PREDICTING HOUSE PRICE USING MACHINE LEARNING

Machine learning algorithms:

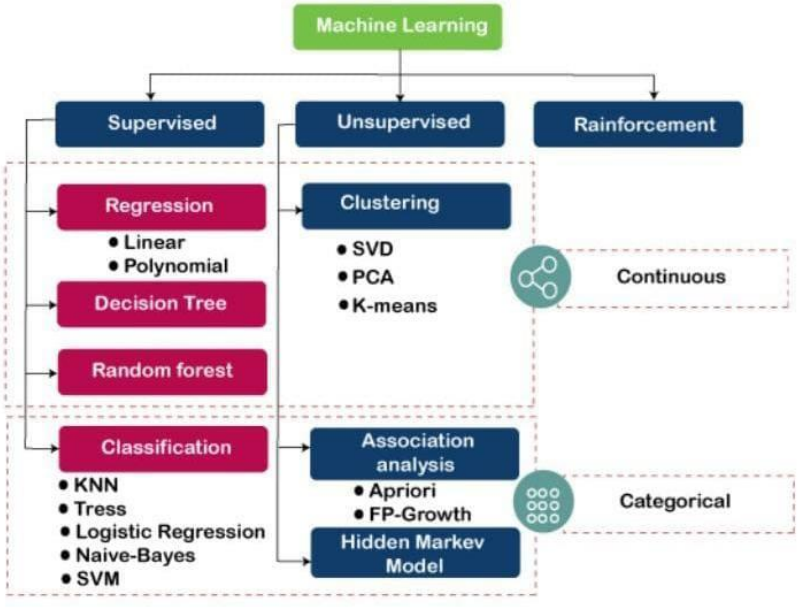
Machine Learning algorithms are the programs that can learn the hidden patterns from the data, predict the output, and improve the performance from experiences on their own. Different algorithms can be used in machine learning for different tasks, such as simple linear regression that can be used for prediction problems like stock market prediction, and the KNN algorithm can be used for classification problems.

In this topic, we will see the overview of some popular and most commonly used machine learning algorithms along with their use cases and categories.

Types of Machine Learning Algorithms:

Machine Learning Algorithm can be broadly classified into three types:

- Supervised Learning Algorithms
- Unsupervised Learning Algorithms
- Reinforcement Learning algorithm



1) Supervised Learning Algorithm:

Supervised learning is a type of Machine learning in which the machine needs external supervision to learn. The supervised learning models are trained using the labeled dataset. Once the training and processing are done, the model is tested by providing a sample test data to check whether it predicts the correct output.

The goal of supervised learning is to map input data with the output data. Supervised learning is based on supervision, and it is the same as when a student learns things in the teacher's supervision. The example of supervised learning is spam filtering.

Supervised learning can be divided further into two categories of problem:

- Classification
- Regression

Examples of some popular supervised learning algorithms are Simple Linear regression, Decision Tree, Logistic Regression, KNN algorithm, etc.

2) Unsupervised Learning Algorithm:

It is a type of machine learning in which the machine does not need any external supervision to learn from the data, hence called unsupervised learning. The unsupervised models can be trained using the unlabelled dataset that is not classified, nor categorized, and the algorithm needs to act on that data without any supervision. In unsupervised learning, the model doesn't have a predefined output, and it tries to find useful insights from the huge amount of data. These are used to solve the Association and Clustering problems.

Hence further, it can be classified into two types:

- Clustering
- Association

Examples of some Unsupervised learning algorithms are K-means Clustering, Apriori Algorithm, Eclat, etc.

3) Reinforcement Learning:

In Reinforcement learning, an agent interacts with its environment by producing actions, and learn with the help of feedback. The feedback is given to the agent in the form of rewards, such as for each good action, he gets a positive reward, and for each bad action, he gets a negative reward. There is no supervision provided to the agent. Q-Learning algorithm is used in reinforcement learning.

List of Popular Machine Learning Algorithm:

1.Linear Regression Algorithm

2.Logistic Regression Algorithm

3.Decision Tree

4.SVM

5.Naïve Bayes

6.KNN

7.K-Means Clustering

8.Random Forest

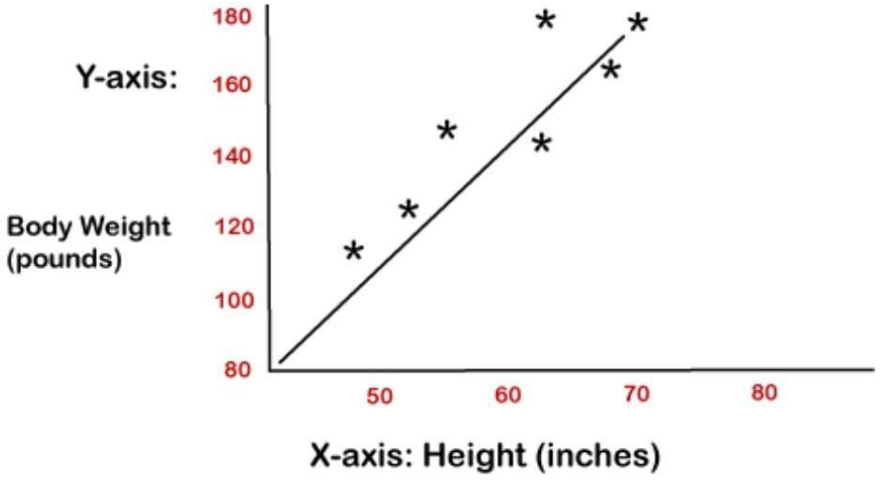
9.Apriori

10.PCA

1. Linear Regression:

Linear regression is one of the most popular and simple machine learning algorithms that is used for predictive analysis. Here, predictive analysis defines prediction of something, and linear regression makes predictions for continuous numbers such as salary, age, etc.. It shows the linear relationship between the dependent and independent variables, and shows how the dependent variable(y) changes according to the independent variable (x).

