini project, submitted for the evaluation of course work of Mini Project to Manipal School of Information Sciences, is (our original work / using / implementing / extending an existing idea / concept / code available at (https://www.kaggle.com/vedavyasv/usa-housing), conducted under the supervision of our guide Arockiaraj S and panel members Deepak Rao B Sir, Sathyendranath Malli References, help and material obtained from other sources have been duly acknowledged.

ABSTRACT

In today's world, everyone wishes for a house that suits their lifestyle and provides amenities according to their needs. While purchasing the house, the price of house is the main factor which is considered by people. The pricing of house not only depends on the size of the property and no. of rooms, but also on the neighbourhoods like transport facility, banks, schools or colleges, shops etc. When a person buys a home, they consider structural features, working accessibility, neighbourhood services. Hence, a house price prediction system is invented to improve estimation of house prices. House prices keep on changing very frequently which proves that house prices are often exaggerated. There are many factors that have to be taken into consideration for predicting house prices such as location, number of rooms, carpet area, how old the property is? and other basic local amenities.

In the present report we discuss about the prediction of future housing prices that is generated by machine learning algorithm. For the selection of prediction methods, we compare and explore various prediction methods like Linear regression, Neural network. Our result exhibit that our approach of the issue need to be successful, and has the ability to process predictions that would be comparative with other house cost prediction models.

Keywords

Machine learning algorithm, House Price Prediction, PCA, Linear regression, Neural network

Contents

LIST OF FIGURES

1. INTRODUCTION

Machine learning plays a major role from past years in image detection, spam reorganization, normal speech command, product recommendation and medical diagnosis. Present machine learning algorithm helps us in enhancing security alerts, ensuring public safety and improve medical enhancements. Machine learning system also provides better customer service and safer automobile systems

Machine learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from data. Machine learning methods are appropriate in big data since attempting to manually process vast volumes of data would be impossible without the support of machines. Machine learning in computer science attempts to solve problems algorithmically rather than purely mathematically. Therefore, it is based on creating algorithms that permit the machine to learn.

Data is the heart of machine learning. Predictive models use data for training which gives somewhat accurate results. Without data we can't train the model. Machine learning involves building these models from data and uses them to predict new data. It gives system capability to learn wherein it automatically learns and improves its performance without being explicitly programmed. It does focus on the development of programs and use it to learn for themselves. As the world is moving forward to using variants technologies, so has automation improved its ways to make our work easier.

Machine learning is closely related to statistics, which focuses on making predictions using computers. There are a variety of applications of Machine Learning such as filtering of emails, where it is difficult to develop a conventional algorithm to perform the task effectively. Machine learning algorithms are purely based on data.
Machine Learning algorithms are an advanced version of the regular algorithm. It makes programs "smarter" by allowing them to automatically learn from the data provided by us.

1.1 Machine Learning Classification

The algorithm is mainly divided into two phases and that is the training phase and the testing phase. Broadly there are three types of algorithms that are mainly used on data and they are supervised, unsupervised and reinforcement learning algorithms.

