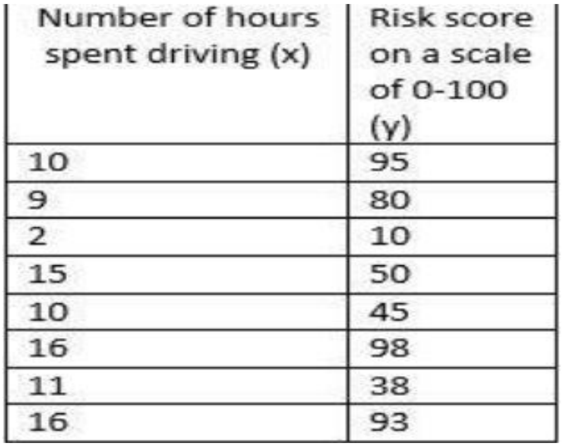
ASSIGNMENT NO. 01

Title:

Assignment based on Linear Regression.

Problem Statement:

The following table shows the results of a recently conducted study on the correlation of the number of hours spent driving with the risk of developing acute backache. Find the equation of the best fit line for this data.



Number of hours spent driving (x)	Risk score on a scale of 0-100 (y)
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

Answer-

y = 4.59x + 12.58

Hints: For each x calculate the value of y using the given equations. Then calculate error for each equation. Equation with lowest error is the desired answer. For error calculation

Given (x1,y1),(x2,y2),…,(xn,yn), best fitting data to y = f(x) by least squares requires minimization of

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2$$

Outcomes:

After completion of this assignment students are able to understand the How to find the correlation between to Two variable, How to Calculate Accuracy of the Linear Model and how to plot graph using **matplotlib.**

Theory:

Linear Regression

Regression analysis is used in stats to find trends in data. For example, you might guess that there's a connection between how much you eat and how much you weight; regression analysis can help you quantify that.

In a cause and effect relationship, the **independent variable** is the cause, and the **dependent variable** is the effect. **Least squares linear regression** is a method for predicting the value of a dependent variable *Y*, based on the value of an independent variable *X*.

Prerequisites for Regression:

Simple linear regression is appropriate when the following conditions are satisfied.

- The dependent variable *Y* has a linear relationship to the independent variable *X*. To check this, make sure that the XY scatterplot is linear and that the residual plot shows a random pattern. For each value of X, the probability distribution of Y has the same standard deviation σ.
- When this condition is satisfied, the variability of the residuals will be relatively constant across all values of X, which is easily checked in a residual plot.
- For any given value of X,
    - The Y values are independent, as indicated by a random pattern on the residual plot.
    - The Y values are roughly normally distributed (i.e., symmetric and unimodal). A little skewness is ok if the sample size is large. A histogram or a dotplot will show the shape of the distribution.

The Least Squares Regression Line:

Linear regression finds the straight line, called the **least squares regression line** or LSRL, that best represents observations in a bivariate data set. Suppose *Y* is a dependent variable, and *X* is an independent variable. The population regression line is:

$$\mathbf{Y = B_0 + B_1X}$$

Where B_0 is a constant, B_1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable.

Given a random sample of observations, the population regression line is estimated by:

$$\mathbf{\hat{y} = b_0 + b_1x}$$

Where b_0 is a constant, b_1 is the regression coefficient, x is the value of the independent variable, and ŷ is the *predicted* value of the dependent variable.

How to Define a Regression Line:

Normally, you will use a computational tool - a software package (e.g., Excel) or a graphing calculator - to find b_0 and b_1. You enter the *X* and *Y* values into your program or calculator, and the tool solves for each parameter. In the unlikely event that you find yourself on a desert island without a computer or a graphing calculator, you can solve for b_0 and b_1 "by hand". Here are the equations.

$$b_1 = \Sigma\ [\ (x_i - x)(y_i - y)\ ]\ /\ \Sigma\ [\ (x_i - x)^2]$$

$$b_1 = r * (s_y / s_x)$$

$$b_0 = y - b_1 * x$$

where b_0 is the constant in the regression equation, b_1 is the regression coefficient, r is the correlation between x and y, x_i is the *X* value of observation *i*, y_i is the *Y* value of observation *i*, x is the mean of *X*, y is the mean of *Y*, s_x is the standard deviation of *X*, and s_y is the standard deviation of *Y*.

Coefficient of determination. The coefficient of determination (R^2) for a linear regression model with one independent variable is:

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] \\ / (\sigma_x * \sigma_y) \}^2$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, x is the mean x value, y_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

If you know the linear correlation (r) between two variables, then the coefficient of determination (R^2) is easily computed using the following formula: $R^2 = r^2$.

Standard Error

The **standard error** about the regression line (often denoted by SE) is a measure of the average amount that the regression equation over- or under-predicts. The higher the coefficient of determination, the lower the standard error; and the more accurate predictions are likely to be.

Example:

Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

- What linear regression equation best predicts statistics performance, based on math aptitude scores?
- If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?
- How well does the regression equation fit the data?

How to Find the Regression Equation?

In the table below, the x_i column shows scores on the aptitude test. Similarly, the y_i column shows statistics grades. The last two columns show deviations scores - the difference between the student's score and the average score on each test. The last two rows show sums and mean scores that we will use to conduct the regression analysis.

Student	x_i	y_i	(x_i-x)	(y_i-y)
1	95	85	17	8
2	85	95	7	18
3	80	70	2	-7
4	70	65	-8	-12
5	60	70	-18	-7
Sum	390	385		
Mean	78	77		

And for each student, we also need to compute the squares of the deviation scores (the last two columns in the table below).

Student	x_i	y_i	$(x_i-x)^2$	$(y_i-y)^2$
1	95	85	289	64
2	85	95	49	324
3	80	70	4	49
4	70	65	64	144
5	60	70	324	49
Sum	390	385	730	630
Mean	78	77		

And finally, for each student, we need to compute the product of the deviation scores.

Student	x_i	y_i	$(x_i-x)(y_i-y)$
1	95	85	136
2	85	95	126
3	80	70	-14
4	70	65	96

5	60	70	126
Sum	390	385	470
Mean	78	77	

The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1x$. To conduct a regression analysis, we need to solve for b_0 and b_1. Computations are shown below. Notice that all of our inputs for the regression analysis come from the above three tables.

First, we solve for the regression coefficient (b_1):

$$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$$

$$b_1 = 470/730$$

$$b_1 = 0.644$$

Once we know the value of the regression coefficient (b_1), we can solve for the regression slope (b_0):

$$b_0 = y - b_1 * x$$

$$b_0 = 77 - (0.644)(78)$$

$$b_0 = 26.768$$

Therefore, the regression equation is: $\hat{y} = 26.768 + 0.644x$

How to Use the Regression Equation:

Once you have the regression equation, using it is a snap. Choose a value for the independent variable (*x*), perform the computation, and you have an estimated value ($\hat{y}$) for the dependent variable. In our example, the independent variable is the student's score on the aptitude test. The dependent variable is the student's statistics grade. If a student made an 80 on the aptitude test, the estimated statistics grade ($\hat{y}$) would be:

$$\hat{y} = b_0 + b_1x$$

$$\hat{y} = 26.768 + 0.644x = 26.768 + 0.644 * 80$$

$$\hat{y} = 26.768 + 51.52 = 78.288$$

Algorithm:

1. Import the Required Packages
2. Read Given Dataset
3. Import the Linear Regression and Create object of it
4. Find the Accuracy of Model using Score Function
5. Predict the value using Regressor Object
6. Take input from user.
7. Calculate the value of y
8. Draw Scatter Plot

Conclusion:

Thus we learn that to how to find the trend of data using X as Independent Variable and Y is and Dependent Variable by using Linear Regression.

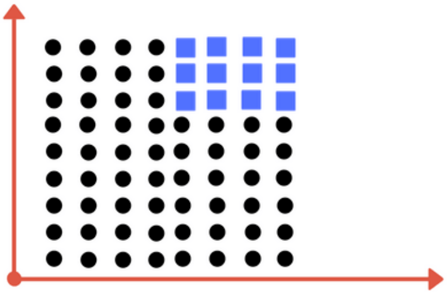
ASSIGNMENT NO. 02

Title:
Assignment based on Decision Tree Classifier

Problem Statement:
A dataset collected in a cosmetics shop showing details of customers and whether or not they responded to a special offer to buy a new lip-stick is shown in table below. Use this dataset to build a decision tree, with Buys as the target variable, to help in buying lip-sticks in the future. Find the root node of decision tree. According to the decision tree you have made from previous training data set, what is the decision for the test data: [Age < 21, Income = Low, Gender = Female, Marital Status = Married]?

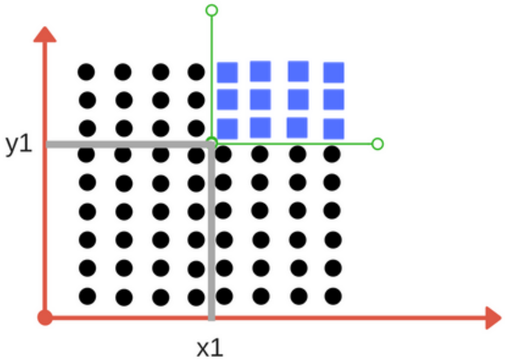
Prerequisite:
Basic of Python, Data Mining Algorithm, Concept of Decision Tree Classifier

Theory:
Suppose we have following plot for two classes represented by black circle and blue squares. Is it possible to draw a single separation line? Perhaps no.



Can you draw single division line for these classes?

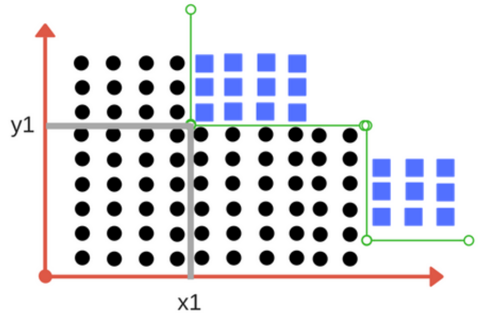
We will need more than one line, to divide into classes. Something similar to following image:



We need two lines one for threshold of x and threshold for y.

