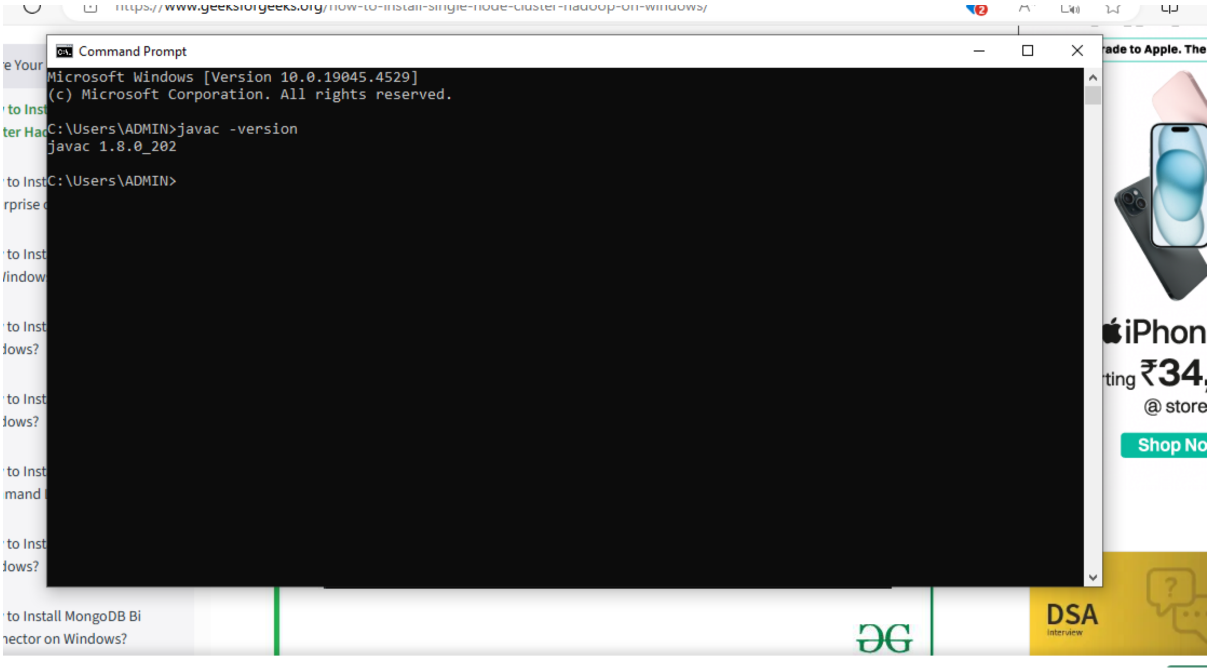
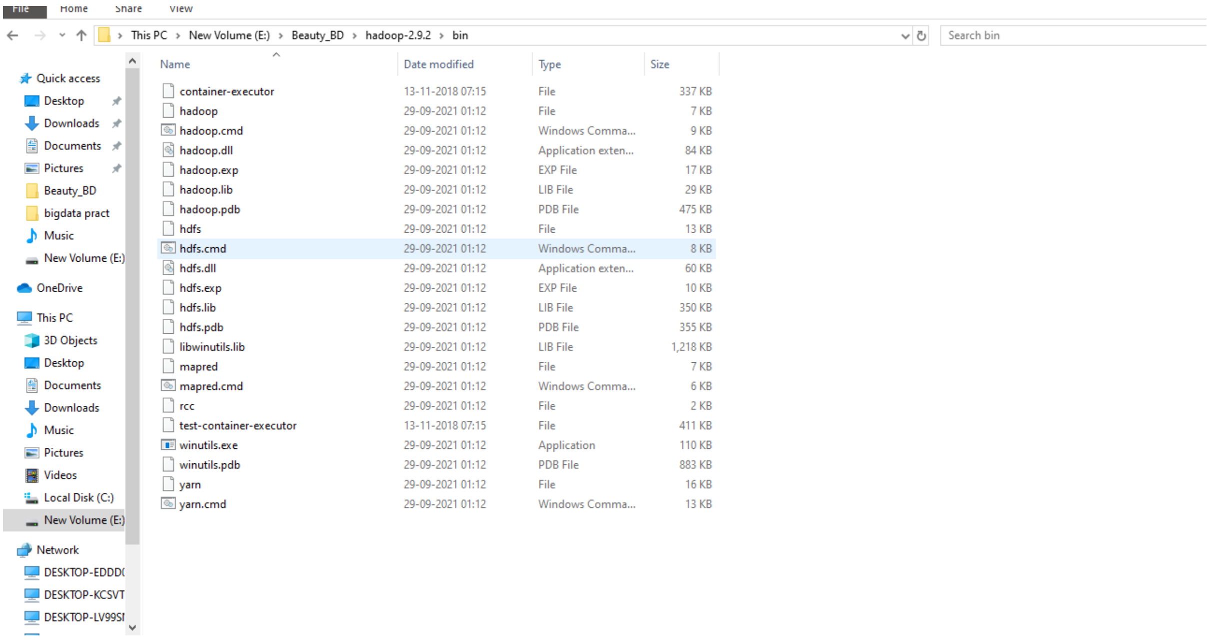
Practical.1