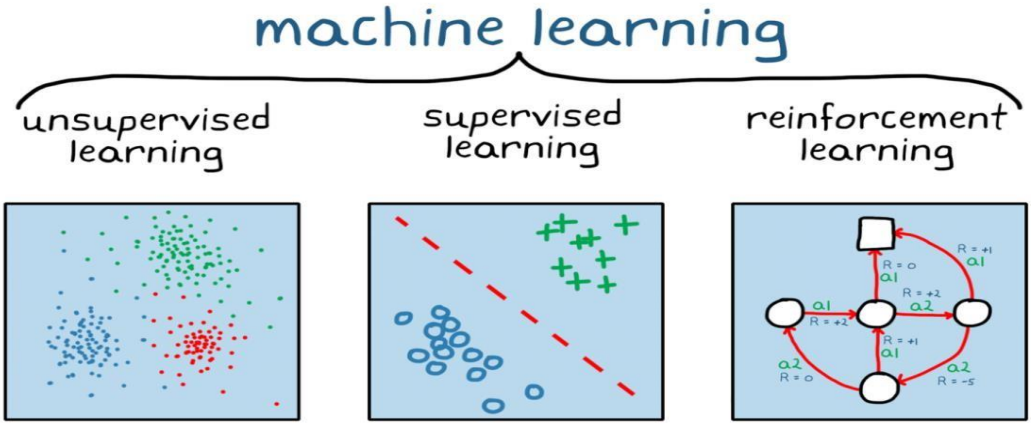


Fig 1: Machine Learning Classification

1. Supervised learning

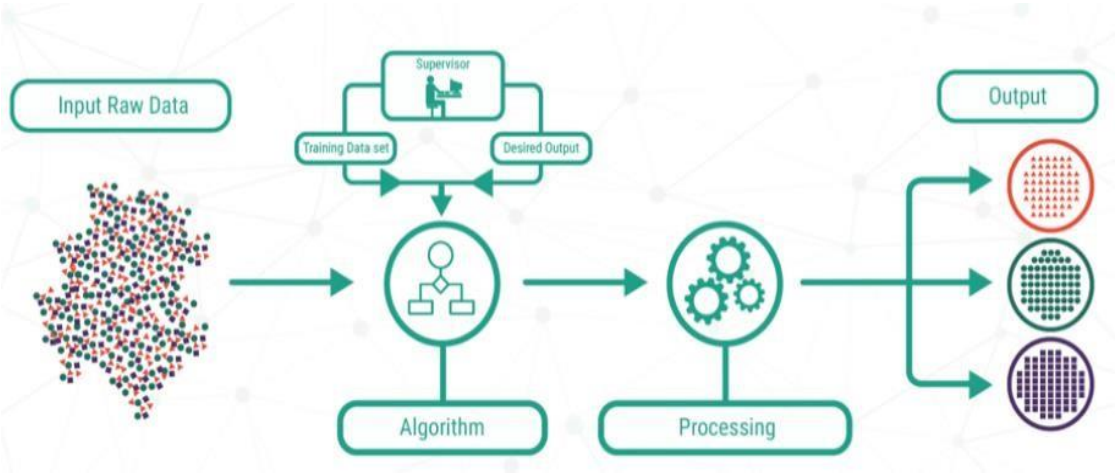


Fig 2: Supervised learning model

Supervised learning, the algorithm consists of a target variable or a dependent variable which is to be predicted from a set of independent variables. Using a function, the inputs are mapped to the desired outputs. Examples of Supervised learning algorithms are Regression, Decision Tree, Random Forest, KNN, Logistic Regression, etc.

2. Unsupervised learning

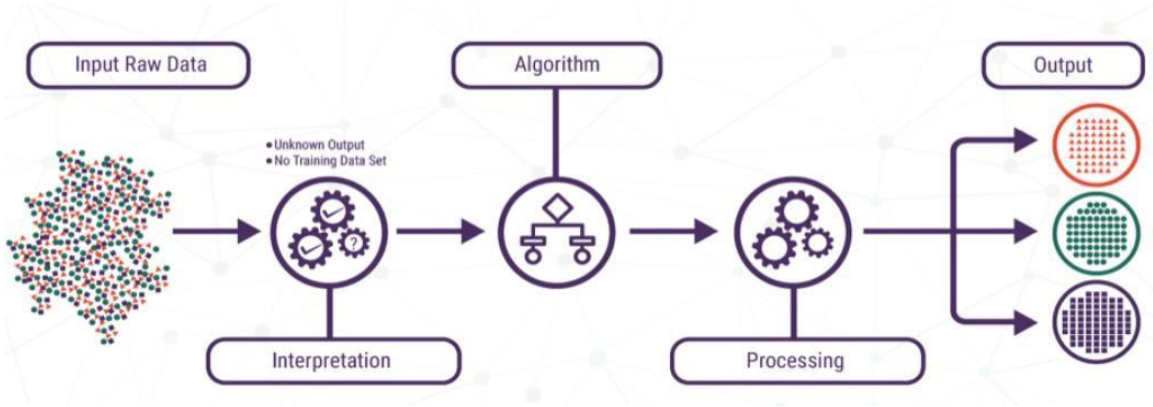


Fig 3: Unsupervised learning model

In Unsupervised learning, the algorithm does not have any target variable. It is used for clustering into different groups. Apriori algorithm, K-means, Principal Component Analysis, Independent Component Analysis are some examples of Unsupervised learning algorithms.

3. Reinforcement learning

Fig 4: Reinforcement Learning model

When the machine is used to make specific decisions, Reinforcement Learning is used. In this, the model is in an environment where it trains itself making it more accurate by using the trial-and-error methodology. The model hence learns from past experiences and it captures the knowledge about that domain to make accurate decisions, Example of Reinforcement Learning: Markov Decision Process-hot encoding is one such Reinforcement learning algorithm.

Several Machine Learning algorithms are used to solve problems in the real world today. However, some of them give better performance in certain circumstances. Thus, this thesis attempts to use linear regression algorithms, decision tree algorithm and neural network to compare their performance when it comes to predicting values of a given dataset. The performance will be measured upon predicting house prices since the prediction in many regression algorithms relies not only on a specific feature but on an unknown number of attributes that result in the value to be predicted. House prices depend on an individual house specification. Houses have a variant number of features that may not have the same cost due to its location. For instance, a big house may have a higher price if it is located in desirable rich area than being placed in a poor neighbourhood.

2. MATERIAL AND METHODS

2.1. Literature survey

There is a vast amount of work that is focused on training models to detect patterns in datasets to predict what the future output could be. However, there are researches where the authors use different machine learning algorithms with a combination of pre-processing data methods.