We need two lines here one separating according to threshold value of ***x*** and other for threshold value of ***y.***

Decision Tree Classifier, repetitively divides the working area (plot) into sub part by identifying lines. (Repetitively because there may be two distant regions of same class divided by other as shown in image below).



So when does it terminate?

1. Either it has divided into classes that are pure (only containing members of single class)
2. Some criteria of classifier attributes are met.

1. Impurity

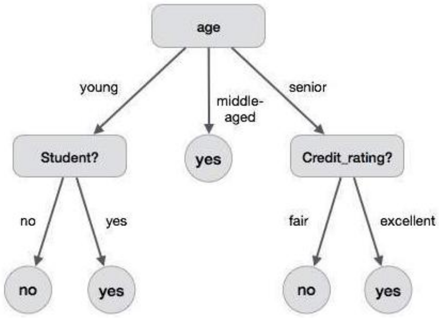
In above division, we had clear separation of classes. But what if we had following case?

Impurity is when we have traces of one class division into other. This can arise due to following reason

1. We run out of available features to divide the class upon.

We tolerate some percentage of impurity (we stop further division) for faster performance. (There is always tradeoff between accuracy and performance).

For example in second case we may stop our division when we have x number of fewer number of elements left. This is also known as ***gini impurity***.



Division based on some features.

2. Entropy

Entropy is degree of randomness of elements or in other words it is measure of impurity. Mathematically, it can be calculated with the help of probability of the items as:

$$H = -\sum p(x) \log p(x)$$

p(x) is probability of item x.

It is negative summation of probability times the log of probability of item x.

For example,

if we have items as number of dice face occurrence in a throw event as 1123,

the entropy is

p(1) = 0.5

p(2) = 0.25

p(3) = 0.25

entropy *= - (0.5 * log(0.5)) - (0.25 * log(0.25)) -(0.25 * log(0.25)*

= ***0.45***

3. Information Gain

Suppose we have multiple features to divide the current working set. What feature should we select for division? Perhaps one that gives us less impurity.

Suppose we divide the classes into multiple branches as follows, the information gain at any node is defined as,

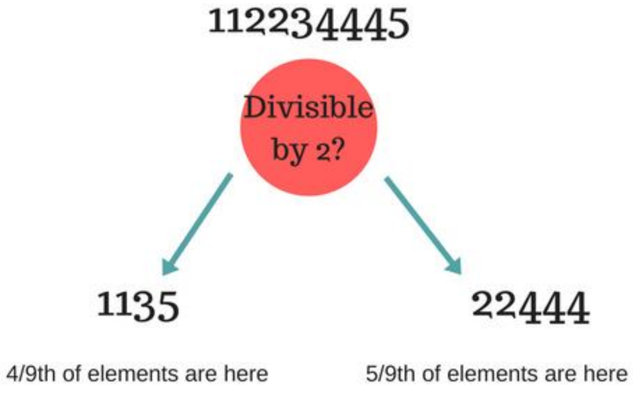
Information Gain (n) = Entropy(x) — ([weighted average] * entropy (children for feature))

This needs a bit explanation!

Suppose we have following class to work with initially

112234445

Suppose we divide them based on property: divisible by 2



Entropy at root level: 0.66

Entropy of left child: 0.45, weighted value = (4/9) * 0.45 = 0.2

Entropy of right child: 0.29, weighted value = (5/9) * 0.29 = 0.16

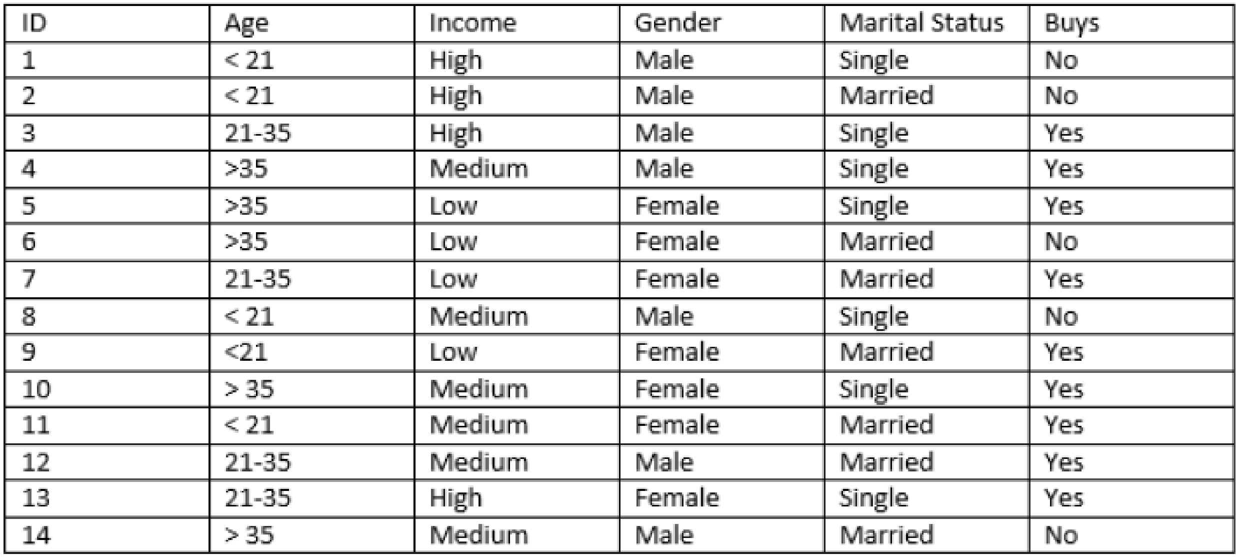
Information Gain = 0.66 - [0.2 + 0.16] = ***0.3***

Check what information gain we get if we take decision as **prime number instead of divide by 2. Which one is better for this case?**

Decision tree at every stage selects the one that gives best information gain.

When information gain is 0 means the feature does not divide the working set at all.

Given Data set in Our Definition



ID	Age	Income	Gender	Marital Status	Buys
1	< 21	High	Male	Single	No
2	< 21	High	Male	Married	No
3	21-35	High	Male	Single	Yes
4	>35	Medium	Male	Single	Yes
5	>35	Low	Female	Single	Yes
6	>35	Low	Female	Married	No
7	21-35	Low	Female	Married	Yes
8	< 21	Medium	Male	Single	No
9	<21	Low	Female	Married	Yes
10	> 35	Medium	Female	Single	Yes
11	< 21	Medium	Female	Married	Yes
12	21-35	Medium	Male	Married	Yes
13	21-35	High	Female	Single	Yes
14	> 35	Medium	Male	Married	No

What is the decision for the test data: [Age < 21, Income = Low, Gender = Female, Marital Status = Married]?

Answer is whether Yes or No??

1. Which of the attributes would be the root node? [Hints: construct the decision tree to answer these questions]

A. Age

B. Income

C. Gender

D. Marital Status

2. What is the decision for the test data [Age < 21, Income = Low, Gender = Female, Marital Status = Married]? [Hints: construct the decision tree to answer these questions]

A. Yes

B. No

Algorithm

1. Import the Required Packages
2. Read Given Dataset
3. Perform the label Encoding Mean Convert String value into Numerical values
4. Import and Apply Decision Tree Classifier
5. Predict value for the given Expression like [Age < 21, Income = Low, Gender = Female, Marital Status = Married]? In encoding Values [1,1,0,0]
6. Import the packages for Create Decision Tree.
7. Check the Decision Tree Created based on Expression.

Conclusion:

Thus, we have studied how to create Decision Tree based on given decision, Find the Root Node of the tree using Decision tree Classifier successfully.

ASSIGNMENT NO. 03

Title:
Assignment based on k-NN Classification

Problem Statement:
In the following diagram let blue circles indicate positive examples and orange squares indicate negative examples. We want to use k-NN algorithm for classifying the points. If k=3, find-the class of the point (6,6). Extend the same example for Distance-Weighted k-NN and Locally weighted Averaging.

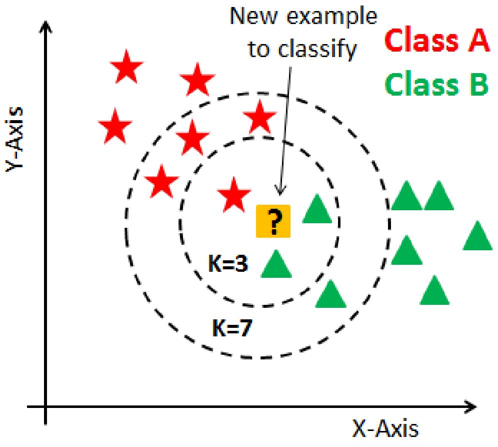
Prerequisite:
Basic of Python, Data Mining Algorithm, Concept of KNN Classification

Theory:

K-Nearest Neighbors (KNN) Algorithm:-

KNN is a ***non parametric lazy learning*** algorithm. That is a pretty concise statement. When you say a technique is non parametric, it means that it does not make any assumptions on the underlying data distribution. This is pretty useful, as in the real world, most of the practical data does not obey the typical theoretical assumptions made (eg Gaussian mixtures, linearly separable etc). Non parametric algorithms like KNN come to the rescue here.

KNN Algorithm is based on **feature similarity**: How closely out-of-sample features resemble our training set determines how we classify a given data point:



KNN can be used for **classification**: - the output is a class membership (predicts a class—a discrete value). An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. It can also be used for **regression**—output is the value for the object (predicts continuous values). This value is the average (or median) of the values of its k nearest neighbors.

KNN for Classification:

Let's see how to use KNN for classification. In this case, we are given some data points for training and also a new unlabelled data for testing. Our aim is to find the class label for the new point. The algorithm has different behavior based on k.

Case 1: k = 1 or Nearest Neighbor Rule

This is the simplest scenario. Let x be the point to be labeled. Find the point closest to x. Let it be y. Now nearest neighbor rule asks to assign the label of y to x. This seems too simplistic and sometimes even counter intuitive. If you feel that this procedure will result a huge error, you are right – but there is a catch. This reasoning holds only when the number of data points is not very large.