It tries to best fit a line between the dependent and independent variables, and this best fit line is knowns as the regression line.



The equation for the regression line is:

$$y = a0 + a*x + b$$

Here, y= dependent variable

x= independent variable

a0 = Intercept of line.

Linear regression is further divided into two types:

1. Simple Linear Regression:

In simple linear regression, a single independent variable is used to predict the value of the dependent variable.

2. Multiple Linear Regression:

In multiple linear regression, more than one independent variables are used to predict the value of the dependent variable.

2. Logistic Regression:

Logistic regression is the supervised learning algorithm, which is used to predict the categorical variables or discrete values. It can be used for the

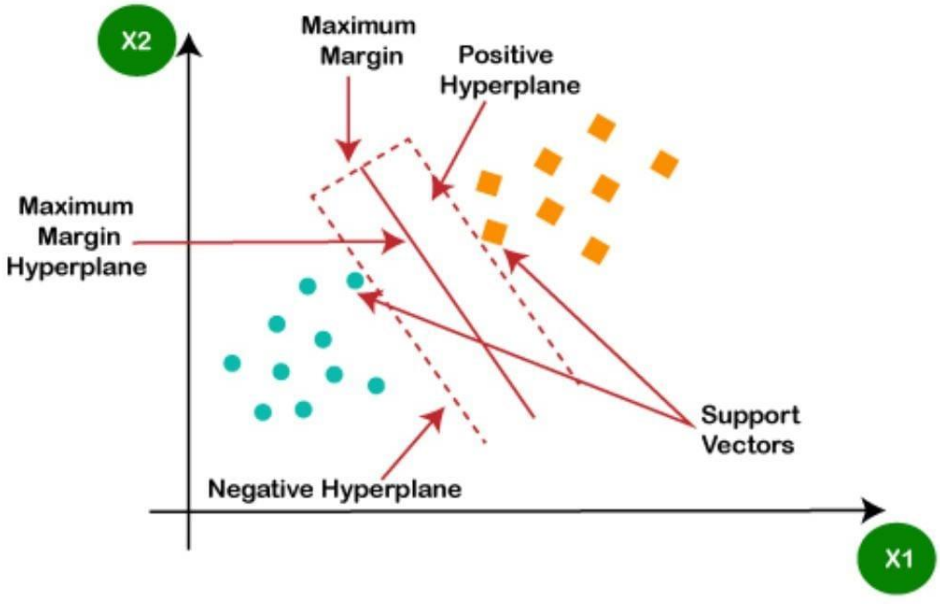
classification problems in machine learning, and the output of the logistic regression algorithm can be either Yes or NO, 0 or 1, Red or Blue, etc.

3. Decision Tree Algorithm:

A decision tree is a supervised learning algorithm that is mainly used to solve the classification problems but can also be used for solving the regression problems. It can work with both categorical variables and continuous variables. It shows a tree-like structure that includes nodes and branches, and starts with the root node that expand on further branches till the leaf node. The internal node is used to represent the features of the dataset, branches show the decision rules, and leaf nodes represent the outcome of the problem.

4. Support Vector Machine Algorithm:

A support vector machine or SVM is a supervised learning algorithm that can also be used for classification and regression problems. However, it is primarily used for classification problems. The goal of SVM is to create a hyperplane or decision boundary that can segregate datasets into different classes.



5. Naïve Bayes Algorithm:

Naïve Bayes classifier is a supervised learning algorithm, which is used to make predictions based on the probability of the object. The algorithm named

as Naïve Bayes as it is based on Bayes theorem, and follows the naïve assumption that says' variables are independent of each other

6. K-Nearest Neighbour (KNN):

K-Nearest Neighbour is a supervised learning algorithm that can be used for both classification and regression problems. This algorithm works by assuming the similarities between the new data point and available data points. Based on these similarities, the new data points are put in the most similar categories. It is also known as the lazy learner algorithm as it stores all the available datasets and classifies each new case with the help of K-neighbours. The new case is assigned to the nearest class with most similarities, and any distance function measures the distance between the data points. The distance function can be Euclidean, Minkowski, Manhattan, or Hamming distance, based on the requirement.

7. K-Means Clustering:

K-means clustering is one of the simplest unsupervised learning algorithms, which is used to solve the clustering problems. The datasets are grouped into K different clusters based on similarities and dissimilarities, it means, datasets with most of the commonalties remain in one cluster which has very less or no commonalities between other clusters. In K-means, K-refers to the number of clusters, and means refer to the averaging the dataset in order to find the centroid.

It is a centroid-based algorithm, and each cluster is associated with a centroid. This algorithm aims to reduce the distance between the data points and their centroids within a cluster.