1. Debanjan Banerjee, Suchibrota Dutta, "Predicting the Housing Price Direction using Machine Learning Techniques" 2018.

This paper considers the issue of changing house price as a classification problem and applies machine learning techniques to predict whether house prices will rise or fall. This work applies various feature selection techniques such as variance influence factor, Information value, principal component analysis and data transformation techniques such as outlier and missing value treatment as well as box-cox transformation techniques. Random Forest provides more accuracy however at the same time this particular type of classifier also prone to over fitting therefore the performance of Support Vector Machine classifier can have said to be reliable and consistent over the rest of the two classifiers

2. Gaurav Kumar, Pradip Kumar Bhatiya, Fourth International Conference on Advanced Computing & Communication Technologies, "A Detailed Review of Feature Extraction in Image Processing Systems", 2014.

In this paper, there is discussion on which feature extraction technique is best by considering various types of features and feature extraction techniques. In case of character recognition application, we are going to refer features and feature extraction methods in this paper. Using this paper, a quick idea of feature extraction techniques may be got and it can be decided that which feature extraction technique will be better for the work to be done based on type of image

3. The Danh Phan, "Housing Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia", 2018.

In this paper, Machine learning techniques are applied to analyze historical property transactions in Australia to discover useful models for house buyers and sellers. There is the high inconsistency between house prices in the most expensive and most affordable suburbs in the city of Melbourne is showed in this paper. Moreover, experiments demonstrate that the combination of Stepwise and

Support Vector Machine that is based on mean squared error measurement is a competitive approach. This paper seeks useful models for house price prediction. It also provides insights into the Melbourne Housing Market. Firstly, the original data is prepared and transformed into a cleaned dataset ready for analysis. Data reduction and transformation are then applied by using Stepwise and PCA techniques. Different methods are then implemented and evaluated to achieve an optimal solution. The evaluation phase indicates that the combination of Step-wise and SVM model is a competitive approach. Therefore, it could be used for further deployment

4. A study was conducted in 2015 by Nils Landberg .

Nils analysed the price development on the Swedish housing market and the influences of qualitative variables on Swedish house prices. Landberg has studied the impact of square meter price, population, new houses, new companies, foreign background, foreign-born, unemployment rate, the number of breaks-in, the total number of crimes, the number of available jobs ranking. According to Nils, unemployment rate, number of crimes, interest rate, and new houses have a negative effect on house prices. Landberg showed that the real estate market is not easy to be analysed compared with goods market because many alternative costs are affecting the increase in house prices. The study shows that the increase in population and qualitative variables have a positive effect on house prices. Besides, it showed unemployment rate effects negatively on house prices, but the sale price and unemployment rate are not directly correlated with each other.

2.2. Hardware/Software Design

Hardware Specification is as follows:

1. Processor: Intel core i5 above
2. Ram: 8 GB
3. Display: Desktop
4. System Type: 64 Bit OS

Software Specification is as follows:

1. IDE: Anaconda(64bit)
2. Google colab or Jupyter Notebook

Algorithms

In this project, a machine learning model is proposed to predict a house price based on data related to the. During the development and evaluation of our model, we will show the code used for each step followed by its output. This will facilitate the reproducibility of our work.
In this study, Python programming language with a number of Python packages will be used.
The main objectives of this project are as follows:

- To apply data pre-processing and preparation techniques in order to obtain clean data.
- To build machine learning models able to predict house price based on different factors.
- To analyse and compare model's performance based on accuracy in order to choose the best model.
-

PCA (Principal Component Analysis)

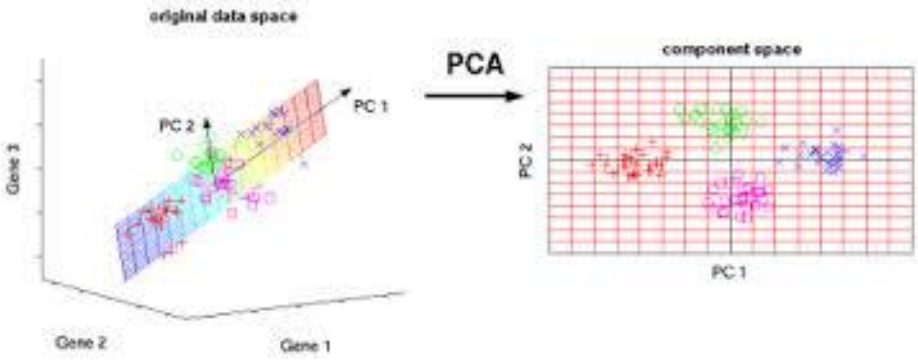


Fig 5: Principal component analysis

- PCA is a technique from linear algebra that can be used to automatically perform dimensionality reduction.
- Dimensionality reduction involves reducing the number of input variables or columns in modeling data.
- This is used because fewer input variables can result in simpler predictive model that may have better performance when making prediction on new data
- It does not select the features but combines old one. Mathematically, it performs linear transformation and moves original data to new space composed by principal component.
- Theoretically, it uses concept of variance matrix, covariance matrix, eigen matrix, eigen value pair to perform PCA.