If the number of data points is very large, then there is a very high chance that label of x and y are same. An example might help – Let's say you have a (potentially) biased coin. You toss it for 1 million time and you have got head 900,000 times. Then most likely your next call will be head. We can use a similar argument here.

Let me try an informal argument here - Assume all points are in a D dimensional plane. The number of points is reasonably large. This means that the density of the plane at any point is fairly high. In other words, within any subspace there is adequate number of points. Consider a point x in the subspace which also has a lot of neighbors. Now let y be the nearest neighbor. If x and y are sufficiently close, then we can assume that probability that x and y belong to same class is fairly same – Then by decision theory, x and y have the same class.

The book **"Pattern Classification**" by Duda and Hart has an excellent discussion about this Nearest Neighbor rule. One of their striking results is to obtain a fairly tight error bound to the Nearest Neighbor rule. The bound is

$$P^* \leq P \leq P^*\left(2 - \frac{c}{c-1}P^*\right)$$

Where P^* is the Bayes error rate, c is the number of classes and P is the error rate of Nearest Neighbor. The result is indeed very striking (at least to me) because it says that if the number of points is fairly large then the error rate of Nearest Neighbor is less than twice the Bayes error rate. Pretty cool for a simple algorithm like KNN.

Case 2: k = K or k-Nearest Neighbor Rule

This is a straightforward extension of 1NN. Basically what we do is that we try to find the k nearest neighbor and do a majority voting. Typically k is odd when the number of classes is 2. Let’s say k = 5 and there are 3 instances of C1 and 2 instances of C2. In this case, KNN says that new point has to label as C1 as it forms the majority. We follow a similar argument when there are multiple classes.

One of the straight forward extensions is not to give 1 vote to all the neighbors. A very common thing to do is **weighted KNN** where each point has a weight which is typically calculated using its distance. For eg under inverse distance weighting, each point has a weight equal to the inverse of its distance to the point to be classified. This means that neighboring points have a higher vote than the farther points.

It is quite obvious that the accuracy *might* increase when you increase k but the computation cost also increases.

Applications of KNN:

KNN is a versatile algorithm and is used in a huge number of fields. Let us take a look at few uncommon and non trivial applications.

1. Nearest Neighbor based Content Retrieval

This is one the fascinating applications of KNN – Basically we can use it in Computer Vision for many cases – You can consider handwriting detection as a rudimentary nearest neighbor problem. The problem becomes more fascinating if the content is a video – given a video find the video closest to the query from the database – Although this looks abstract, it has lot of practical applications – Eg :

Consider **ASL** (American Sign Language) . Here the communication is done using hand gestures.

2. Gene Expression

This is another cool area where many a time, KNN performs better than other state of the art techniques . In fact a combination of KNN-SVM is one of the most popular techniques there. This is a huge topic on its own and hence I will refrain from talking much more about it.

3. Protein-Protein interaction and 3D structure prediction

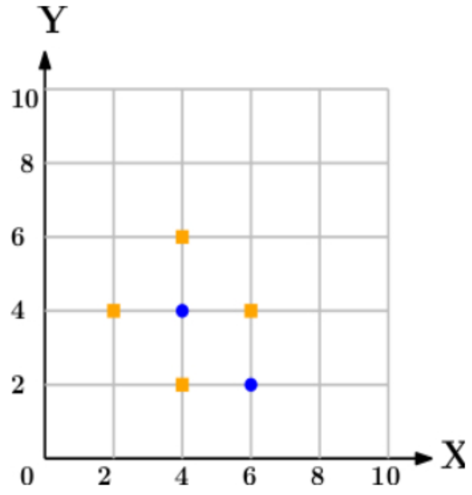
Graph based KNN is used in protein interaction prediction. Similarly KNN is used in structure prediction.

4. Credit ratings—collecting financial characteristics vs. comparing people with similar financial features to a database. By the very nature of a credit rating, people who have similar financial details would be given similar credit ratings. Therefore, they would like to be able to use this existing database to predict a new customer’s credit rating, without having to perform all the calculations.

5. Should the bank give a loan to an individual? Would an individual default on his or her loan? Is that person closer in characteristics to people who defaulted or did not default on their loans?

Given Diagram Represent Positive and Negative Point with Color

In the following diagram let blue circles indicate positive examples and orange squares indicate negative examples. We want to use k-NN algorithm for classifying the points. If k=3, find-the class of the point (6,6). Extend the same example for Distance-Weighted k-NN and Locally Weighted Averaging



Algorithm:

1. Import the Required Packages
2. Read Given Dataset
3. Import KNeighborshood Classifier and create object of it.
4. Predict the class for the point (6, 6) w.r.t to General KNN.
5. Predict the class for the point (6, 6) w.r.t to Distance Weighted KNN.

Conclusion:

Thud we have studied how KNN Classification to predict the General and Distance Weighted KNN for Given data point in term of Positive or Negative.

ASSIGNMENT NO. 04

Title:
Assignment based on k-mean Clustering

Problem Statement:

We have given a collection of 8 points. P1=[0.1,0.6] P2=[0.15,0.71] P3=[0.08,0.9] P4=[0.16,0.85] P5=[0.2,0.3] P6=[0.25,0.5] P7=[0.24,0.1] P8=[0.3,0.2]. Perform the k-mean clustering with initial centroids as m1=P1 =Cluster#1=C1 and m2=P8=cluster#2=C2. Answer the following
1] Which cluster does P6 belongs to?
2] What is the population of cluster around m2?
3] What is updated value of m1 and m2?

Prerequisite:
Basic of Python, Data Mining Algorithm, Concept of K-mean Clustering

Theory:
A Hospital Care chain wants to open a series of Emergency-Care wards within a region. We assume that the hospital knows the location of all the maximum accident-prone areas in the region. They have to decide the number of the Emergency Units to be opened and the location of these Emergency Units, so that all the accident-prone areas is covered in the vicinity of these Emergency Units.
The challenge is to decide the location of these Emergency Units so that the whole region is covered. Here is when K-means Clustering comes to rescue!
A cluster refers to a small group of objects. Clustering is grouping those objects into clusters. In order to learn clustering, it is important to understand the scenarios that lead to cluster different objects. Let us identify a few of them.

What is Clustering?
Clustering is dividing data points into homogeneous classes or clusters:
Points in the same group are as similar as possible
Points in different group are as dissimilar as possible
When a collection of objects is given, we put objects into group based on similarity.

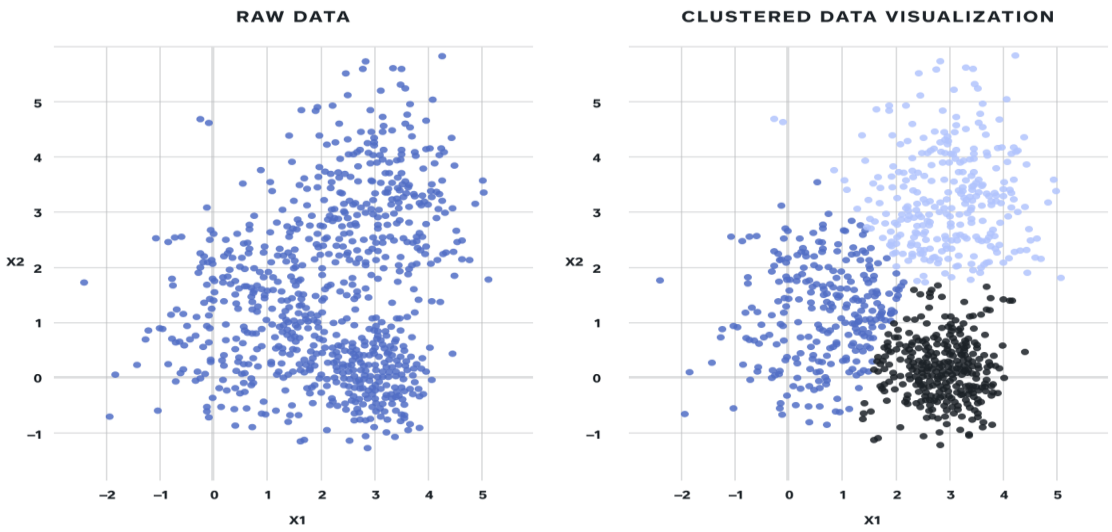
Application of Clustering:
Clustering is used in almost all the fields. You can infer some ideas from Example 1 to come up with lot of clustering applications that you would have come across.
Listed here are few more applications, which would add to what you have learnt.

- ✓ Clustering helps marketers improve their customer base and work on the target areas. It helps group people (according to different criteria’s such as willingness, purchasing power etc.) based on their similarity in many ways related to the product under consideration.
- ✓ Clustering helps in identification of groups of houses on the basis of their value, type and geographical locations.
- ✓ Clustering is used to study earth-quake. Based on the areas hit by an earthquake in a region, clustering can help analyse the next probable location where earthquake can occur.

Clustering Algorithms:

A Clustering Algorithm tries to analyse natural groups of data on the basis of some similarity. It locates the centroids of the group of data points. To carry out effective clustering, the algorithm evaluates the distance between each point from the centroids of the cluster.

The goal of clustering is to determine the intrinsic grouping in a set of unlabelled data.



What is K-means Clustering?

K-means (Macqueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining.

K-means Clustering – Example 1:

A pizza chain wants to open its delivery centres across a city. What do you think would be the possible challenges?

- They need to analyse the areas from where the pizza is being ordered frequently.
- They need to understand as to how many pizza stores has to be opened to cover delivery in the area.
- They need to figure out the locations for the pizza stores within all these areas in order to keep the distance between the store and delivery points minimum.

Resolving these challenges includes a lot of analysis and mathematics. We would now learn about how clustering can provide a meaningful and easy method of sorting out such real life challenges. Before that let's see what clustering is.