This algorithm starts with a group of randomly selected centroids that form the clusters at starting and then perform the iterative process to optimize these centroids' positions. It can be used for spam detection and filtering, identification of fake news, etc.

8. Random Forest Algorithm:

Random forest is the supervised learning algorithm that can be used for both classification and regression problems in machine learning. It is an ensemble learning technique that provides the predictions by combining the multiple classifiers and improve the performance of the model.

It contains multiple decision trees for subsets of the given dataset, and find the average to improve the predictive accuracy of the model. A random-forest should contain 64-128 trees. The greater number of trees leads to higher accuracy of the algorithm.

9. Apriori Algorithm:

Apriori algorithm is the unsupervised learning algorithm that is used to solve the association problems. It uses frequent itemsets to generate association rules, and it is designed to work on the databases that contain transactions. With the help of these association rule, it determines how strongly or how weakly two objects are connected to each other. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset efficiently.

The algorithm process iteratively for finding the frequent itemsets from the large dataset.The apriori algorithm was given by the R. Agrawal and Srikant in the year 1994. It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions in patients.

10. Principle Component Analysis:

Principle Component Analysis (PCA) is an unsupervised learning technique, which is used for dimensionality reduction. It helps in reducing the dimensionality of the dataset that contains many features correlated with each other. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. It is one of the popular tools that is used for exploratory data analysis and predictive modeling.

PCA works by considering the variance of each attribute because the high variance shows the good split between the classes, and hence it reduces the dimensionality.Some real-world applications of PCA are image processing, movie recommendation system, optimizing the power allocation in various communication channels.

Feature selection in machine learning:

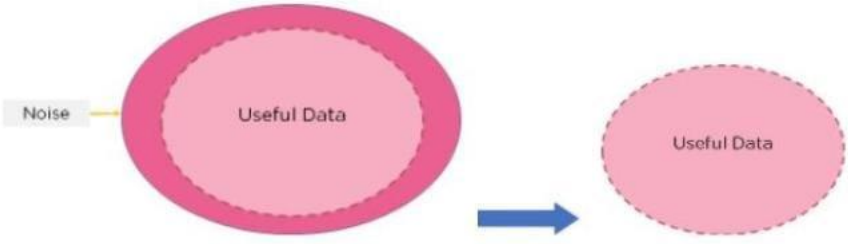
The input variables that we give to our machine learning models are called features. Each column in our dataset constitutes a feature. To train an optimal model, we need to make sure that we use only the essential features. If we have too many features, the model can capture the unimportant

patterns and learn from noise. The method of choosing the important parameters of our data is called Feature Selection.

Why feature selection?

Machine learning models follow a simple rule: whatever goes in, comes out. If we put garbage into our model, we can expect the output to be garbage too. In this case, garbage refers to noise in our data.

Feature selection is what separates good data scientists from the rest. Given the same model and computational facilities, why do some people win in competitions with faster and more accurate models? The answer is Feature Selection. Apart from choosing the right model for our data, we need to choose the right data to put in our model.



Techniques used are,

- Correlation and Mutual Information for Numerical Features
- SelectFromModel
- SelectKBest

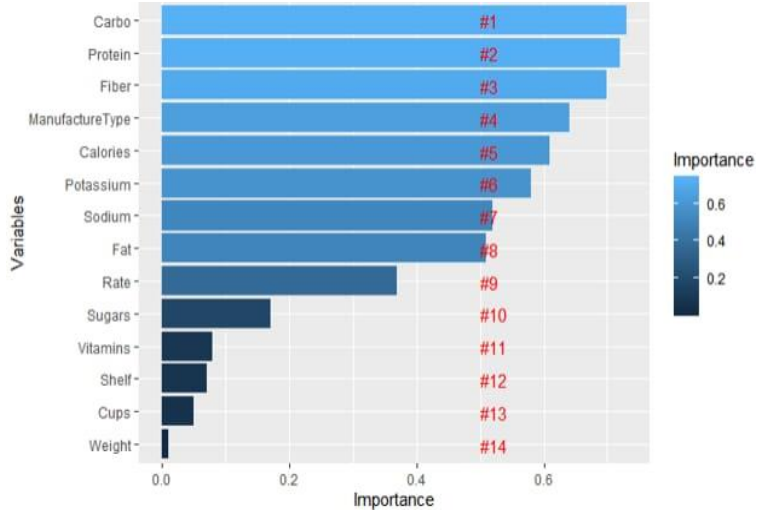
Program input:

```
feature_with_year = []

for feature in x_train.columns:
```

```
        if "Yr" in feature or "Year" in  feature:
            feature_with_year.append(feature)
    feature_with_year
```

OUTPUT :