LINEAR REGRESSION

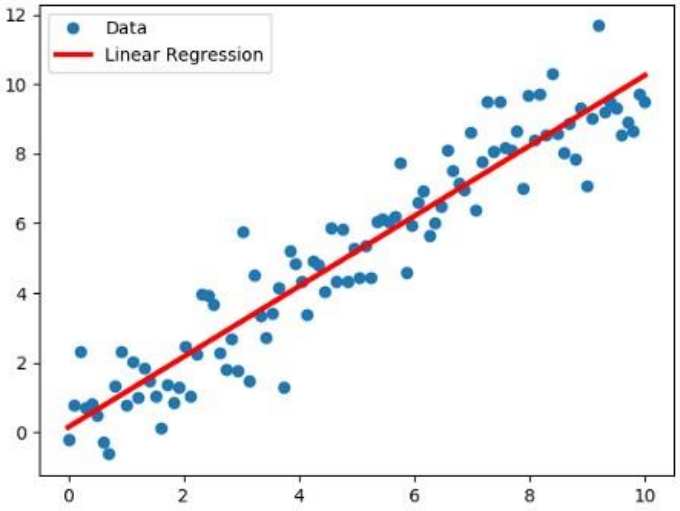


Fig 6: Linear Regression Example

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

A linear regression line has an equation of the form $\boldsymbol{Y = a + bX}$ where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses.

- If the goal is prediction, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
- If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

ARTIFICIAL NEURAL NETWORKS

Artificial neural network (ANN) is an attempt to simulate the work of a biological brain. The brain learns and evolves through the experiments that it faces through time to make decisions and predict the result of particular actions. Thus, ANN tries to simulate the brain to learn the pattern in a given data to predict the output of that data whether the expected data was provided in the learning process or not.

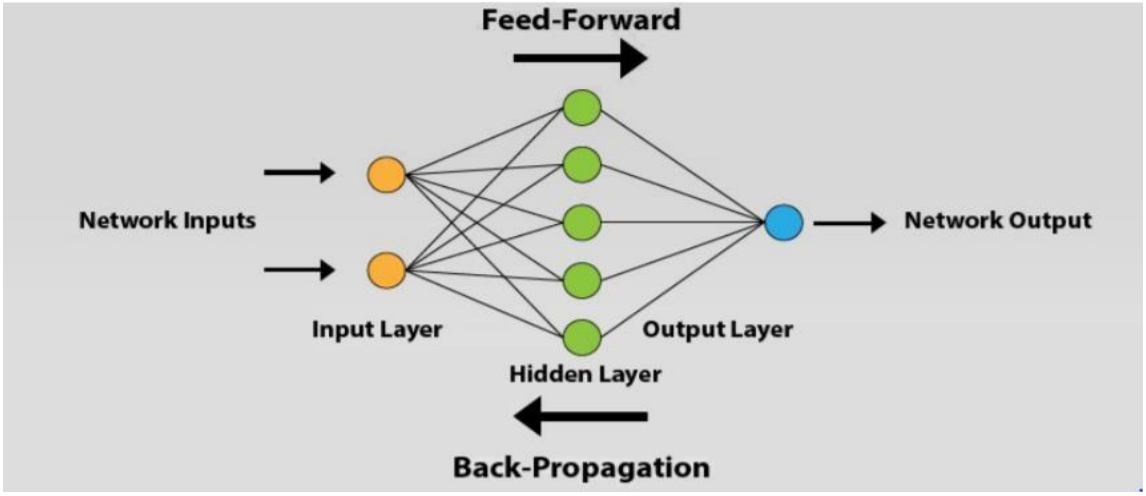


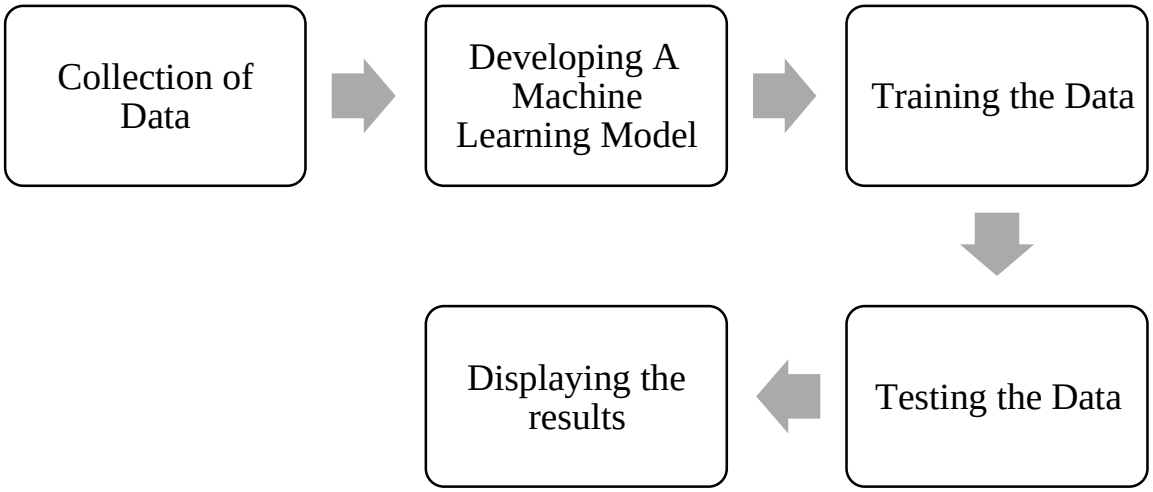
Fig 7: Artificial Neural Network model

Layers consist of at least three layers, input layer, one or more of hidden layers and output layer. Each layer holds a set of neurons that takes input and process data and finally pass the output to other neurons in the next layer.