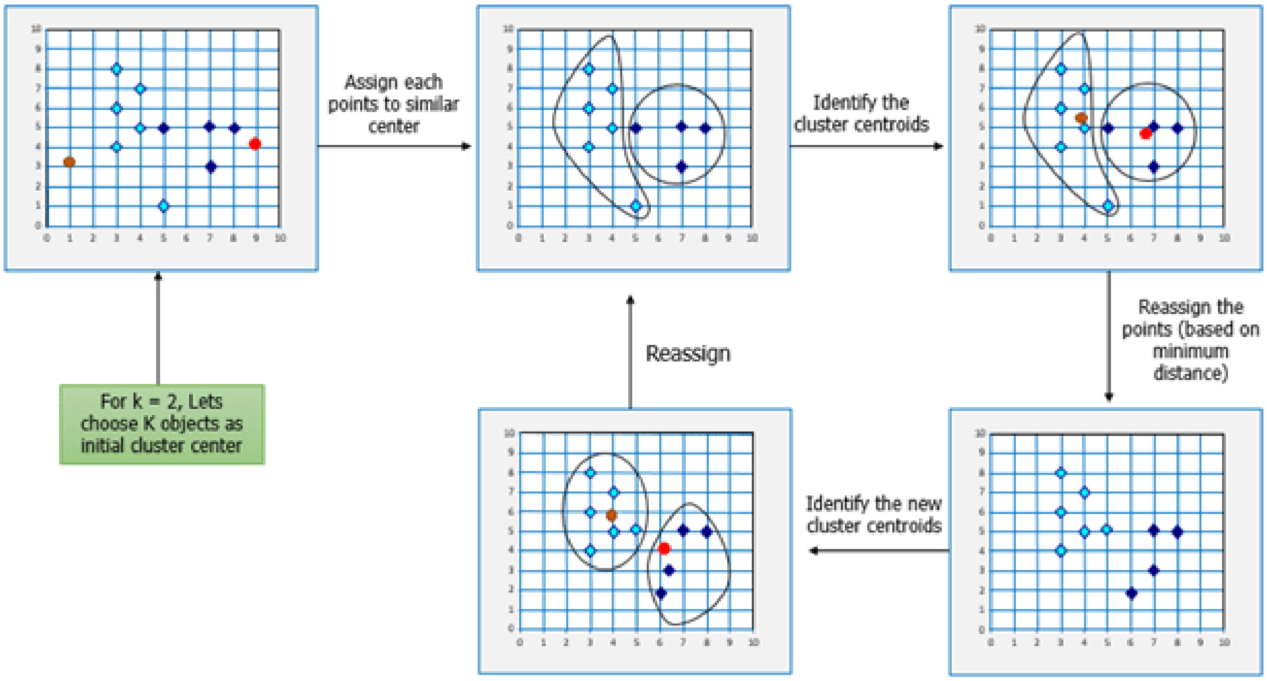
K-means Clustering Method:

If k is given, the K-means algorithm can be executed in the following steps:

- Partition of objects into k non-empty subsets
- Identifying the cluster centroids (mean point) of the current partition.
- Assigning each point to a specific cluster
- Compute the distances from each point and allot points to the cluster where the distance from the centroids is minimum.

- After re-allotting the points, find the centroids of the new cluster formed.

The step by step process of clustering:


Assign each points to similar center
Identify the cluster centroids
Reassign the points (based on minimum distance)
Identify the new cluster centroids
Reassign
For k = 2, Lets choose K objects as initial cluster center

Now, let's consider the problem in Example 1 and see how we can help the pizza chain to come up with centres based on K-means algorithm.

Similarly, for opening Hospital Care Wards:

K-means Clustering will group these locations of maximum prone areas into clusters and define a cluster center for each cluster, which will be the locations where the Emergency Units will open. These Clusters centers are the centroids of each cluster and are at a minimum distance from all the points of a particular cluster, henceforth, the Emergency Units will be at minimum distance from all the accident prone areas within a cluster.

Here is another example for you, try and come up with the solution based on your understanding of K-means clustering.

K-means Clustering – Example 2:

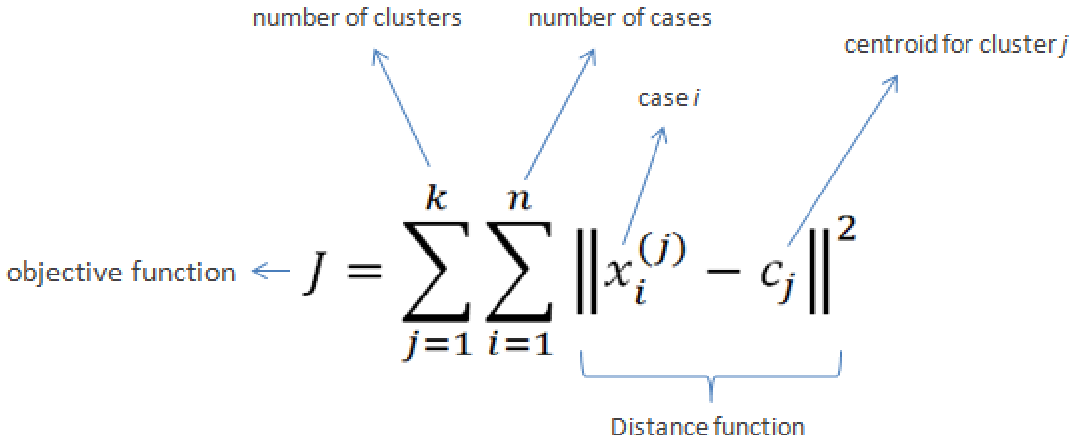
Let's consider the data on drug-related crimes in Canada. The data consists of crimes due to various drugs that include, Heroin, Cocaine to prescription drugs, especially by underage people. The crimes resulted due to these substance abuse can be brought down by starting de-addiction centres in areas most afflicted by this kind of crime. With the available data, different objectives can be set. They are:

- Classify the crimes based on the abuse substance to detect prominent cause.
- Classify the crimes based on age groups.
- Analyze the data to determine what kinds of de-addiction centre are required.
- Find out how many de-addiction centres need to be setup to reduce drug related crime rate.

The K-means algorithm can be used to determine any of the above scenarios by analyzing the available data.
Following the K-means Clustering method used in the previous example, we can start off with a given k, following by the execution of the K-means algorithm.

Mathematical Formulation for K-means Algorithm:

K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading h separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means cluster total intra-cluster variance, or, the squared error function



Algorithm:

1. Import the Required Packages
2. Create dataset using DataFrame
3. Find centroids points
4. Plot the given points
5. For i in centroids ():
6. Plot given elements with centroids elements
7. Import KMeans class and create object of it
8. Using labels find population around centroids
9. Find new centroids

Conclusion:

Thus, in this experiment we have studied and implemented Kmean Clustering Algorithm successfully.

ASSIGNMENT NO. 05

Mini-Project 1 on Genetic Algorithm:

Apply the Genetic Algorithm for optimization on a dataset obtained from UCI ML repository.

For Example: IRIS Dataset or Travelling Salesman Problem or KDD Dataset

OR

Mini-Project 2 on SVM:

Apply the Support vector machine for classification on a dataset obtained from UCI ML repository.

For Example: Fruits Classification or Soil Classification or Leaf Disease Classification.

OR

Mini-Project 3 on PCA:

Apply the Principal Component Analysis for feature reduction on any Company Stock Market Dataset.

ASSIGNMENT NO. 06

Title:

Implementation of S-DES (Data Encryption Standard)

Problem Statement:

Implementation of S-DES

Theory:

Data Encryption Standard (DES)

The Data Encryption Standard (DES) is a Symmetric-key block cipher issued by the national Institute of Standards & Technology (NIST).

DES is an implementation of a Feistel Cipher. It uses 16 round Feistel structure. The block size is 64-bit. Though, key length is 64-bit, DES has an effective key length of 56 bits, since 8 of the 64 bits of the key are not used by the encryption algorithm (function as check bits only). **General Structure of DES is depicted in the following illustration**

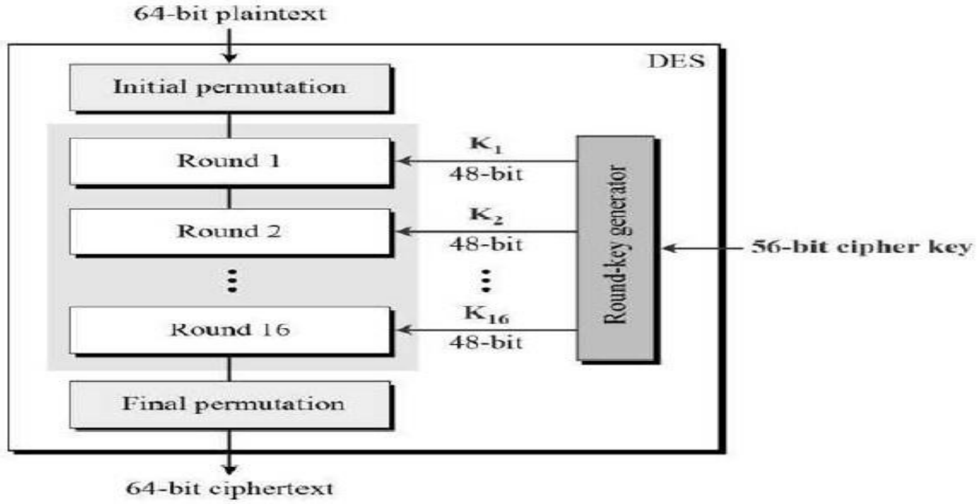


Fig. General Structure of DES

Since DES is based on the Feistel Cipher, all that is required to specify DES is −

- Round function
- Key schedule
- Any additional processing − Initial and final permutation

Initial and Final Permutation

The initial and final permutations are straight Permutation boxes (P-boxes) that are inverses of each other. They have no cryptography significance in DES. The initial and final permutations are shown as follows,