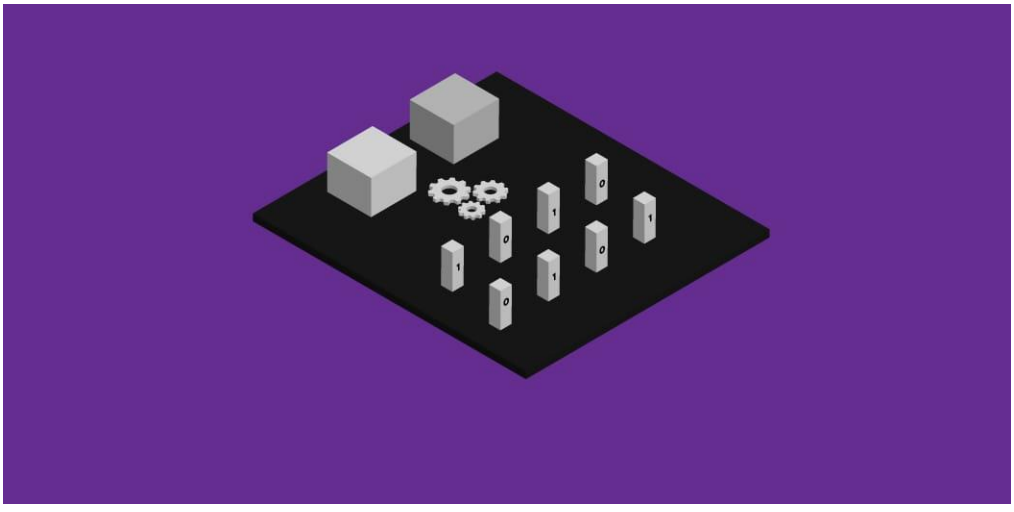
['YearBuilt', 'YearRemodAdd', 'GarageYrBlt', 'YrSold']



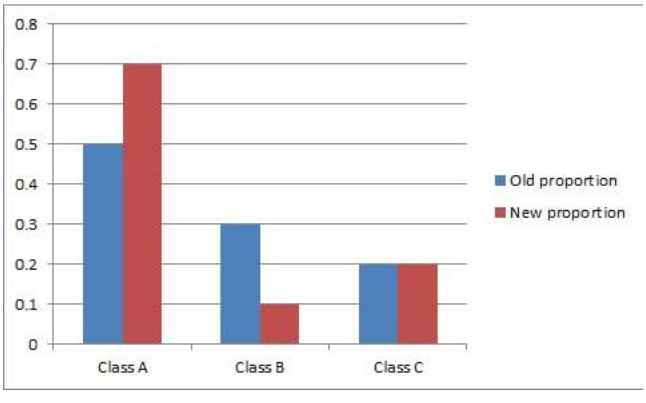
NUMERICAL FEATURES:

Numerical features are continuous values that can be measured on a scale. Examples of numerical features include age, height, weight, and income. Numerical features can be used in machine learning algorithms directly.

Categorical features are discrete values that can be grouped into categories. Examples of categorical features include gender, color, and zip code. Categorical features typically need to be converted to numerical features before they can be used in machine learning algorithms. This can be done using a variety of techniques, such as one-hot encoding, label encoding, and ordinal encoding.



The type of feature that is used in feature engineering depends on the specific machine learning algorithm that is being used. Some machine learning algorithms, such as decision trees, can handle both numerical and categorical features. Other machine learning algorithms, such as linear regression, can only handle numerical features.

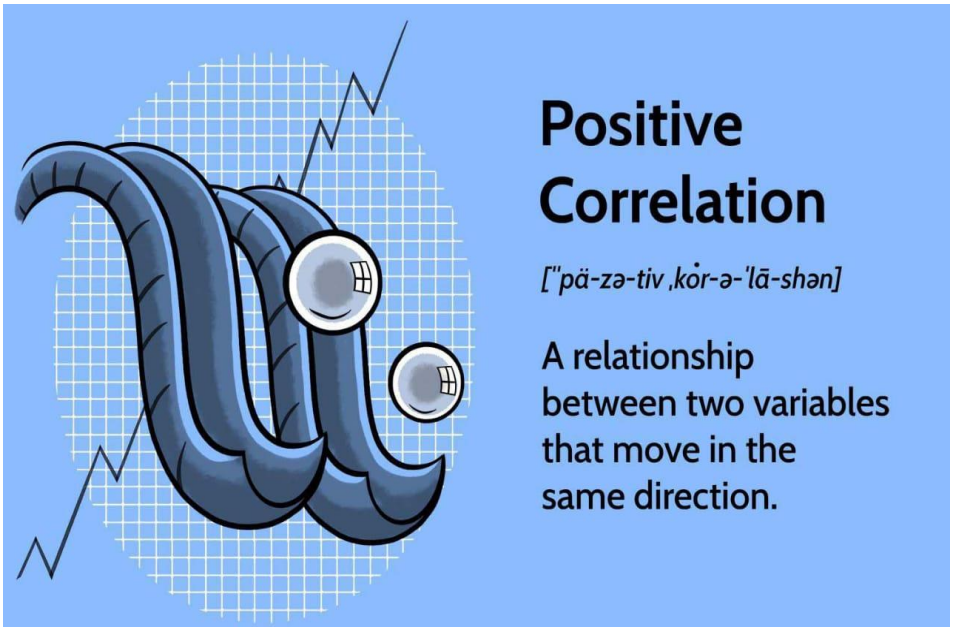


Correlation:

It is a numerical measure of the direction and magnitude of the mutual relationship between the variables(X and Y). Causation: X is the cause of change in Y i.e, the change of Y is the effect of change in X. If X and Y are correlated then X and Y may or may not have a casual relationship.