- An input layer that accepts the independent variables or inputs of the model
- One hidden layer
- An output layer that generates predictions

Artificial Neural Networks are used for a variety of tasks, a popular use is for classification. You can collect datasets of images for example of different breeds of dogs and then train a neural network on the images, then if you supply a new image of a dog it will give a statistical score on how closely the new image matches the model and then will output what breed of dog the image is. Neural Networks are also used in Self Driving cars, Character Recognition, Image Compression, House price Prediction, and lots of other interesting applications.

2.3. Block Diagram



3. RESULTS

1. COLLECTION OF DATA:

Dataset used is USA Housing Data set provided by the Kaggle. This data gives different sales price with respect to type of houses in USA.

```
uh=pd.read_csv('USA_Housing')
uh
```

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386
...	...	...	...	...	...	...	...
4995	60567.944140	7.830362	6.137356	3.46	22837.361035	1.060194e+06	USNS Williams\nFPO AP 30153-7653
4996	78491.275435	6.999135	6.576763	4.02	25616.115489	1.482618e+06	PSC 9258, Box 8489\nAPO AA 42991-3352
4997	63390.686886	7.250591	4.805081	2.13	33266.145490	1.030730e+06	4215 Tracy Garden Suite 076\nJoshualand, VA 01...
4998	68001.331235	5.534388	7.130144	5.44	42625.620156	1.198657e+06	USS Wallace\nFPO AE 73316
4999	65510.581804	5.992305	6.792336	4.07	46501.283803	1.298950e+06	37778 George Ridges Apt. 509\nEast Holly, NV 2...

5000 rows × 7 columns

Fig 8: Collection of Data

2. DATA PRE-PROCESSING:

Our approach is to input USA Housing dataset provided by the Kaggle and do the primary analysis on the dataset. This is the small step which involves taking input dataset as data frames. The dataset has 7 columns and 5000 rows.

```
print('Housing Data: \n')
print("Number of columns: "+ str(uh.shape[1]))
print("Number of rows: "+ str(uh.shape[0]))
```

```
Housing Data:

Number of columns: 7
Number of rows: 5000
```

```
[ ]  uh.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 7 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   Avg_Area_Income              5000 non-null   float64
 1   Avg_Area_House_Age           5000 non-null   float64
 2   Avg_Area_Number_of_Rooms     5000 non-null   float64
 3   Avg_Area_Number_of_Bedrooms  5000 non-null   float64
 4   Area_Population              5000 non-null   float64
 5   Price                        5000 non-null   float64
 6   Address                      5000 non-null   object 
dtypes: float64(6), object(1)
memory usage: 273.6+ KB
```

	Avg_Area_Income	Avg_Area_House_Age	Avg_Area_Number_of_Rooms	Avg_Area_Number_of_Bedrooms	Area_Population	Price
0	79545.45857	5.682861	7.009188	4.09	23086.80050	1.059034e+06
1	79248.64245	6.002900	6.730821	3.09	40173.07217	1.505891e+06
2	61287.06718	5.865890	8.512727	5.13	36882.15940	1.058988e+06
3	63345.24005	7.188236	5.586729	3.26	34310.24283	1.260617e+06
4	59982.19723	5.040555	7.839388	4.23	26354.10947	6.309435e+05

Fig 9: Description of Data

3. DATA VISUALIZATION:

We study the data closely through various graphs, to determine any trends in the data. We use the seaborn package to plot the graph.

```
sns.heatmap(uh.corr(),annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f47f80b0390>
```

