BANGALORE HOUSING PRICE PREDICTION

The main aim of the Project is to perform Exploratory Data Analysis on the dataset and further predict the correct housing prices for a given locality in Bangalore using different Machine Learning Techniques.

MOTIVATION:

Nowadays buying a perfect house needs several precalculated measurements and personal requirements as well as the most important thing, the prices of housing.

So for this Project, the main motivation is :

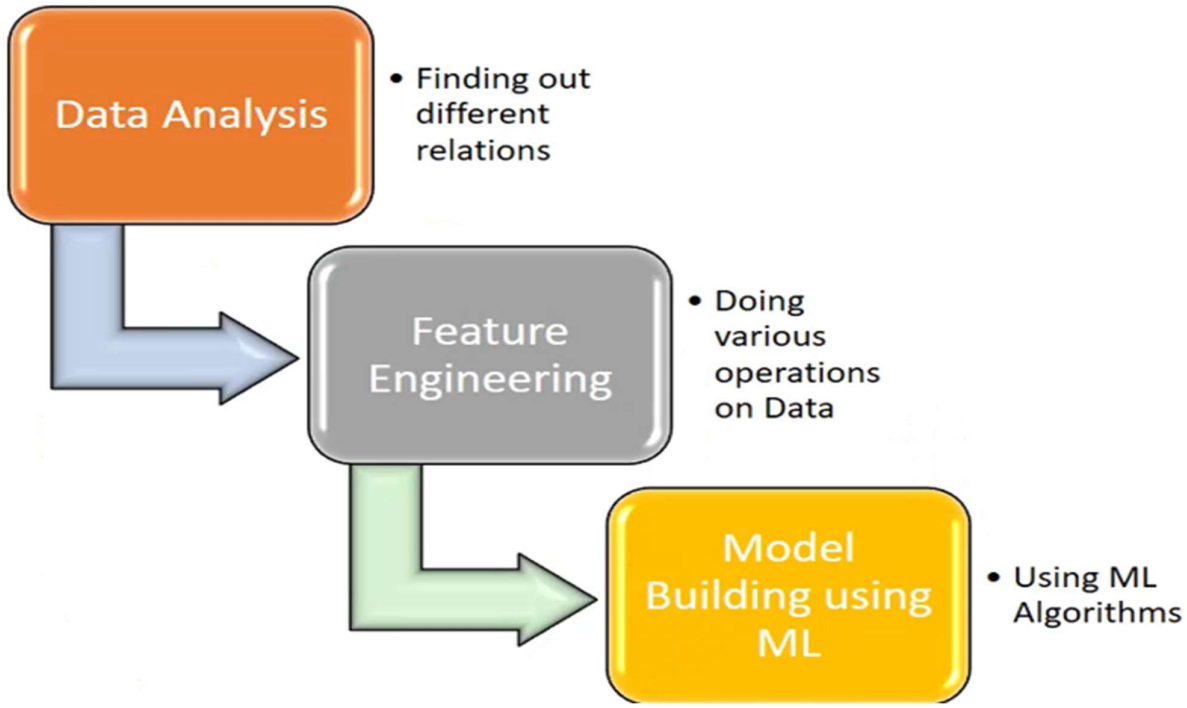
What are the things that a potential home buyer considers before purchasing a house? The Location, the Size of the Property, vicinity to offices, schools, parks, restaurants, hospitals or the stereotypical white picket fence?

What about the most important factor – The Price??

ABOUT THE PROJECT:

We will create a Machine Learning Model which will help us in predicting the prices of the housing prices of a locality in Bangalore on iputting some of the attributes. Here we will be provided with a dataset containing information about the Housing and its Prices.

TIME LINE THE PROJECT:



ABOUT THE DATASET:

Dataset is downloaded from here: https://www.kaggle.com/amitabhajoy/bengaluru-house-price-data

Predicting the price of houses in Bengaluru is based on the factors like:

1) Location,
2) Size(bhk),
3) Total Square Feet (Area),
4) Price (Most Important), and many more.

Thus this Project will revolve around Supervised Machine Learning Algorithms to predict the price of house in Bangalore based on location, number of bedrooms, number of bathrooms, etc.

Some of the Machine Learning Algorithms used are mentioned as follows:

1) Linear Regression
2) Lasso Regression
3) Decision Tree Regression

In the dataset, the dependent variable according to our proposed models is 'price' and rest columns are independent though some of them are reduntant for prediction of price.