Step 1 :-**Verify the Java installed**



```
Microsoft Windows [Version 10.0.19045.4529]
(c) Microsoft Corporation. All rights reserved.

C:\Users\ADMIN>javac -version
javac 1.8.0_202

C:\Users\ADMIN>
```

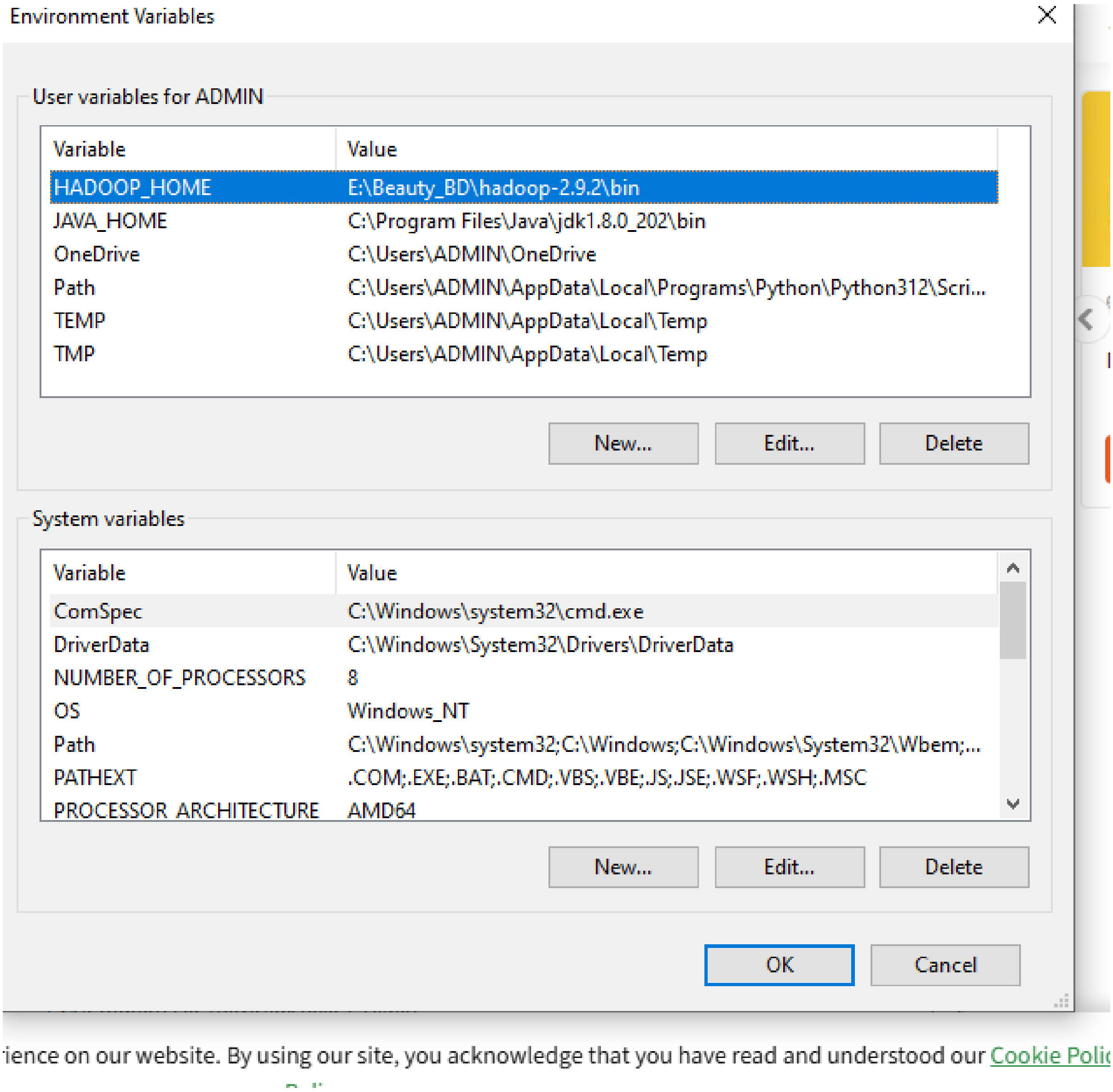
Step 2: Extract Hadoop at C:\Hadoop



Name	Date modified	Type	Size
container-executor	13-11-2018 07:15	File	337 KB
hadoop	29-09-2021 01:12	File	7 KB
hadoop.cmd	29-09-2021 01:12	Windows Comma...	9 KB
hadoop.dll	29-09-2021 01:12	Application exten...	84 KB
hadoop.exp	29-09-2021 01:12	EXP File	17 KB
hadoop.lib	29-09-2021 01:12	LIB File	29 KB
hadoop.pdb	29-09-2021 01:12	PDB File	475 KB
hdfs	29-09-2021 01:12	File	13 KB
hdfs.cmd	29-09-2021 01:12	Windows Comma...	8 KB
hdfs.dll	29-09-2021 01:12	Application exten...	60 KB
hdfs.exp	29-09-2021 01:12	EXP File	10 KB
hdfs.lib	29-09-2021 01:12	LIB File	350 KB
hdfs.pdb	29-09-2021 01:12	PDB File	355 KB
libwinutils.lib	29-09-2021 01:12	LIB File	1,218 KB
mapred	29-09-2021 01:12	File	7 KB
mapred.cmd	29-09-2021 01:12	Windows Comma...	6 KB
rcc	29-09-2021 01:12	File	2 KB
test-container-executor	13-11-2018 07:15	File	411 KB
winutils.exe	29-09-2021 01:12	Application	110 KB
winutils.pdb	29-09-2021 01:12	PDB File	883 KB
yarn	29-09-2021 01:12	File	16 KB
yarn.cmd	29-09-2021 01:12	Windows Comma...	13 KB