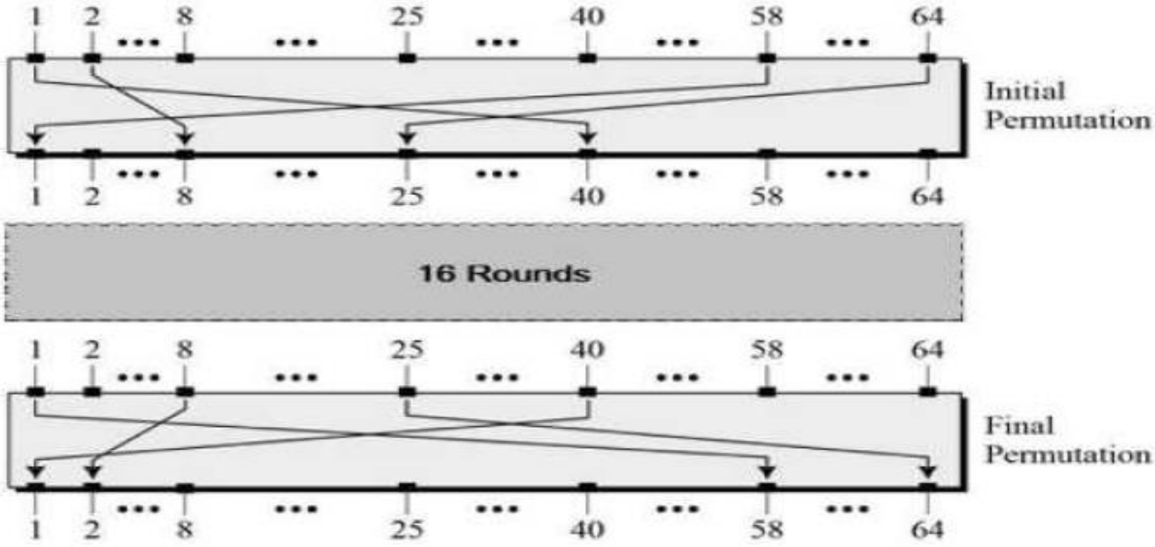
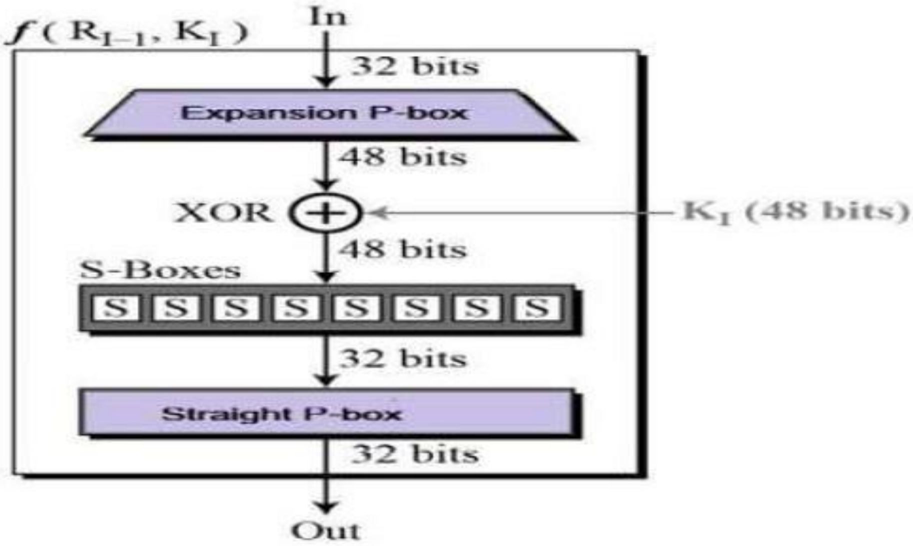


Fig. initial and final permutations

Round Function:

The heart of this cipher is the DES function, *f*. The DES function applies a 48-bit key to the right most 32 bits to produce a 32-bit output.



Expansion Permutation Box

Since right input is 32-bit and round key is a 48-bit, we first need to expand right input to 48 bits.

XOR(Whitener)

After the expansion permutation, DES does XOR operation on the expanded right section and the round key. The round key is used only in this operation.

Substitution Boxes:

The S-boxes carry out the real mixing (confusion). DES uses 8 S-boxes, each with a 6-bit input and a 4-bit output.

Straight Permutation: The 32 bit output of S-boxes is then subjected to the straight permutation with rule

Key Generation:

The round-key generator creates sixteen 48-bit keys out of a 56-bit cipher key. The process of key generation is depicted in the following illustration

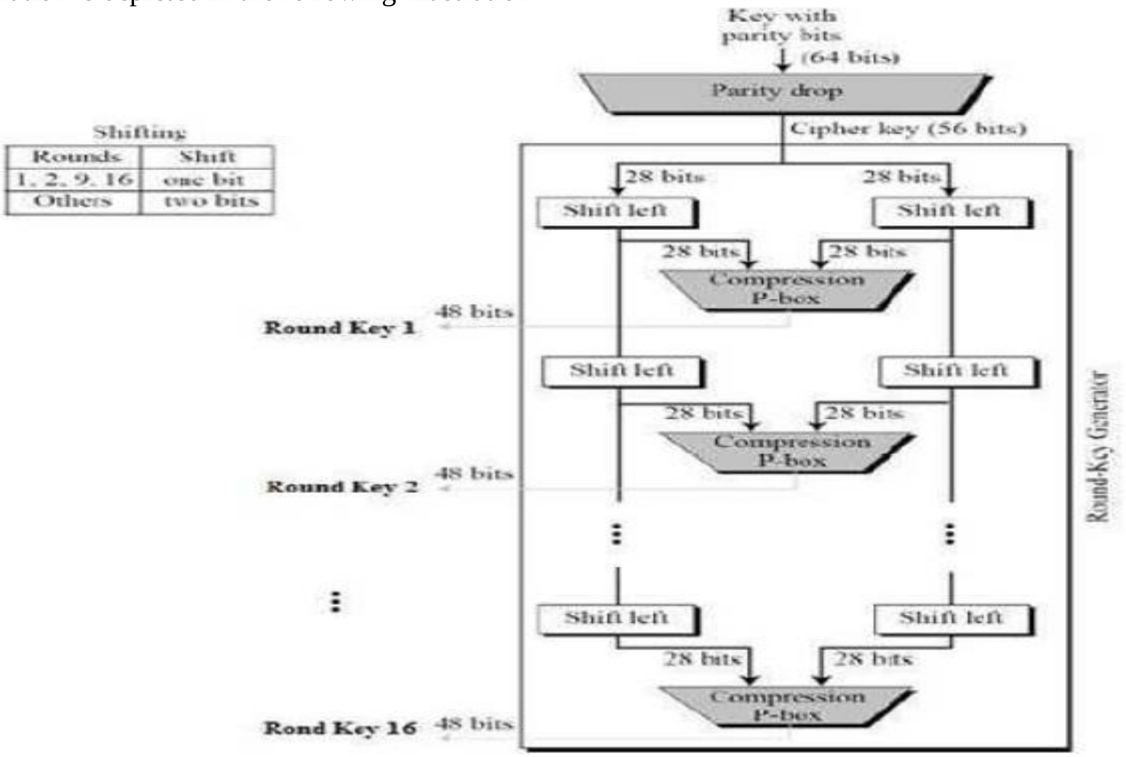


Fig. The process of key generation

DES Analysis:

The DES satisfies both the desired properties of block cipher. These two properties make cipher very strong.

- **Avalanche effect:-**A small change in plaintext results in the very great change in the cipher text.
- **Completeness: -** Each bit of cipher text depends on many bits of plaintext.

During the last few years, cryptanalysis has found some weaknesses in DES when key selected are weak keys. These keys shall be avoided.

DES has proved to be a very well designed block cipher. There have been no significant cryptanalytic attacks on DES other than exhaustive key search.

Conclusion:

Thus we have studied and implemented S-DES in detail successfully.

ASSIGNMENT NO. 07

Title:

Implementation of S-AES (Advanced Encryption Standard)

Problem Statement:

Implementation of S-AES

Theory:

The more popular and widely adopted symmetric encryption algorithm likely to be encountered Now a days is the Advanced Encryption Standard (AES). It is found at least six times faster than Triple DES. A replacement for DES was needed as its key size was too small. With increasing computing power, it was considered vulnerable against exhaustive key search attack. Triple DES was designed to overcome this drawback but it was found slow.

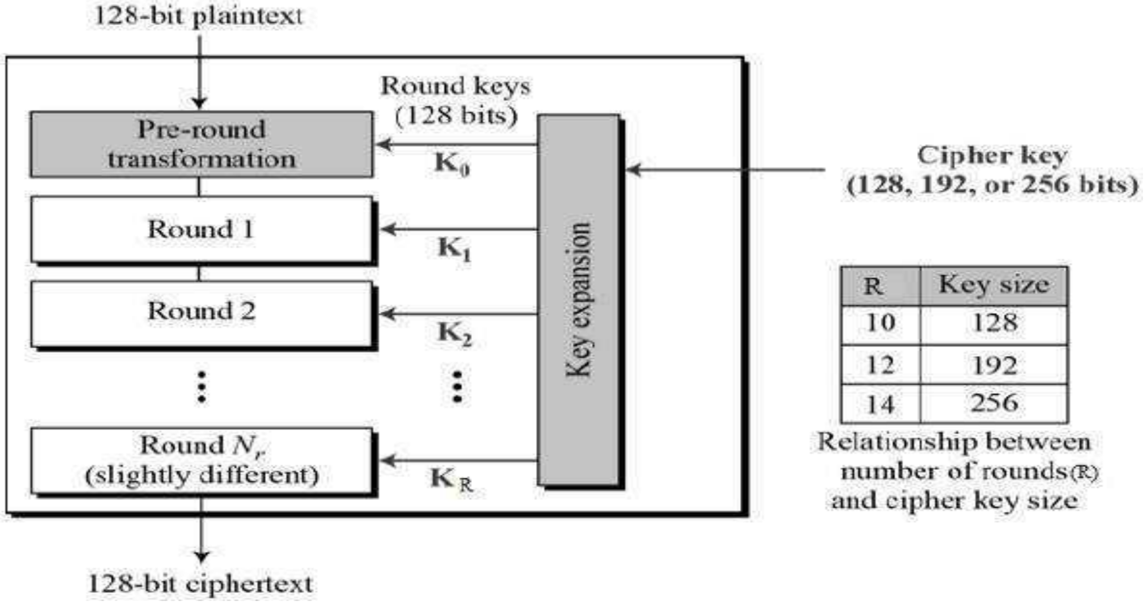
The features of AES are as follows:

- Symmetric key symmetric block cipher
- 128-bit data, 128/192/256-bit keys
- Stronger and faster than Triple-DES
- Provide full specification and design details
- Software implementable in C ,Java and Python

Operation of AES

AES is an iterative rather than Feistel cipher. It is based on ‘substitution–permutation network’. It comprises of a series of linked operations, some of which involve replacing inputs by specific Outputs (substitutions) and others involve shuffling bits around (permutations). Interestingly, AES performs all its computations on bytes rather than bits. Hence, AES treats the 128 bits of a plaintext block as 16 bytes. These 16 bytes are arranged in four columns and four rows for processing as a matrix .Unlike DES, the number of rounds in AES is variable and depends on the length of the key. AES uses 10 rounds for 128-bit keys, 12 rounds for 192-bit keys and 14 rounds for 256-bit keys. Each of these rounds uses a different 128-bit round key, which is calculated from the original AES key.