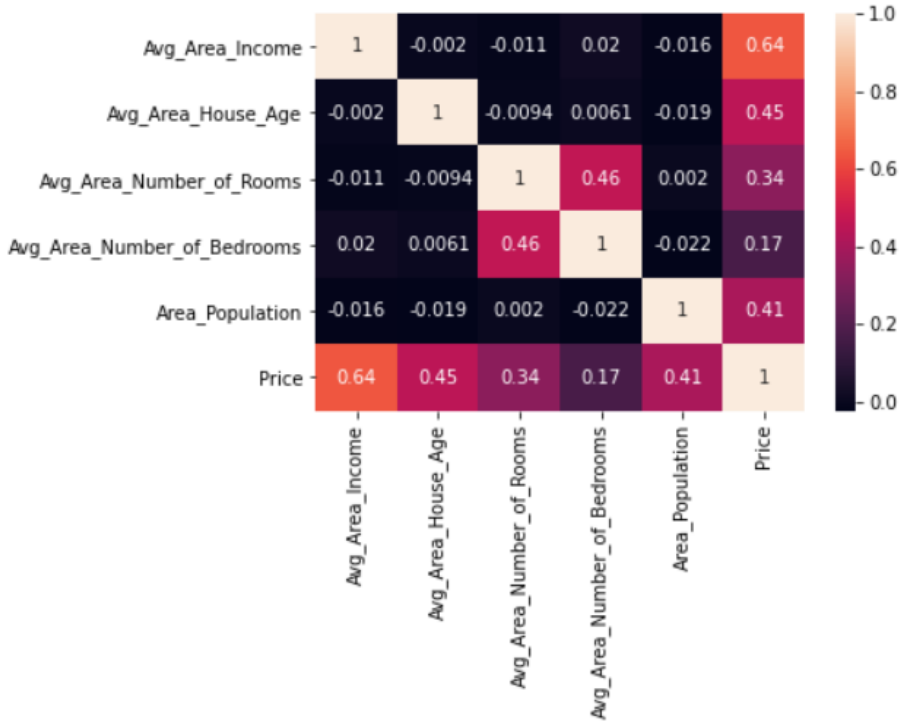


Fig 10: HeatMap of Data

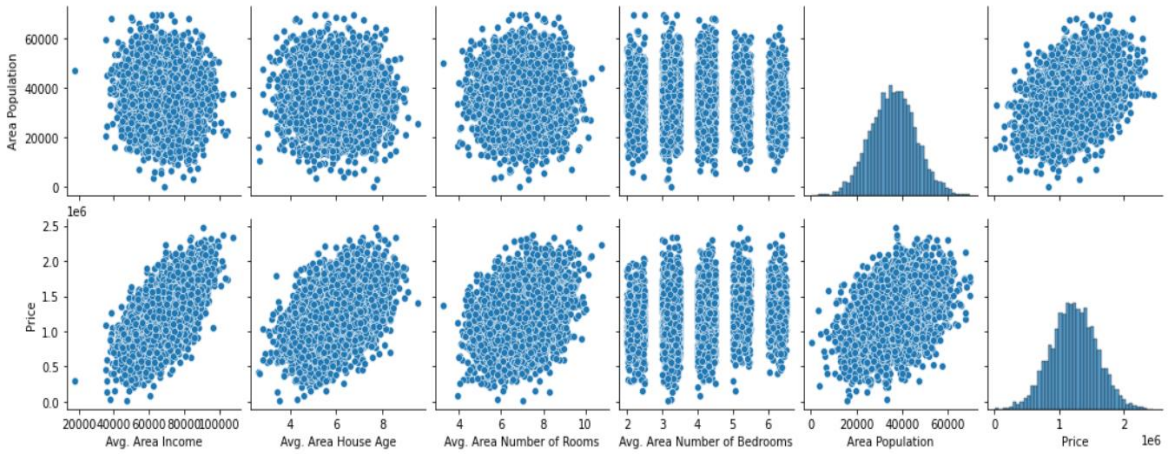


Fig 11: Pair Plot of Data

The second graph is a distribution of Sale price. The below graph appears to be somewhat right skewed. This suggests that mean for sale price is greater than its median.

4. APPLYING ALGORITHMS

LINEAR REGRESSION AND PCA

Input Code:

```
# splitting the dataset in to training set and test set
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = 1/3.0, random_state = 0)
#Fitting simple linear Regression to training set

# Applying PCA
from sklearn.decomposition import PCA
pca = PCA(n_components = 1)
x_train = pca.fit_transform(x_train)
x_test = pca.transform(x_test)
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(x_train, y_train)

#predicting the test set results
y_pred = regressor.predict(x_test)

#visualising the training set results
plt.figure(figsize=(8, 8))
plt.scatter(x_train, y_train, color = 'green')
plt.scatter(x_test, y_test, color = 'red')
plt.plot(x_train, regressor.predict(x_train), color = 'blue')
plt.title('House Price Predictions')
plt.xlabel('features')
plt.ylabel('price')
plt.show()
```

Output:

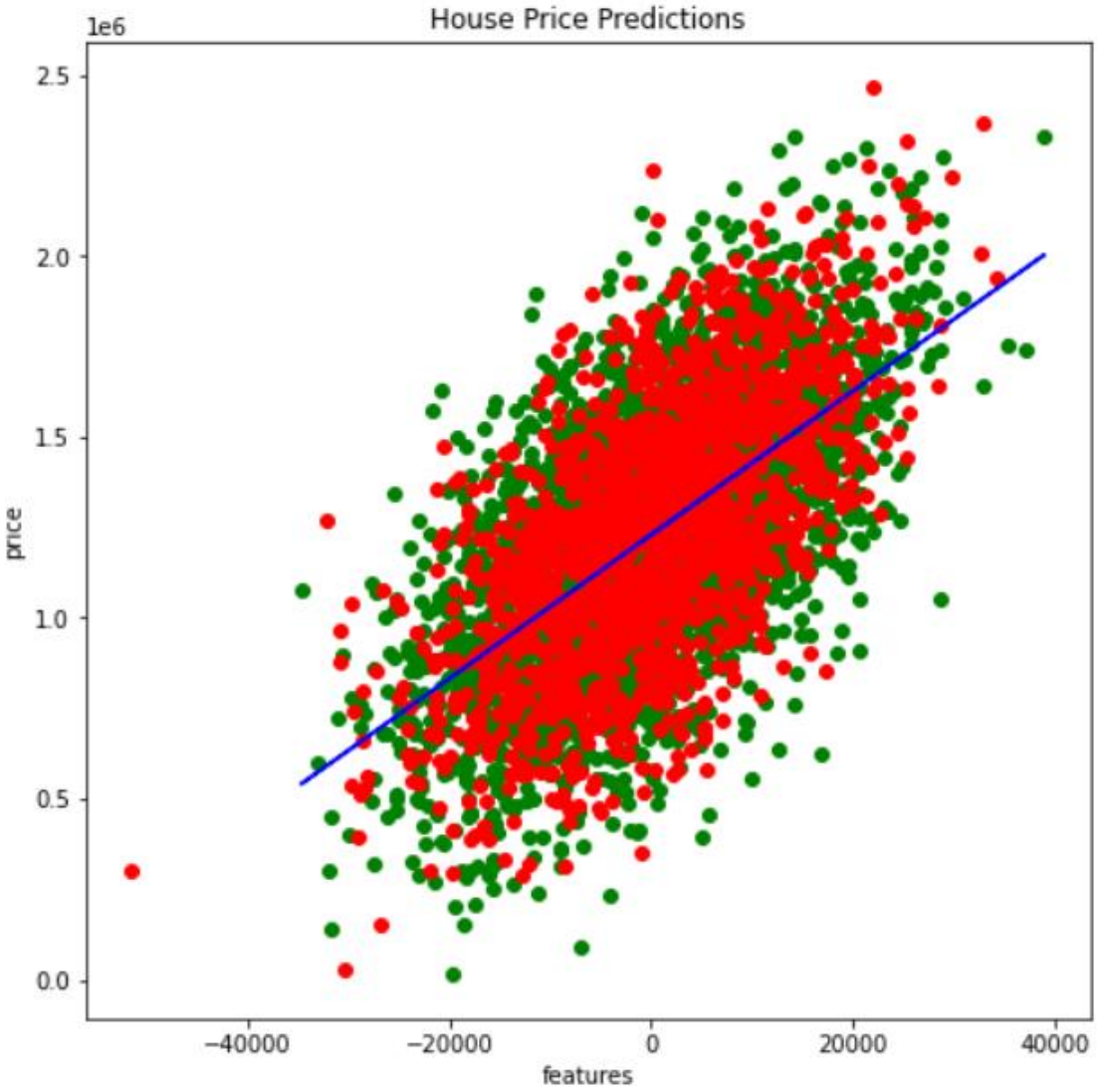


Fig 12: Output of Linear Regression

NEURAL NETWORK:

Neural networks are applied using two models in this project: TensorFlow and keras. TensorFlow is an end-to-end open-source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.

Keras is an open-source neural-network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, R, Theano, or PlaidML. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible.

For creating a baseline neural network model, we have to Start with all of the needed functions and objects. Since we have 5 features, let's insert 5 neurons as a start, 2 hidden layers and 1 output layer due to predict house Price.

Also, ADAM optimization algorithm is used for optimizing loss function (Mean squared error). Then, we train the model for 100 epochs, and each time record the training and validation accuracy in the history object. To keep track of how well the model is performing for each epoch, the model will run in both train and test data along with calculating the loss function.

Input Code:

```
# Create a model
def get_model():
  model = Sequential([
      Dense(10, input_shape = (5,), activation = 'relu'),
      Dense(10, activation = 'relu'),
      Dense(5, activation = 'relu'),
      Dense(1)
      ])
  model.compile(
      loss = 'mse',
      optimizer = 'adam'
  )
  return model
```

```
ann = get_model()
ann.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 10)                60
_________________________________________________________________
dense_1 (Dense)              (None, 10)                110
_________________________________________________________________
dense_2 (Dense)              (None, 5)                 55
_________________________________________________________________
dense_3 (Dense)              (None, 1)                 6
=================================================================
Total params: 231
Trainable params: 231
Non-trainable params: 0
```

```
plot_loss(history)
```

