Practical No: 01

Aim: Install, configure and run Hadoop and HDFS and explore HDFS.

1] Hadoop is an open source distributed computed Framework that enables the processing of large data set across cluster of computers using simple programming models.

2] It provides highly scalable and Fault tolerent enviroment for processing and storing Large data sets.

3] Hadoop distributed file system (HDFS) is distributed file system that stores data across multiple maching in a hadoop cluster.

4] HDFS is highly Fault tolerant and provides high throughput access to application data.

practical NO:02

Aim: Implement word count / Frequency programs Using MapReduce.

1] MapReduce is programming model used for processing and generating large data sets in parallel and distributed fusion.

2] It divesles a large dataset into smaller chunks. distribute them across cluster, and processes them in parallel.

3] The MapReduce process consists of two phoses. The map phase and The Reduce phase.

4] In the map phase, data is transformed in to key value pairs, and in reduced phase, this key value pairs are aggregated to generate final output.

5] The wordcount program is a classic example of mapReduce that counts the occurances of words in Large datasets.

6] It involves mapping each word to key value pair and then reducing counts of same word.

Practical NO: 03

Aim: Implement an application that stores big data in Hbase / MongoDB and manipulate it using R / Python.

1] HBase is distributed NoSQL database that is built on top of the Hadoop distributed file system (HDFS)

2] It provides real time read and write access to large dataset, and it is designed to store and manage Structured data.

3] HBase is widely used for storing and processing large amount of data in a scalable and fault tolerent manner.

4] MongoDB is popular NoSQL Database that uses a document oriented data model.

5] It stores data in JSON-like documents and supports dynamic scheme design.

6] MongoDB provides high performance scalibility and availibility, and it is often used for building modern web applications and managing Large database.

Practical NO : 05

Aim: Implement Decision tree classification techniques.

* Describe Decision tree classification in deta.

Decision tree is popular classification algorithm that recursively splits tha data into smaller subsets based on most importaut Features.

Importaut function

1] Vpart (Formula, data) : This is function is used to fit decision tree model to the training set. The 'Formula' argument is Formula of from 'dependaut variable ~ independeut variables.

2] predict (objeet, newdata, type) : This Function is used to predict the test set result, it takes the decision tree model filted on training set as objeet.

3] table (x.y) : This function creates a configency table of x and y variables. It is used to evaluate accuraay of model.

Practical NO: 06

Aim: Implement SVM classification techniques.

Explain sum classification in data:1

1] support vector machine (SUM) is common classification method that combines linear models with instance based learning techniques.

2] SUM select small number of critical boundry instances called support vectors from each class and build linear decision function that seperates them widely as possible.

3] SUM choose the extreme points that help in creating the hyperplane these extreme cases are called a support vectors.

4] Consider diagram in which there are two different categaries that are classification using decision bondary.

practical NO:07

Aim: REGRESSION MODEL Import a data from web storage. Name the dataset and now do Logistic Regression to Find out relation between variables that are affecting the admission of a student in an institute based on his or her GRE score, GPA obtained and rank of the student. Also check the model is fit or not. require (foreign), requires (MASS).

Describe logistic regression

1] logistic regression is a statical method used to analyize relationship between a binary outcame variable and one or more predictor variables.

2] It is type of regression analysis that models probability of binary outcome as a function of predictor variable.

3] The logistic function is used to transform output to range between 0 & 1, representing probabitiy of outcome.

4] Logistic regression is commonly used in fields such as healthcame, finance and marketing to predict likehood of an event occuring based on various input factors.

practical NO: 08

Aim: MULTIPLE REGRESSION MODEL Apply multiple regressions, if data have a continious independent variable. Apply on above dataset.

- Explain multiple regression in detail.

1] multiple regression is statical techniques used to analyze relationship between a dependent variable and two or more independant variable.

2] It extends simple linear regression to include multiple predictors, allowing for the examination of how predictors jointly affect outcome.

3] The goal is to estimate the strength and direction of relationship between the independant data dependant variable as well as to predict value of dependant variable based on values of independent variable.

Practical NO : 09

Practical No: 10

Aim: CLUSTERING MODEL a. Clustering algorithms for unsupervised classification. b. Plot the cluster data using R visualization.

- Describe K means clustering

1] K means algorithm is clustering techniques used to partition a collection of m object into K district cluster.

step 1: Choose the value of K and initial gusses for centroids.

step 2: Calculate distance from each cluster and update centroid accordingly

step 3: Compute centroid of each ~~out~~ cluster and update centroid accordingly.

step 4: Repeat step 2 and 3 until assignments of data points to cluster no longer change.

Important Function

1] woss = vector (): This is used to create empty vector 'woss

2] This is used to implement elbow method
- It runs loop from 1 to 10 and comptes sum of squard distance of data.

3] K'mean's = kmeans (x = dataset, centers = 5):
This is used to perform kmeans clustering with 5 clusters on 'dataset'.