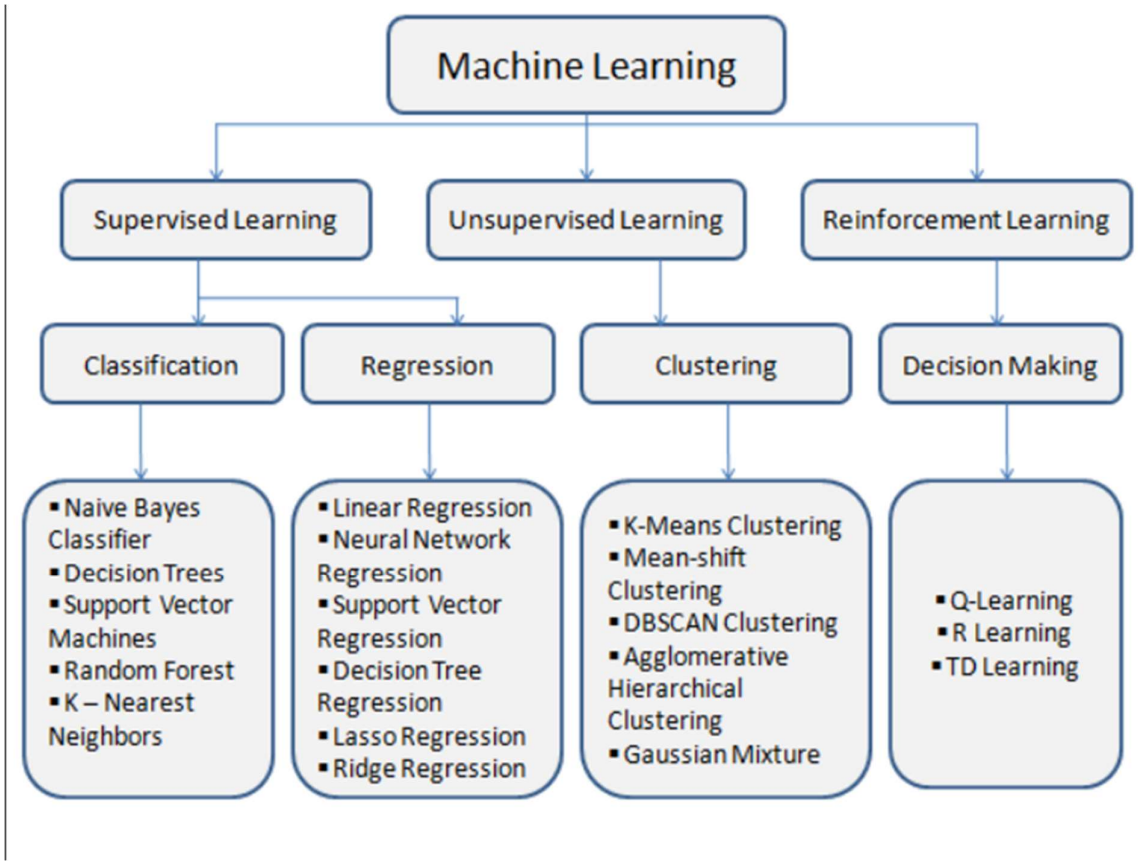
But before all these, we should have basic understanding of Machine Learning and Deep Learning Concepts!

WHAT IS MACHINE LEARNING ?

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

Supervised learning is when the model is getting trained on a labelled dataset. A labelled dataset is one that has both input and output parameters. In this type of learning both training and validation, datasets are labelled.

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.



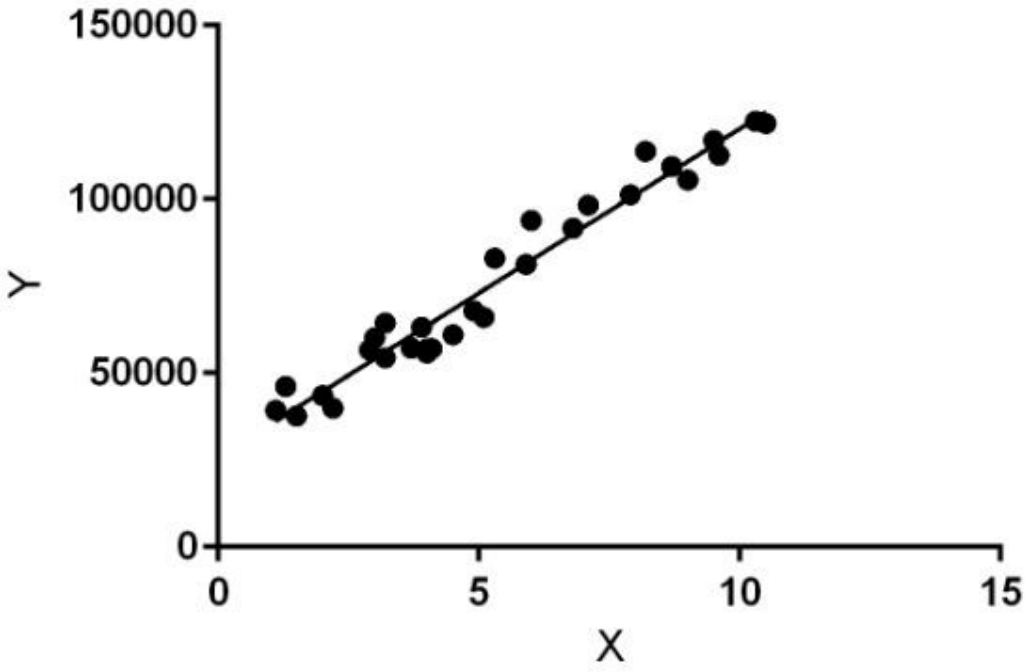
LINEAR REGRESSION

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behavior of a particular variable. For example, in finance, linear regression might be used to understand the relationship between a company's stock price and its earnings or to predict the future value of a currency based on its past performance.