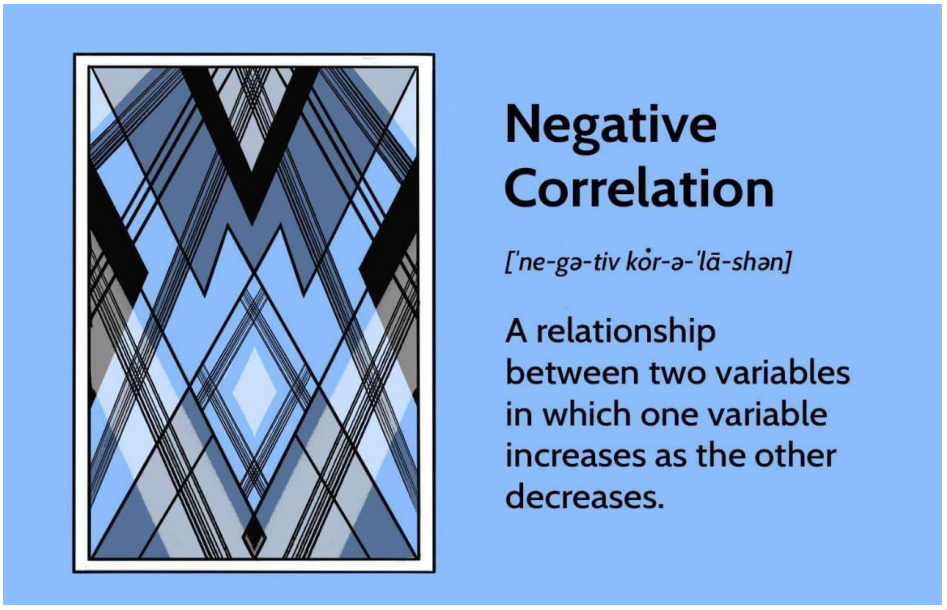
- It is a way to understand the relationship between multiple variables and attributes in your dataset.

- It helps in predicting one attribute from another.
- Can Indicate presence of causal relationship.
- Can help us identify redundant information/features.
- Positive correlation means if A increases, B also increases. Usually indicated by positive value.

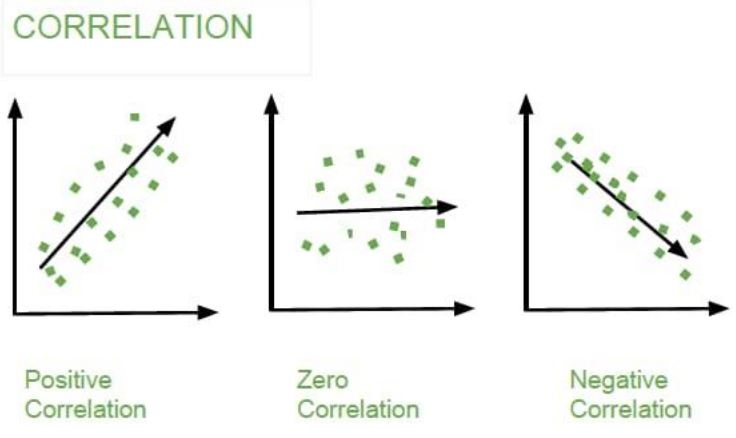


- Multicollinearity happens when one predictor variable in a multiple regression model can linearly predict others with a high degree of accuracy.
- Decision trees and boosted tree algorithms are immune to this.

- Negative correlation means if A increases, B decreases.

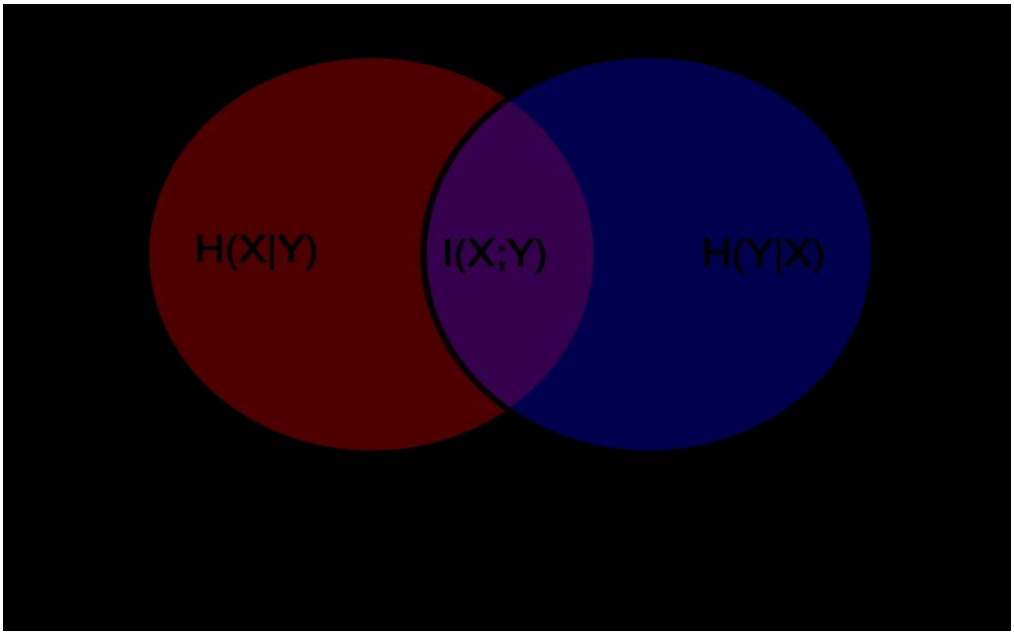


- To deal with it use PCA.
- Note that some models like Linear Regression must have correlation between dependent and independent features.