The schematic of AES structure is given in the following illustration



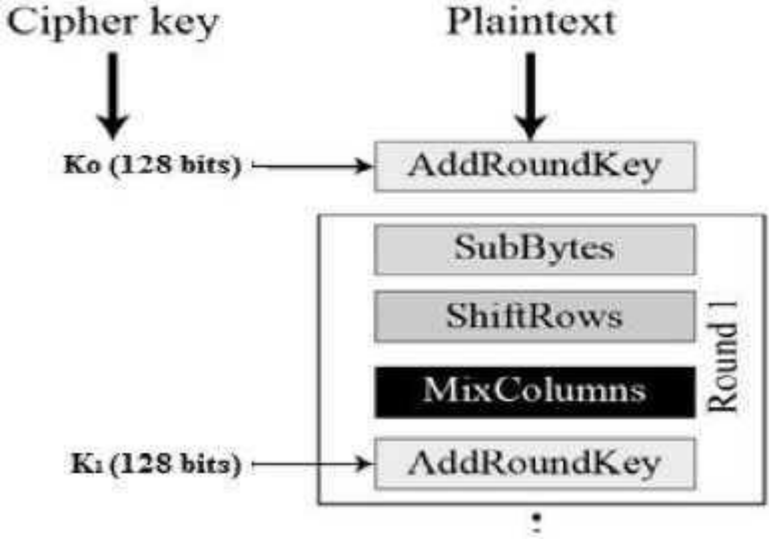
R	Key size
10	128
12	192
14	256

Relationship between number of rounds(R) and cipher key size

Figure 6.1 AES structure

Encryption Process:

Here, we restrict to description of a typical round of AES encryption. Each round comprise of four Sub-processes. The first round process is depicted below,



Conclusion:

Thus in this experiment we have studied s-aes and implemented successfully.

ASSIGNMENT NO. 08

Title:

Implementation of Diffie-Hellman key Exchange (DH)

Problem Statement:

Implementation of Diffie-Hellman key Exchange (DH)

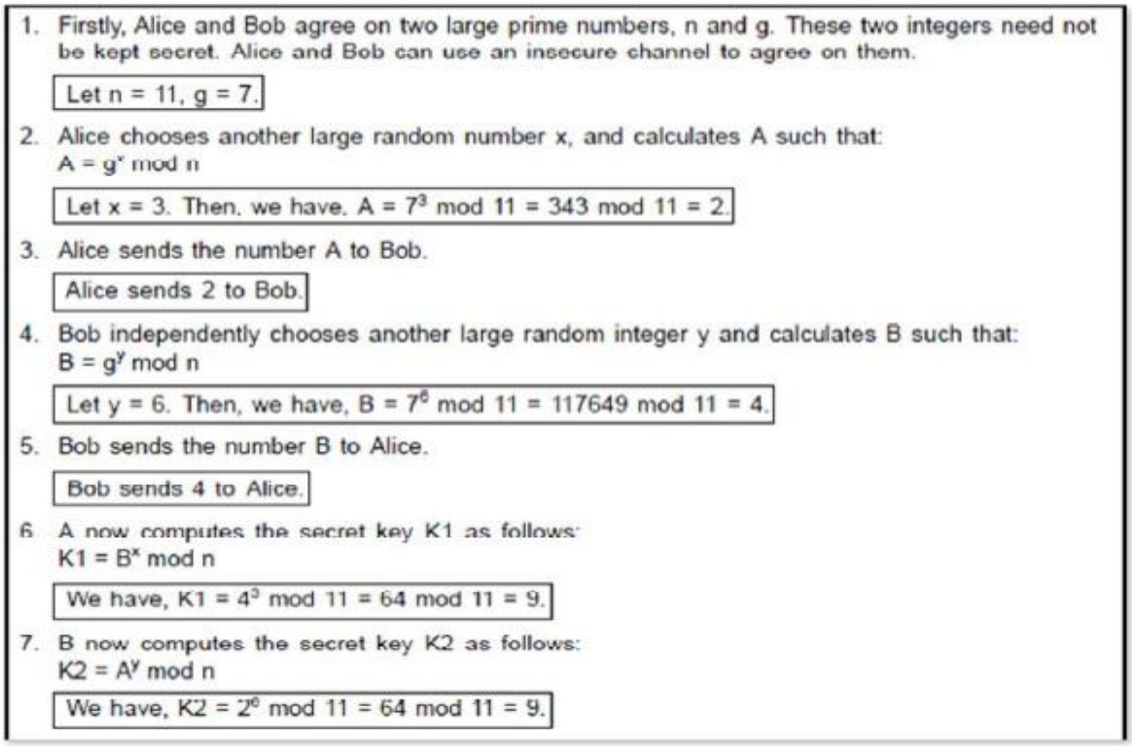
Diffie-Hellman key Exchange (DH)

In the mid- 1970's, Whitefield Diffie, a student at the Stanford University met with Martin Hellman, his professor &the two began to think about it. After some research & complicated mathematical analysis, they came up with the idea of AKC. Many experts believe that this development is the first & perhaps the only truly revolutionary concept in the history of cryptography.

Silent Features of Diffie-Hellman key Exchange (DH)

1. Developed to address shortfalls of *key distribution* in symmetric key distribution.
2. A *key exchange algorithm*, not an encryption algorithm
3. Allows two users to share a *secret key* securely over a public network
4. Once the key has been shared Then both parties can use it to encrypt and decrypt messages using symmetric cryptography
5. Algorithm is based on "difficulty of calculating discrete logarithms in a finite field"
6. These keys are mathematically related to each other.
7. ''Using the public key of users, the session key is generated without transmitting the private key of the users.''

Diffie-Hellman Key Exchange/Agreement Algorithm with Example



1. Firstly, Alice and Bob agree on two large prime numbers, n and g. These two integers need not be kept secret. Alice and Bob can use an insecure channel to agree on them.

   Let n = 11, g = 7.

2. Alice chooses another large random number x, and calculates A such that:
   $A = g^x \bmod n$

   Let x = 3. Then, we have, $A = 7^3 \bmod 11 = 343 \bmod 11 = 2$.

3. Alice sends the number A to Bob.

   Alice sends 2 to Bob.

4. Bob independently chooses another large random integer y and calculates B such that:
   $B = g^y \bmod n$

   Let y = 6. Then, we have, $B = 7^6 \bmod 11 = 117649 \bmod 11 = 4$.

5. Bob sends the number B to Alice.

   Bob sends 4 to Alice.

6 A now computes the secret key K1 as follows:
   $K1 = B^x \bmod n$

   We have, $K1 = 4^3 \bmod 11 = 64 \bmod 11 = 9$.

7. B now computes the secret key K2 as follows:
   $K2 = A^y \bmod n$

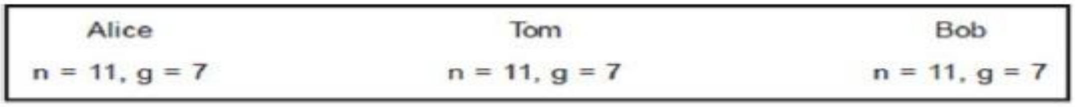
   We have, $K2 = 2^6 \bmod 11 = 64 \bmod 11 = 9$.

7.8.4 Diffie-Hellman Key exchange

1. Public values:
   large prime p, generator g (primitive root of p)
2. Alice has secret value x, Bob has secret y
3. Discrete logarithm problem: given x, g, and n, find A
4. A B: gx (mod n)
5. B A: gy (mod n)
6. Bob computes (gx)y = gxy(mod n)
7. Alice computes (gy)x = gxy (mod n)
8. Symmetric key= gxy (mod n)

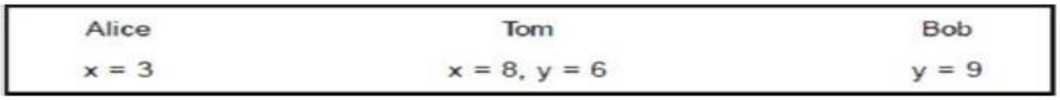
7.8.5 Limitation: Vulnerable to "man in the middle" attacks*