One of the most important supervised learning tasks is regression. In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.



Linear Regression

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best-fit line for our model.

ASSUMPTIONS FOR LINEAR REGRESSION MODEL:

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

1) **Linearity**: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.

2) **Independence:** The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
3) **Homoscedasticity**: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.

4) **Normality:** The errors in the model are normally distributed.

5) **No multicollinearity**: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

LASSO REGRESSION

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Lasso Regression uses L1 regularization technique (will be discussed later in this article). It is used when we have more features because it automatically performs feature selection.

DECISION TREE REGRESSION

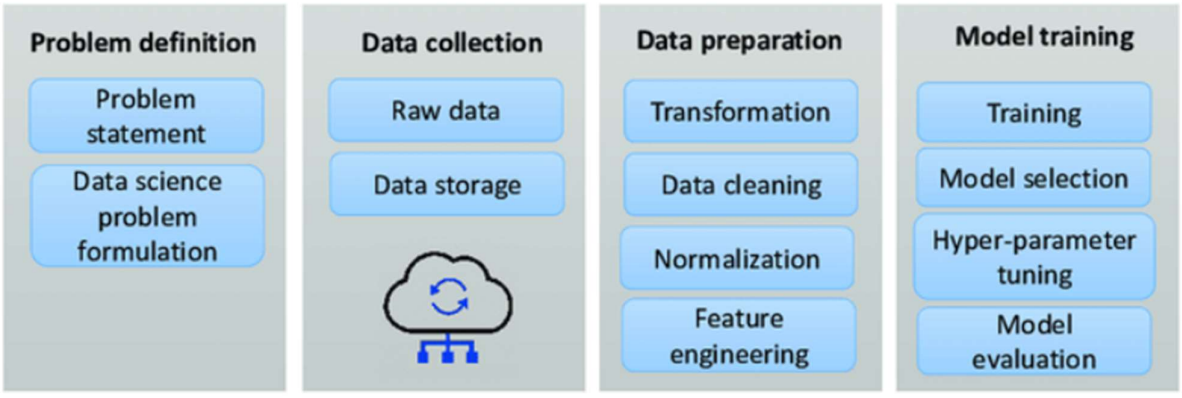
Decision tree learning is a supervised learning approach used in statistics, data mining and machine learning. In this formalism, a classification or regression decision tree is used as a predictive model to draw conclusions about a set of observations.

Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. More generally, the concept of regression tree can be extended to any kind of object equipped with pairwise dissimilarities such as categorical sequences.

Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making).

BASIC WORKING PIPELINE OF THE PROJECT



The following steps are used to create a Machine Learning Project using predefined dataset:

1) **Data Collection:** The very first step is to collect the data and required dependencies.

2) **Data Analysis and Data Preprocessing**

2.1 Data Analysis: After that, we used to analyze the dataset, about it's behaviour, trend and changes with respect to independent attributes. We also have to analyze the dependencies between every attribute, such that our model must be fitted perfectly.

2.2 Data Cleaning and Preprocessing: This is the most important step in model creation. The Cleaning of our dataset, such that the good data can be used for next process and unrequired/bad data should be removed so that the condition of overfitting and underfitting won't occur.

3) **Feature Engineering:** It is a machine learning technique that leverages data to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

4) **Data Splitting (Train-Test):** Now we will spilt the dataset into two parts, the training part will be used to create the model and testing part will be used to verify the results of our model.

5) At last, we will calculate the accuracy of the Machine Learning Models and see which one performs the best.

REQUIRED DEPENDENCIES

1) Numpy
2) Pandas
3) Matplotlib
4) Scikit-learn
5) Seaborn
6) GridSearch CV

RESULTS

Best Fit – Linear Regression -->78.5% accuracy

We have obtained same results i.e, Linear Regression gives the best fit, through GridSearch CV which has compared all the three Algorithms used.

	model	best_score	best_params
0	linear_regression	-7.903557e+16	{'normalize': False}
1	lasso	6.274809e-01	{'alpha': 1, 'selection': 'random'}
2	decision_tree	6.526613e-01	{'criterion': 'friedman_mse', 'splitter': 'best'}

CONCLUSION

Since Linear Regression model gives the best fit and explains nearly 78.5% of the variation of the data which is by far the highest among all the Algorithms used, I am considering Linear Regression Model as our predictive model.