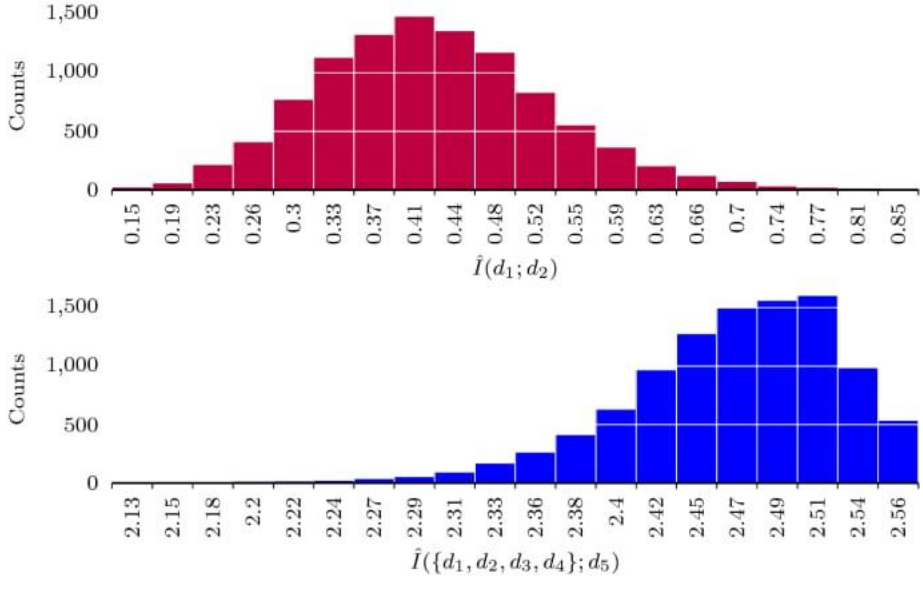


Mutual information:

Mutual information is a measure of the inherent dependence expressed in the joint distribution of and relative to the marginal distribution of and under the assumption of independence. Mutual information therefore measures dependence in the following sense: if and only if and are independent random variables.



- There is no redundancy when two variables are independent.
- As the variables get more and more dependent, redundancy increases and to quantify this, we use mutual information.
- Mutual information is the amount of extra information needed if we are representing the true joint distribution with the independent factorization.
- If the mutual information is high, the feature is a strong indicator of the class.
- Measure of mutual dependence between two variables.

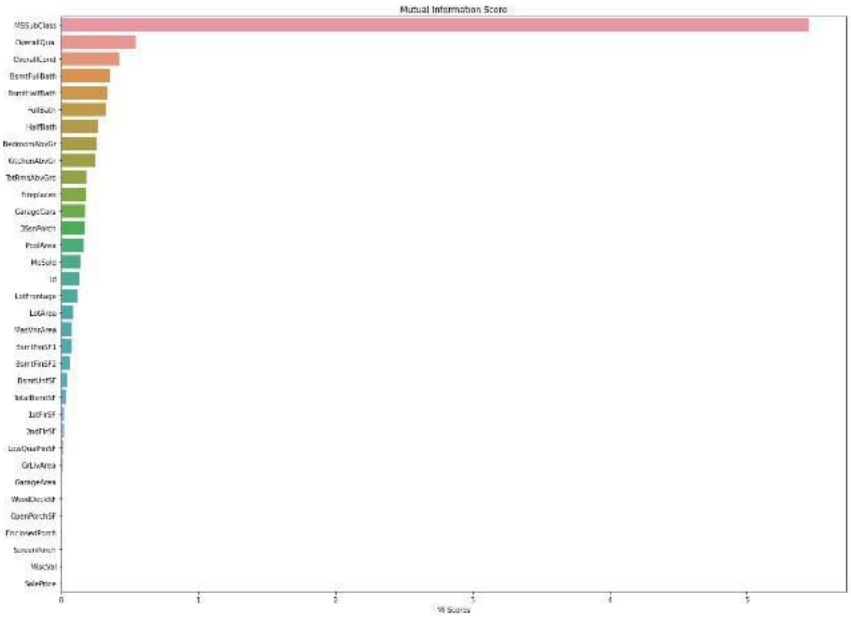


Program input:

```
plt.figure(figsize=(20, 15))
plt.title("Mutual Information Score")
sns.barplot(x=mi_scores, y=final_columns)
```

OUTPUT:

<AxesSubplot:title={'center':'Mutual Information Score'}, xlabel='MI Scores'>



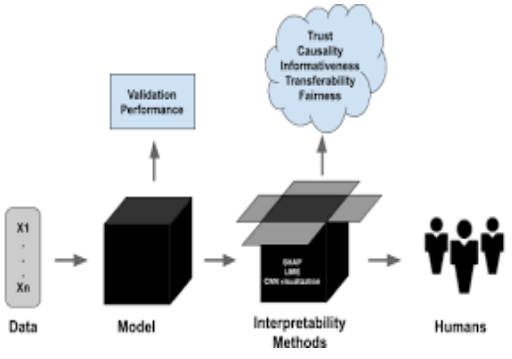
Interpretation:

Feature selection is an important problem in statistical machine learning, and is a common method for dimensionality reduction that encourages model interpretability. Classical feature selection asks for a subset of features that are most informative for the entire data set.