7.8.5.1 Man-in-the-Middle Attack:



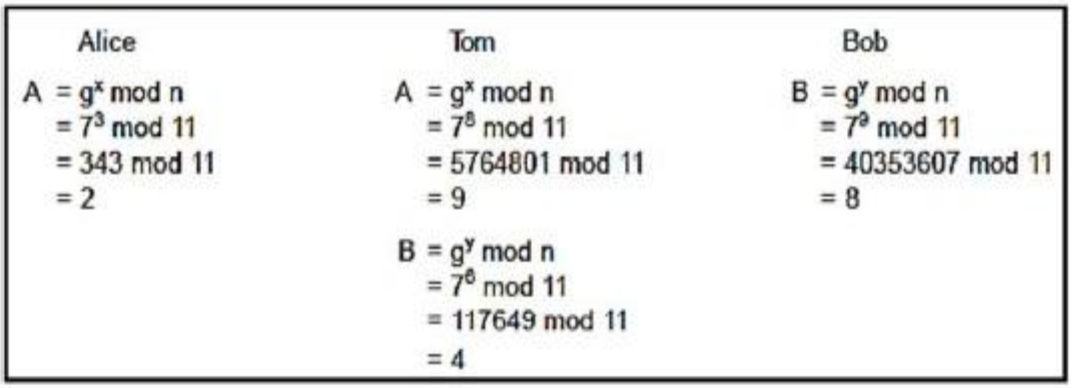
Alice	Tom	Bob
n = 11, g = 7	n = 11, g = 7	n = 11, g = 7

Figure 7.1 Man-in-the-Middle Attack Part-I



Alice	Tom	Bob
x = 3	x = 8, y = 6	y = 9

Figure 7.2 Man-in-the-Middle Attack Part-II



Alice	Tom	Bob
A $= g^x$ mod n	A $= g^x$ mod n	B $= g^y$ mod n
$= 7^3$ mod 11	$= 7^8$ mod 11	$= 7^9$ mod 11
= 343 mod 11	= 5764801 mod 11	= 40353607 mod 11
= 2	= 9	= 8
	B $= g^y$ mod n	
	$= 7^6$ mod 11	
	= 117649 mod 11	
	= 4	

Figure 7.3 Man-in-the-Middle Attack Part-III

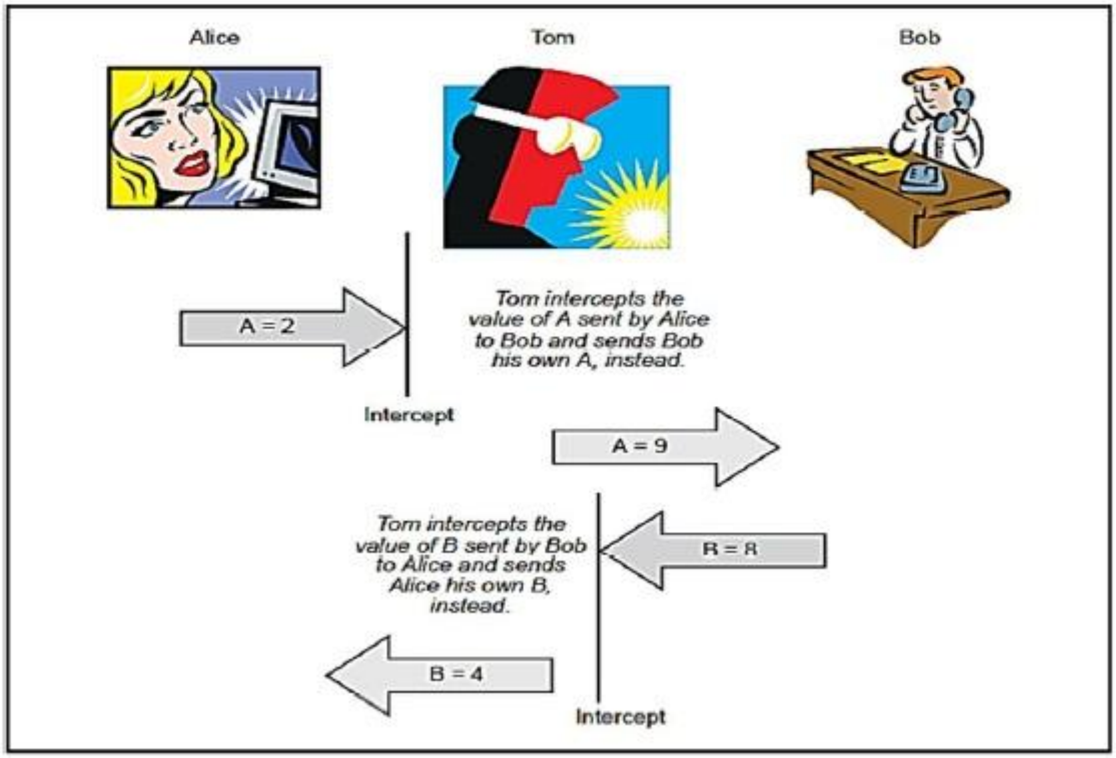
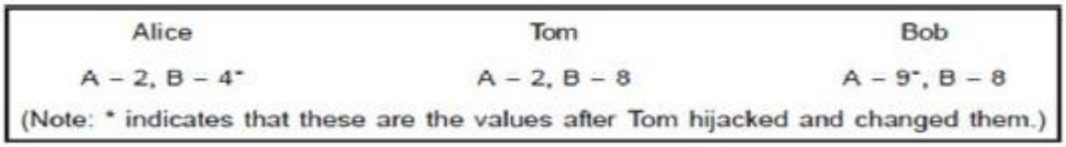

Alice
Tom
Bob
A = 2
Intercept
Tom intercepts the value of A sent by Alice to Bob and sends Bob his own A, instead.
A = 9
Tom intercepts the value of B sent by Bob to Alice and sends Alice his own B, instead.
B = 8
B = 4
Intercept

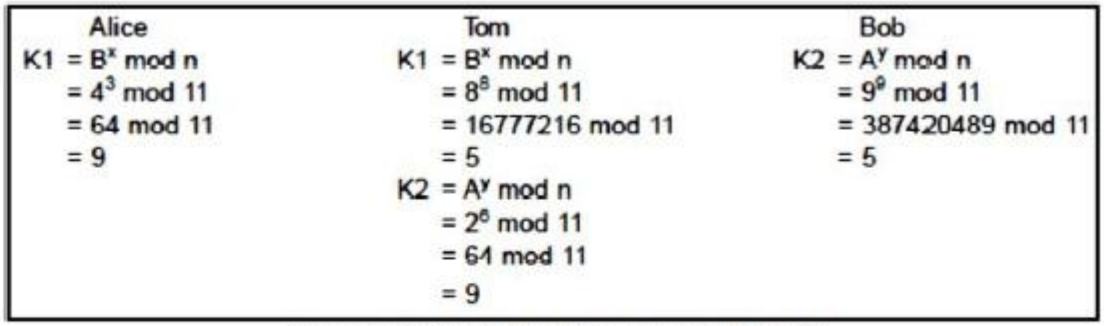
Figure 7.4 Man-in-the-Middle Attack Part-IV



Alice	Tom	Bob
A – 2, B – 4*	A – 2, B – 8	A – 9*, B – 8

(Note: * indicates that these are the values after Tom hijacked and changed them.)

Figure 7.5 Man-in-the-Middle Attack Part-V



Alice	Tom	Bob
K1 $= B^x \bmod n$	K1 $= B^x \bmod n$	K2 $= A^y \bmod n$
$= 4^3 \bmod 11$	$= 8^8 \bmod 11$	$= 9^9 \bmod 11$
$= 64 \bmod 11$	$= 16777216 \bmod 11$	$= 387420489 \bmod 11$
$= 9$	$= 5$	$= 5$
	K2 $= A^y \bmod n$	
	$= 2^6 \bmod 11$	
	$= 64 \bmod 11$	
	$= 9$	

Figure 7.6 Man-in-the-Middle Attack Part-VI

Preventing a Man-in-the-Middle Attack with Hashing

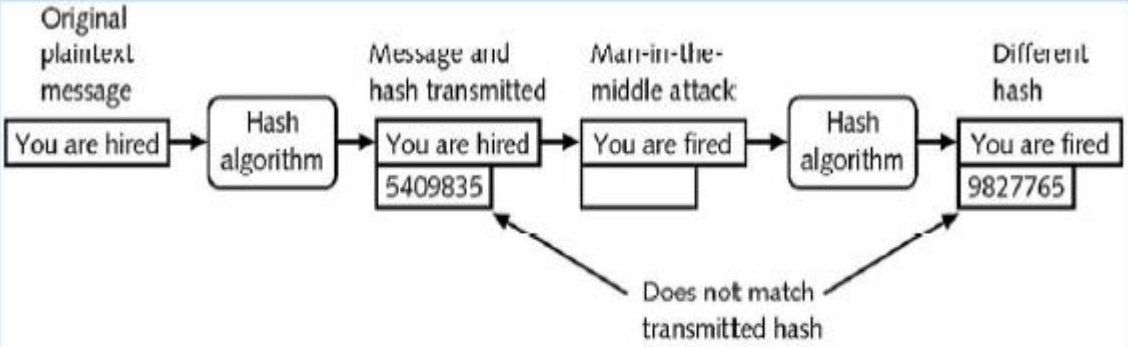


Figure 11-4 Man-in-the-middle attack defeated by hashing

Conclusion: Thus in this experiment we have studied and implement Diffie-Hellmen key exchange algorithm and how to prevent Man-in-the-Middle Attack

ASSIGNMENT NO. 9

Title:

Implementation of RSA Algorithm

Problem Statement:

Implementation of RSA Algorithm

Theory:

RSA (Rivest, Shamir & Adleman)

RSA is an algorithm used by modern computers to encrypt and decrypt messages. It is an asymmetric cryptographic algorithm. Asymmetric means that there are two different keys. This is also called public key cryptography, because one of the keys can be given to anyone. The other key must be kept private. The algorithm is based on the fact that finding the factors of a large composite number is difficult: when the integers are prime numbers, the problem is called prime factorization. It is also a key pair (public and private key) generator.

RSA makes the public and prívate keys by multiplying two large prime numbers *p* and *q*

- It’s easy to find & multiply large prime No. (n=pq)
- It is very difficult to factor the number n to find p and q
- Finding the private key from the public key would require a factoring operation
- The real challenge is the selection & generation of keys.
- RSA is complex and slow, but secure
- 100 times slower than DES on s/w & 1000 times on h/w

The Rivest-Shamir-Adleman (RSA) algorithm is one of the most popular and secures public key encryption methods. The algorithm capitalizes on the fact that there is no efficient way to factor very large (100-200 digit) numbers.

