A Synopsis

Submitted in Partial fulfillment of the requirements for the Final Year

Project of **HOUSE PRICE PRIDICTION USING MACHINE LEARNING**

**Bachelor of Computer Application**

**By**

**Gaurav Soni (Roll no.: 0204696 Sec: B)**

**Gyanendra Kumar (Roll no.: 0204698 Sec: B)**

**Harshit Shukla (Roll no.: 0204708 Sec: B)**

Under Guidance Of

Dr. **Mayur Rahul Sir**

****

Department Of Computer Application

UIET

CSJM UNIVERSITY Kanpur

2019-2022

**ACKNOWLEDGEMENT**

I express my deep sense of gratitude to my **AMIT VIRMANI SIR (Computer Application)** for the valuable guidance and for permitting us to carry out this project.

With gratitude,

## Name : ***GAURAV SONI*(Roll no.: 0204696 Sec: B)**

###### Name : **GYANENDRA KUMAR(Roll no.: 0204698 Sec: B)**

## Name : **HARSHIT SHUKLA(Roll no.: 0204708 Sec: B)**

**CONTENTS**

1. Introduction Of Project
2. Objective Of The Project
3. Scope Of The Project
4. Project Plan ( Initial Phase)
5. The Existing System
6. Current Problems
7. Areas For Improvement
8. Proposed System
9. Input/output Requirement
10. Software Requirement
11. Conclusion
12. References

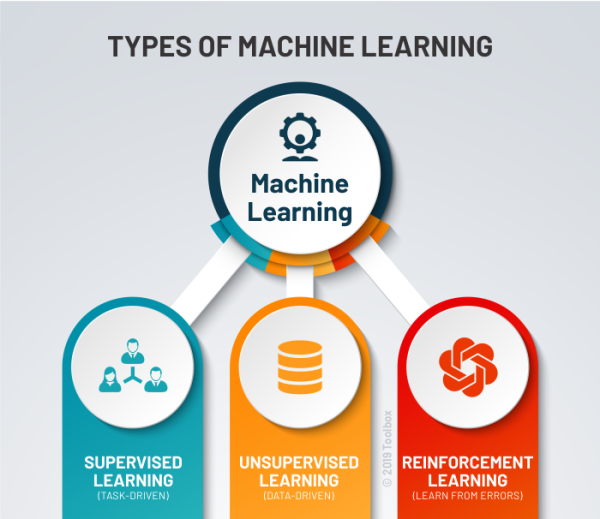
**  
 Introduction of The House Price Priidiction**

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. However, there are two general groups in machine learning which are supervised and unsupervised. Supervised is where the program gets trained on pre-determined set to be able to predict when a new data is given. Unsupervised is where the program tries to find the relationship and the hidden pattern between the data.

**Types Of Machine Learning**

There are two types of machine learning. These are following

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Reinforcement Machine Learning



1. **Supervised Machine Learning**

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labeled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

**For instance**, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all different fruits one by one like this: 

* If the shape of the object is rounded and has a depression at the top, is red in color, then it will be labeled as –**Apple**.
* If the shape of the object is a long curving cylinder having Green-Yellow color, then it will be labeled as –**Banana**.

Now suppose after training the data, you have given a new separate fruit, say Banana from the basket, and asked to identify it. 

Since the machine has already learned the things from previous data and this time has to use it wisely. It will first classify the fruit with its shape and color and would confirm the fruit name as BANANA and put it in the Banana category. Thus the machine learns the things from training data(basket containing fruits) and then applies the knowledge to test data(new fruit).

Supervised learning is classified into two categories of algorithms: 

* **Classification**: A classification problem is when the output variable is a category, such as “Red” or “blue” or “disease” and “no disease”.
* **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Supervised learning deals with or learns with “labeled” data. This implies that some data is already tagged with the correct answer.

1. **Unsupervised Machine Learning**

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.   
**For instance**, suppose it is given an image having both dogs and cats which it has never seen. 

Thus the machine has no idea about the features of dogs and cats so we can’t categorize it as ‘dogs and cats ‘. But it can categorize them according to their similarities, patterns, and differences, i.e., we can easily categorize the above picture into two parts. The first may contain all pics having **dogs** in them and the second part may contain all pics having **cats** in them. Here you didn’t learn anything before, which means no training data or examples.

 It allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with unlabelled data.

Unsupervised learning is classified into two categories of algorithms: 

* **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
* **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

1. **Reinforcement Machine Learning**

Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.

In Reinforcement Learning, the agent learns automatically using feedbacks without any labeled data, unlike supervised learning.