- From the above figure, we can infer couple of points.

1. Mutual Information between target variable and itself is 0.

2. MSSubClass shares the most information about SalePrice and thus must be selected.
3. Even though ID is unique, it does contibute more than some other features which shows that sometimes MI scores can be misleading.



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.

Lasso Regression is a model that uses l1 norm to penalizes the terms that have higher weights.

It can be used for Feature selection because it shrinks the coefficients of useless features to 0.

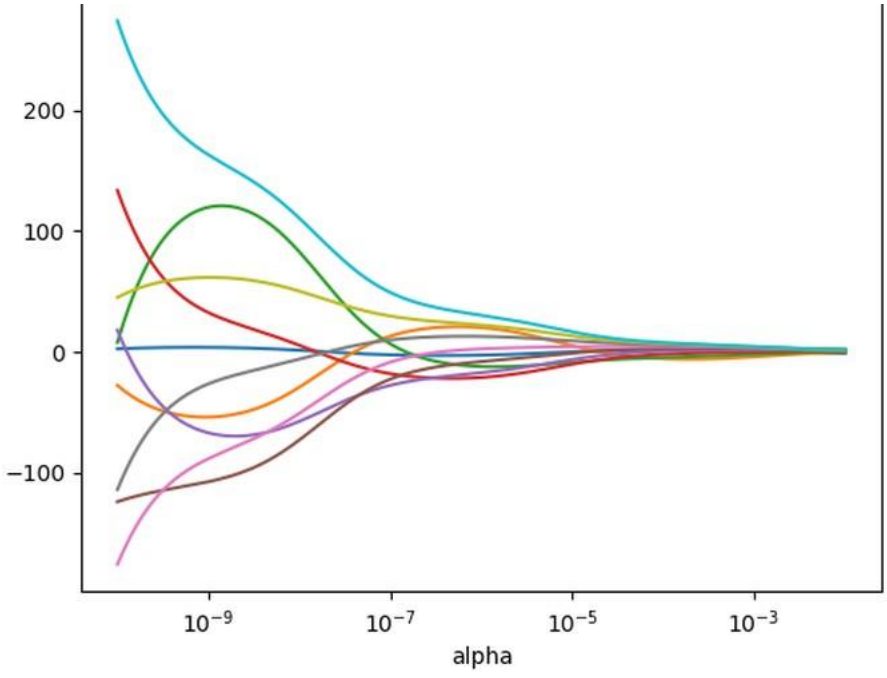
Program input:

```
# Select from model method

from sklearn.linear_model import Lasso

from sklearn.feature_selection import SelectFromModel, SelectKBest
```

```
# Drop SalePrice and ID Column
X_train = X_train.drop(["Id", "SalePrice"], axis=1, errors="ignore")
# Choose alpha value to be low because it will allow less number of
Features to shrink to 0.
alpha = 0.005
SEED_VALUE = 42
selected_features_support = SelectFromModel(Lasso(alpha=alpha,
random_state=SEED_VALUE))
selected_features_support.fit(X_train, y_train)
```

OUTPUT:

```
SelectFromModel(estimator=Lasso(alpha=0.005, random_state=42))
```



SELECTKBEST:

- Uses univariate statistical test.

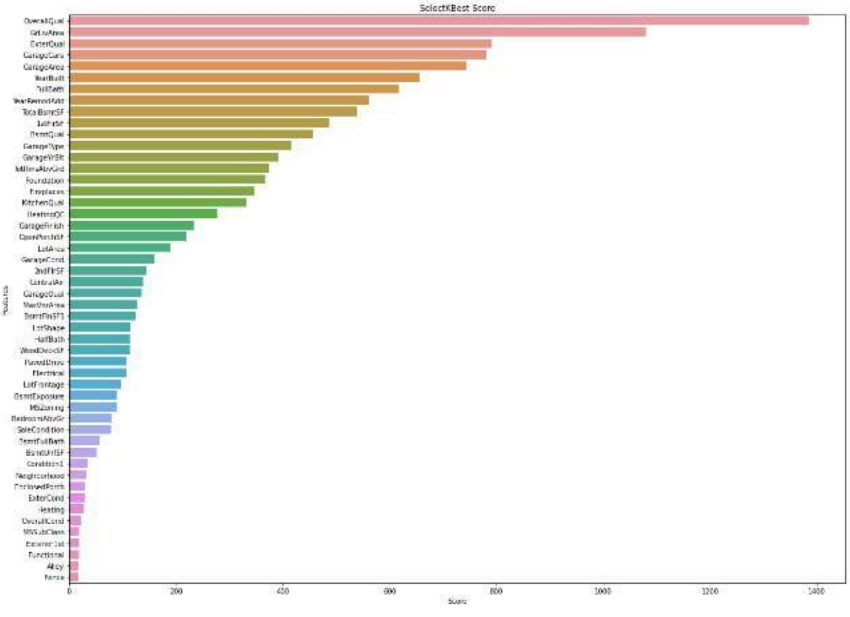
- Selects those features that have the strongest relationshup with the output variable.
- For this, we are using f_regression which is a metric for calculating f-statistics for regression problem.