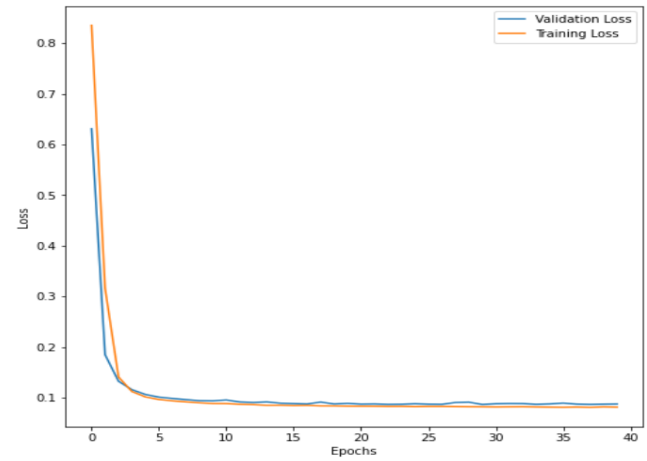


Fig 13: plot the loss function measure on the training and validation sets

```
compare_predictions(preds_on_untrained, preds_on_trained, y_test)
```

```
/usr/local/lib/python3.6/dist-packages/numpy/core/_asarray.py:136: VisibleDeprecationWarning:
  return array(a, dtype, copy=False, order=order, subok=True)
```

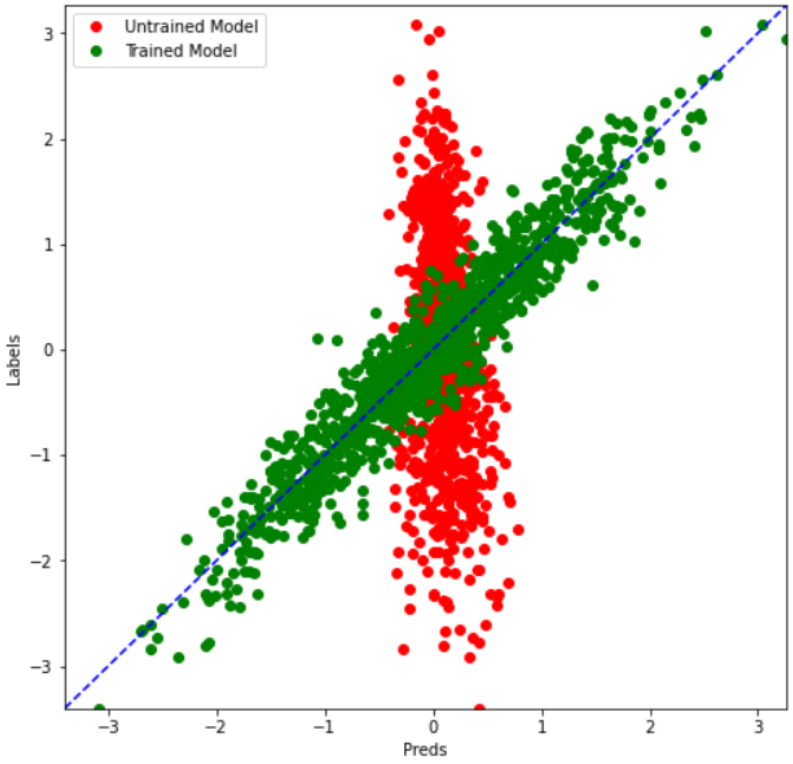


Fig 14: Models before Training and after training

Predicting the Prices of House Based on The Given Inputs

```
Avg_Area_Income = 79545
Avg_Area_House_Age = 6
Avg_Area_Number_of_Rooms = 7
Avg_Area_Number_of_Bedrooms = 4
Area_Population = 23085

input = np.array([[Avg_Area_Income,
                   Avg_Area_House_Age,
                   Avg_Area_Number_of_Rooms,
                   Avg_Area_Number_of_Bedrooms,
                   Area_Population]])

input = (input-input_mean) / (input_std)
output = model.predict(input)

output = output*output_std + output_mean

print("Price of the house will be", output[0,0])
```

```
Price of the house will be 1254640.0
```

4. CONCLUSIONS

The goal is to achieve the system which will reduce the human effort to find a house having reasonable price. The proposed system. House Price Prediction model approximately try to achieve the same one. Proposed system focused on predict the house price according to the area for that image processing and machine learning methods are used. The experimental results showed that this technique that are used while developing system will give accurate prediction of house price.

This study aims to analyse the accuracy of predicting house prices when using Linear Regression and Artificial neural network (ANN).

5. SCOPE FOR FURTHER WORK

The project is correctly classified into training, testing and Prediction process. More data will help in getting a better accuracy as we have enough experience to give to our model.

Once data is augmented and there is enough data to make our model, Machine will be trained using the training data then test with testing data and then predicts the house price. Different algorithms like linear regression and neural network accuracy will be recorded, and which will further go for implementation of the different algorithms and later the models will be compared based on accuracy of prediction.

6. REFERENCES

1. https://www.kaggle.com/vedavyasv/usa-housing
2. David E. Rapach, Jack K. Strauss “Forecasting real housing price growth in the Eighth District states”
3. https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471