RL solves a specific type of problem where decision making is sequential, and the goal is long-term, such as **game-playing, robotics**, etc.

The agent interacts with the environment and explores it by itself. The primary goal of an agent in reinforcement learning is to improve the performance by getting the maximum positive rewards.

The agent learns with the process of hit and trial, and based on the experience, it learns to perform the task in a better way. Hence, we can say that ***"Reinforcement learning is a type of machine learning method where an intelligent agent (computer program) interacts with the environment and learns to act within that."*** How a Robotic dog learns the movement of his arms is an example of Reinforcement learning.

Since there is no labeled data, so the agent is bound to learn by its experience only.

## **Terms used in Reinforcement Learning**

* **Agent():** An entity that can perceive/explore the environment and act upon it.
* **Environment():** A situation in which an agent is present or surrounded by. In RL, we assume the stochastic environment, which means it is random in nature.
* **Action():** Actions are the moves taken by an agent within the environment.
* **State():** State is a situation returned by the environment after each action taken by the agent.
* **Reward():** A feedback returned to the agent from the environment to evaluate the action of the agent.
* **Policy():** Policy is a strategy applied by the agent for the next action based on the current state.
* **Value():** It is expected long-term retuned with the discount factor and opposite to the short-term reward.
* **Q-value():** It is mostly similar to the value, but it takes one additional parameter as a current action .

Several Machine Learning algorithms are used to solve problems in the real world today. However, some of them give better performance in certain circumstances, as stated in the No Free Lunch Theorem. Thus, this thesis attempts to use regression algorithms and artificial neural network (ANN) to compare their performance when it comes to predicting values of a given dataset.

**No Free Lunch Theorem:** The “no free lunch” (NFL) theorem for supervised machine learning is a theorem that essentially implies that **no single machine learning algorithm is universally the best-performing algorithm for all problems**.

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.

The data used in the experiment will be handled by using a combination of pre-processing methods to improve the prediction accuracy. In addition, some factors will be added to the local dataset in order to study the relationship between these factors and the sale price.

**Objective Of Project**

There are following objective of House Price Prediction Machine Learning. These are

1. Predict the sale price for each house.
2. Minimize the difference between predicted and actual rating.

**Scope Of Project**

It is a website at which you have to enter required input to get price.

**Project Plan ( Initial Phase)**

1. **Data Collection:** We take data from online data repository site. The information includes many variables such as Price, Rera(**Real Estate Regulatory Authority**) approved or Not, BHK \_No, Total Square Feet - size of the property in square feet , Category marking Ready to move or Not, Address of the property, Under Construction or Not, Category marking Resale or not, Type of property.
2. **Data Analysis:**  We analyze the data by making graphs, checking missing data points and removing outliers.

**We use different regression model to know which model predict best price.**

1. **Linear Regression:** Linear regression is a supervised learning technique. It is responsible for predicting the value of a dependent variable (Y) based on a given independent variable (X). It is the connection between the input (X) and the output (Y). It is one of the most well-known and well-understood machine learning algorithms. Simple linear regression, ordinary least squares, Gradient Descent, and Regularization are the linear regression models.
2. **Decision Tree Regression:** It is an object that trains a tree-structured model to predict data in the future in order to provide meaning ful continuous output. The core principles of decision trees, Maximizing Information Gain, Classification trees, and Regression trees are the processes involved in decision tree regression. The essential notion of decision trees is that they are built via recursive partitioning. Each node can be divided into child nodes, beginning with the root node, which is known as the parent node. These nodes have the potential to become the parent nodes of their resulting offspring nodes. The nodes at the informative features are specified as the maximizing information gain, to establish an objective function that is to optimize the tree learning method.
3. **Random Forest Regression** It is an essential learning approach for classification and regression to create a large number of decision trees. Preliminaries of decision trees are common approaches for a variety of machine learning problems. Tree learning is required for serving n off the self-produce for data mining since it is invariant despite scaling and several other changes. The trees are grown very deep in order to learn a high regular pattern. Random forest is a method of averaging several deep decision trees trained on various portions of the same training set. This comes at the price of a slight increase in bias and some interoperability
4. **Performance evaluation** **:** We evaluate performance using following method.
5. **Mean Square Error** **:** Mean Square Error is a measure for how close the estimation is relative to the actual data. It measures the average of the square of the errors deviation of the estimated values with respect to the actual values. It is measured by:

MSE = (1 /n) ∑ni=1 (y1−y) 2

Where y1 is the estimated value from the regression and y is the actual value. The lower the MSE, the better the estimation model.