Program input:

```
# Plotting Bar Chart

plt.figure(figsize=(20, 15))

plt.title("SelectKBest Score")

sns.barplot(x=k_best_scores["Score"], y=k_best_scores['Features'])
```

OUTPUT:

<AxesSubplot:title={'center':'SelectKBest Score'}, xlabel='Score', ylabel='Features'>

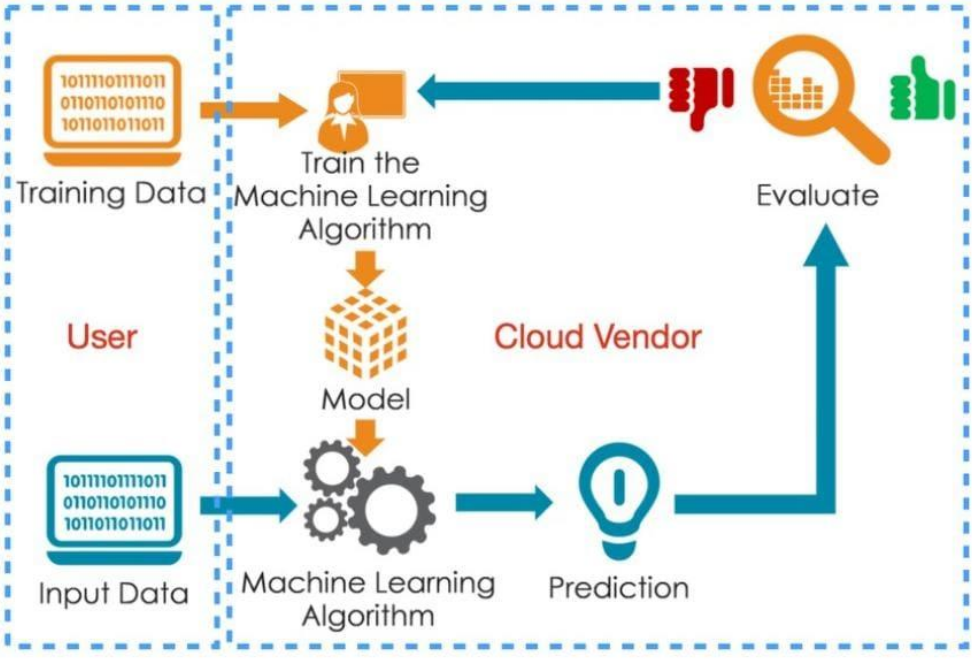


What Is Model Training In Machine Learning?

A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have

an influence on the output. The training model is used to run the input data through the algorithm to correlate the processed output against the sample output. The result from this correlation is used to modify the model.This iterative process is called “model fitting”. The accuracy of the training dataset or the validation dataset is critical for the precision of the model.

Model training in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved. There are several types of machine learning models, of which the most common ones are supervised and unsupervised learning.



Creating A Model In Machine Learning:

There are 7 primary steps involved in creating a machine learning model. Here is a brief summarized overview of each of these steps:

1.Defining The Problem:

Defining the problem statement is the first step towards identifying what an ML model should achieve. This step also enables recognizing the appropriate inputs and their respective outputs; Questions like “what is the main objective?”

“what is the input data?” and “what is the model trying to predict?” must be answered at this stage.

2.Data Collection:

After defining the problem statement, it is necessary to investigate and gather data that can be used to feed the machine. This is an important stage in the process of creating an ML model because the quantity and quality of the data used will decide how effective the model is going to be. Data can be gathered from pre-existing databases or can be built from the scratch.

3.Preparing The Data:

The data preparation stage is when data is profiled, formatted and structured as needed to make it ready for training the model. This is the stage where the appropriate characteristics and attributes of data are selected. This stage is likely to have a direct impact on the execution time and results. This is also at the stage where data is categorized into two groups – one for training the ML model and the other for evaluating the model. Pre-processing of data by normalizing, eliminating duplicates and making error corrections is also carried out at this stage.

4.Assigning Appropriate Model / Protocols:

Picking and assigning a model or protocol has to be done according to the objective that the ML model aims to achieve. There are several models to pick from, like linear regression, k-means and bayesian. The choice of models largely depends on the type of data that is being used. For instance, image processing convolutional neural networks would be the ideal pick and k-means would work best for segmentation.

5.Training The Machine Model Or “The Model Training”:

This is the stage where the ML algorithm is trained by feeding datasets. This is the stage where the learning takes place. Consistent training can significantly improve the prediction rate of the ML model. The weights of the model must be initialized randomly. This way the algorithm will learn to adjust the weights accordingly.

6.Evaluating And Defining Measure Of Success:

The machine model will have to be tested against the “validation dataset”. This helps assess the accuracy of the model. Identifying the measures of

success based on what the model is intended to achieve is critical for justifying correlation.

7.Parameter Tuning:

Selecting the correct parameter that will be modified to influence the ML model is key to attaining accurate correlation. The set of parameters that are selected based on their influence on the model architecture are called hyperparameters. The process of identifying the hyperparameters by tuning the model is called parameter tuning. The parameters for correlation should be clearly defined in a manner in which the point of diminishing returns for validation is as close to 100% accuracy as possible.

Conclusion:

Machine Learning can be a Supervised or Unsupervised. If you have lesser amount of data and clearly labelled data for training, opt for Supervised Learning. Unsupervised Learning would generally give better performance and results for large data sets.