Using an encryption key (*e*, *n*), the algorithm is as follows:

1. Represent the message as an integer between 0 and (n-1). Large messages can be broken up into a number of blocks. Each block would then be represented by an integer in the same range.

2. Encrypt the message by raising it to the *e*th power modulo n. The result is a cipher text message C.

3. To decrypt cipher text message C, raise it to another power d modulo n

The encryption key (e, n) is made public. The decryption key (d, n) is kept private by the user.

How to Determine Appropriate Values for *e*, *d*, and *n*

1. Choose two very large (100+ digit) prime numbers. Denote these numbers as p and q.
2. Set n equal to p * q.
3. Choose any large integer, d, such that GCD (d, ((p-1) * (q-1))) = 1
4. Find e such that e * d = 1 (**mod** ((p-1) * (q-1)))

Rivest, Shamir, and Adleman provide efficient algorithms for each required operation [4].

How secure is a communication using RSA?

Cryptographic methods cannot be proven secure. Instead, the only test is to see if someone can figure out how to decipher a message without having direct knowledge of the decryption key.

The RSA method's security rests on the fact that it is extremely difficult to factor very large numbers. If 100 digit numbers are used for *p* and *q*, the resulting *n* will be approximately 200 digits. The fastest known factoring algorithm would take far too long for an attacker to ever break the code. Other methods for determining *d* without factoring *n* are equally as difficult.

Any cryptographic technique which can resist a concerted attack is regarded as secure. At this point in time, the RSA algorithm is considered secure.

How Does RSA Works?

RSA is an **asymmetric** system, which means that a key pair will be generated (we will see how soon) , a **public** key and a **private** key , obviously you keep your private key secure and pass around the public one.

The algorithm was published in the 70's by Ron **R**ivest, Adi **S**hamir, and Leonard **A**dleman, hence RSA, and it sort of implement's a trapdoor function such as Diffie's one.

RSA is rather slow so it's hardly used to encrypt data, more frequently it is used to encrypt and pass around **symmetric** keys which can actually deal with encryption at a **faster** speed.

RSA Security:

• It uses prime number theory which makes it difficult to find out the key by reverse

Engineering.

• Mathematical Research suggests that it would take more than 70 years to find P & Q if

N is a 100 digit number.

Algorithm

The RSA algorithm holds the following features −

- RSA algorithm is a popular exponentiation in a finite field over integers including prime numbers.
- The integers used by this method are sufficiently large making it difficult to solve.
- There are two sets of keys in this algorithm: private key and public key.

You will have to go through the following steps to work on RSA algorithm

Step 1: Generate the RSA modulus

The initial procedure begins with selection of two prime numbers namely p and q, and then calculating their product N, as shown,

N=p*q

Here, let N be the specified large number.

Step 2: Derived Number (e)

Consider number e as a derived number which should be greater than 1 and less than (p-1) and (q-1). The primary condition will be that there should be no common factor of (p-1) and (q-1) except 1

Step 3: Public key

The specified pair of numbers **n** and **e** forms the RSA public key and it is made public.

Step 4: Private Key

Private Key **d** is calculated from the numbers p, q and e. The mathematical relationship between the numbers is as follows −

ed = 1 mod (p-1) (q-1)

The above formula is the basic formula for Extended Euclidean Algorithm, which takes p and q as the input parameters.

Encryption Formula

Consider a sender who sends the plain text message to someone whose public key is **(n,e).** To encrypt the plain text message in the given scenario, use the following syntax −

C = Pe mod n

Decryption Formula

The decryption process is very straightforward and includes analytics for calculation in a systematic approach. Considering receiver **C** has the private key **d**, the result modulus will be calculated as

Plaintext = Cd mod n

Example

1. P=7, Q=17
2. 119=7*17
3. (7-1)*(17-1)= 6*16 =96 factor 2 & 3, so E=5
4. (D*5) mod (7-1)*(17-1)=1, so D=77
5. CT=105 mod 119 =100000 mod 119 =40
6. Send 40
7. PT=4077 mod 119 = 10

Conclusion

Thus in this experiment we learn that how to Encrypt and Decrypt the message by using RSA Algorithm.

ASSIGNMENT NO. 10

Title:

Implementation of ECC Algorithm

Problem Statement:

Implementation of ECC Algorithm

Theory:

ECC (Elliptic Curve Cryptography) is a modern and efficient type of public key cryptography. Its security is based on the difficulty to solve discrete logarithms on the field defined by specific equations computed over a curve.

ECC can be used to create digital signatures or encrypting data.

The main benefit of ECC is that the size of a key is significantly smaller than with more traditional algorithms like RSA or DSA.

For instance, consider the security level equivalent to AES128: an RSA key of similar strength must have a modulus of 3072 bits (therefore the total size is 768 bytes, comprising modulus and private exponent). An ECC private needs as little as 256 bits (32 bytes).

Elliptic-curve cryptography (ECC) is an approach to public-key cryptography based on the algebraic structure of elliptic curves over finite fields. ECC requires smaller keys compared to non-EC cryptography (based on plain Galois fields) to provide equivalent security.

Elliptic curves are applicable for key agreement, digital signatures, pseudo-random generators and other tasks. Indirectly, they can be used for encryption by combining the key agreement with a symmetric encryption scheme. They are also used in several integer factorization algorithms based on elliptic curves that have applications in cryptography, such as Lenstra elliptic-curve factorization.

Key Benefits of ECC:

ECC key is very helpful for the current generation as more people are moving to the Smartphone. As the utilization of Smartphone extends to grow, there is an emerging need for a more flexible encryption for business to meet with increasing security requirements.

Stronger Keys:

ECC stands for Elliptic Curve Cryptography is the latest encryption method offers stronger security. If we compare to the RSA and DSA algorithms, then 256-bit ECC is equal to 3072-bit RSA key. The reason behind keeping short key is the use of less computational power, fast and secures connection, ideal for Smartphone and tablet too.

The US government and the National Security Agency have certified ECC encryption method. The mathematical problem of the ECC algorithm, It is harder to break for hackers compare to RSA and DSA, which means the ECC algorithm ensures web site and infrastructure safety than traditional methods in a more secure manner.

Shorter Key Size:

The elliptic curve cryptography (ECC) certificates allow key size to remain small while providing a higher level of security. ECC certificates key creation method is entirely different from previous algorithms, while relying on the use of a public key for encryption and a private key for decryption. By starting small and with a slow growth potential, ECC has longer potential lifespan. Elliptic curves are likely to be the next generation of cryptographic algorithms, and we are seeing the beginning of their use now.

Why Elliptic Curve Cryptography is Required ?

Encryption experts are pressed to find ever more effective methods, measured in security and performance, because the threats presented by hackers are ever greater – partly because the hackers themselves become more sophisticated in their attacks, and also because the fallout from an attack gets ever more dangerous as our use of data grows. It creates an urgency of new algorithms with a goal to provide a higher level of security by having keys that are more difficult to break, while offering better performance across the network and while working with large data sets.

Example:

Elliptic Curve Cryptography (ECC) was discovered in 1985 by Victor Miller (IBM) and Neil Koblitz (University of Washington) as an alternative mechanism for implementing public-key cryptography.

We assume that those who are going through this article will have a basic understanding of cryptography (terms like encryption and decryption) .

The equation of an elliptic curve is given as,

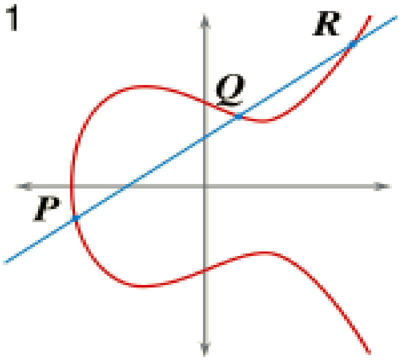
$$y^2 = x^3 + ax + b$$

Few terms that will be used,

E -> Elliptic Curve

P -> Point on the curve

n -> Maximum limit (This should be a prime number)



$$y^2 = x^3 + ax + b$$

The figure above shows that simple elliptic curve.

Key Generation:

Key generation is an important part where we have to generate both public key and private key. The sender will be encrypting the message with receiver's public key and the receiver will decrypt its private key.

Now, we have to select a number 'd' within the range of 'n'.

Using the following equation we can generate the public key

$$\mathbf{Q = d * P}$$

d = The random number that we have selected within the range of (1 to n-1). P is the point on the curve.

'Q' is the public key and'd' is the private key.

Encryption:

Let 'm' be the message that we are sending. We have to represent this message on the curve. These have in-depth implementation details. All the advance research on ECC is done by a company called certicom.

Conside 'm' has the point 'M' on the curve 'E'. Randomly select 'k' from [1 – (n-1)].

Two cipher texts will be generated let it be C1 and C2.

$$\mathbf{C1 = k*P}$$

$$\mathbf{C2 = M + k*Q}$$

C1 and C2 will be send.

Decryption:

We have to get back the message 'm' that was send to us,

M = C2 – d * C1

M is the original message that we have send.

Proof

How does we get back the message,

M = C2 – d * C1

‘M’ can be represented as ‘C2 – d * C1’

C2 – d * C1 = (M + k * Q) – d * (k * P) (C2 = M + k * Q and C1 = k * P)

= M + k * d * P – d * k *P (canceling out k * d * P)

= M (Original Message)

Conclusion:

Thus in this experiment we have studied and implemented ECC algorithm in detail successfully.

ASSIGNMENT NO. 11

Mini Project 1: SQL Injection attacks and Cross -Site Scripting attacks are the two most common attacks on web application. Develop a new policy based Proxy Agent, which classifies the request as a scripted request or query based request, and then, detects the respective type of attack, if any in the request. It should detect both SQL injection attack as well as the Cross-Site Scripting attacks

OR

Mini Project 2: This task is to demonstrate insecure and secured website. Develop a web site and demonstrate how the contents of the site can be changed by the attackers if it is http based and not secured. You can also add payment gateway and demonstrate how money transactions can be hacked by the hackers. Then support your website having https with SSL and demonstrate how secured website is.