### Root Mean Squared Error

Root mean squared error (RMSE) is the square root of the mean of the square of all of the error. The use of RMSE is very common, and it is considered an excellent general-purpose error metric for [numerical predictions](https://www.sciencedirect.com/topics/engineering/numerical-prediction).

RMSE=(1/n(∑i=1n(Si-Oi)2))1/2

where Oi are the observations, Si predicted values of a variable, and n the number of observations available for analysis. RMSE is a good measure of accuracy, but only to compare prediction errors of different models or model configurations for a particular variable and not between variables, as it is scale-dependent.

* **Steps to Create Model**

1. **Import Libraries:** library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs**.**

We use following library in python

Pandas

Numpy

Sklearn

seaborn

1. **Load Dataset:** It is process of reading and loading of data from csv file.
2. **Exploratory Data Analysis:** EDA is a phenomenon under data analysis used for gaining a better understanding of data aspects like:   
   – main features of data   
   – variables and relationships that hold between them   
   – identifying which variables are important for our problem
3. **Data Cleaning:** Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.
4. **Feature Engineering: Feature engineering** (or **feature extraction**) is the process of using [domain knowledge](https://en.wikipedia.org/wiki/Domain_knowledge) to extract [features](https://en.wikipedia.org/wiki/Feature_(machine_learning)) (characteristics, properties, attributes) from raw [data](https://en.wikipedia.org/wiki/Data). The motivation is to use these extra features to improve the quality of results from a [machine learning](https://en.wikipedia.org/wiki/Machine_learning)  process, compared with supplying only the raw data to the machine learning process.
5. **Outlier Removal:** An Outlier is a data-item/object that deviates significantly from the rest of the (so-called normal) objects. They can be caused by measurement or execution errors. The analysis for outlier detection is referred to as outlier mining.
6. **Data Visualization:** Data visualization is the graphical representation of information and data.
7. **Building a Model:** After passing through above steps, we create appropriate model.
8. **Test the Model for few properties**: After building model now we test model and check rmse.
9. **Export the tested model to a joblib file:**Now we save tested model using joblib library.

* **Technologies used**

1. Python
2. HTML
3. CSS

* **Tools Used**

1. Visual Studio
2. Jupyter Notebook

**The Existing System**

The present system is not dunce proof and has certain drawbacks. Being a manual system the possible limitations and loopholes in the present system is large. Some of them are:-

1. **HUMAN resource: -** The current system has too much manual work from filling a form to filing a document, delivering manifesto. This increases burden on workers but does not yield the results it should.
2. **THORNY Job: -** In current system if any modification is to be made it increases manual work and is error prone.
3. **ERROR: -** As the system is managed and maintained by workers errors are some of the possibilities.

**Current Problems**

The Existing System cannot predict the exact price of house that is why sometime many company faces loss.

**Areas For Improvement**

There are following areas for improvement.

1. Minimize the difference between predicted and actual rating.
2. Design such model to predict best price of house .

**Proposed System**

* Proposed system is an website.
* Proposed system uses the parameters such as BHK\_NO, size of the property, location and other parameters for house price prediction.
* Proposed system is built using visual studio as front-end technology.

**Input/Output Requirements**

**Input:**

1. Rera(**Real Estate Regulatory Authority**) approved or Not
2. BHK \_No
3. Total Square Feet - size of the property in square feet
4. Category marking Ready to move or Not
5. Address of the property
6. Under Construction or Not
7. Category marking Resale or not
8. Type of property(BHK stands for **B= bed, H= hall and K= kitchen &** RK means basically **a 1 room kitchen set**.)

**Output:**

House price prediction using ML supervised learning technique.

**Software Requirement**

We use following software to develop our project

1. Python interpreter
2. Visual Studio/Jupyter Notebook

**Conclusion**

As we discussed about machine learning and its type briefly and collect information from various online sources. We conclude that Supervised learning is used to design House Price Prediction. Here we have to know real entity i.e. price so use regression to solve this. After a analysis now we start implementation of this project.

**References**

1. <https://datasetsearch.research.google.com/search?query=house%20price%20%20pridiction&docid=L2cvMTFxbDFxMjVxaw%3D%3D>
2. <https://www.kaggle.com/anmolkumar/house-price-prediction-challenge>
3. <https://towardsdatascience.com/tagged/house-price-prediction?p=dea265cc3154>
4. <https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/>
5. <https://www.geeksforgeeks.org/machine-learning/>