Step 3: Setting up the HADOOP_HOME variable And

Step 4: Set JAVA_HOME variable



User variables for ADMIN

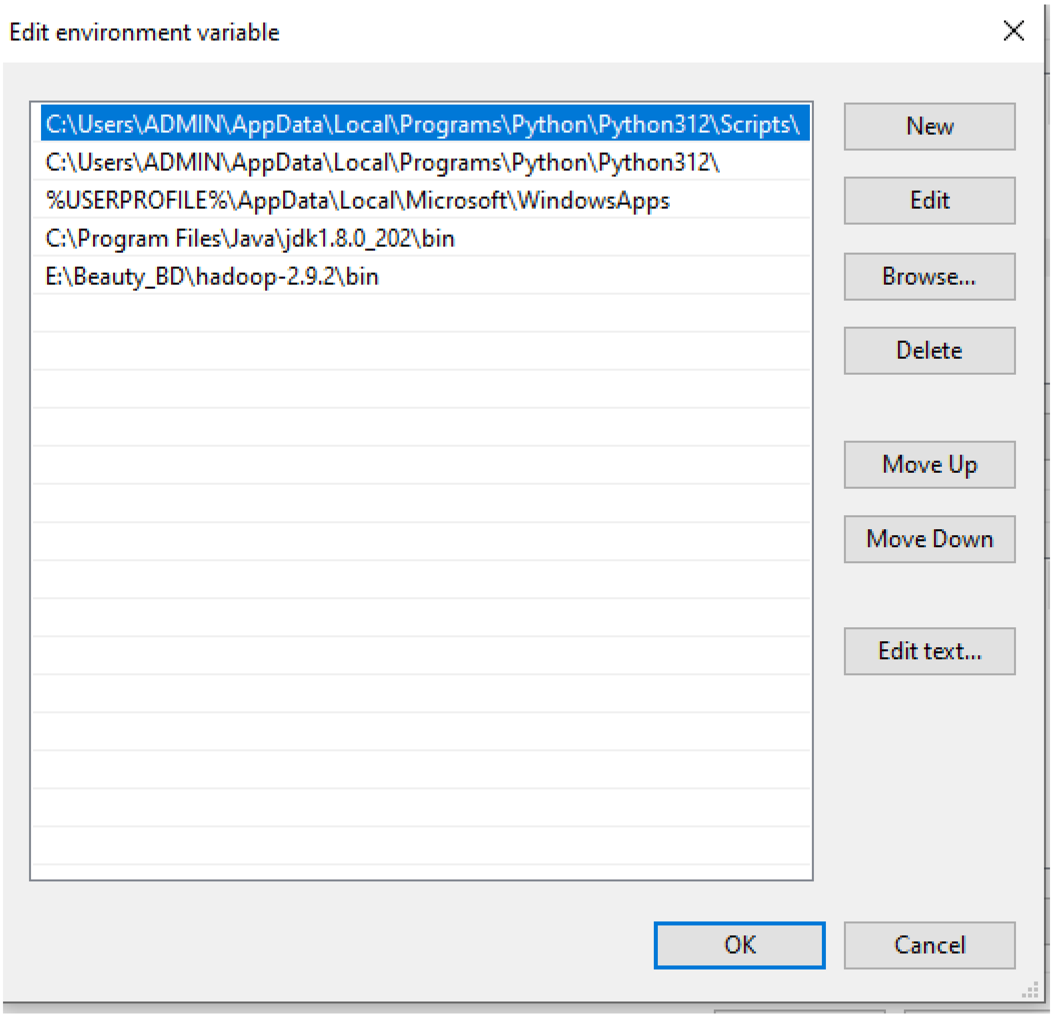
Variable	Value
HADOOP_HOME	E:\Beauty_BD\hadoop-2.9.2\bin
JAVA_HOME	C:\Program Files\Java\jdk1.8.0_202\bin
OneDrive	C:\Users\ADMIN\OneDrive
Path	C:\Users\ADMIN\AppData\Local\Programs\Python\Python312\Scri...
TEMP	C:\Users\ADMIN\AppData\Local\Temp
TMP	C:\Users\ADMIN\AppData\Local\Temp

System variables

Variable	Value
ComSpec	C:\Windows\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	8
OS	Windows_NT
Path	C:\Windows\system32;C:\Windows;C:\Windows\System32\Wbem;...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64

ience on our website. By using our site, you acknowledge that you have read and understood our Cookie Polic

Step 5: Set Hadoop and Java bin directory path



```
C:\Users\ADMIN\AppData\Local\Programs\Python\Python312\Scripts\
C:\Users\ADMIN\AppData\Local\Programs\Python\Python312\
%USERPROFILE%\AppData\Local\Microsoft\WindowsApps
C:\Program Files\Java\jdk1.8.0_202\bin
E:\Beauty_BD\hadoop-2.9.2\bin
```

Step 6: Hadoop Configuration :

```
1. Core-site.xml
2. Mapred-site.xml
3. Hdfs-site.xml
4. Yarn-site.xml
5. Hadoop-env.cmd
6. Create two folders datanode and namenode
```

Step 6.1: Core-site.xml configuration

```
<configuration>
   <property>
       <name>fs.defaultFS</name>
       <value>hdfs://localhost:9000</value>
   </property>
</configuration>
```

Step 6.2: Mapred-site.xml configuration

```
<configuration>
   <property>
       <name>mapreduce.framework.name</name>
       <value>yarn</value>
   </property>
</configuration>
```

Step 6.3: Hdfs-site.xml configuration

```
<configuration>
   <property>
       <name>dfs.replication</name>
       <value>1</value>
   </property>
   <property>
       <name>dfs.namenode.name.dir</name>
       <value>C:\hadoop-2.8.0\data\namenode</value>
   </property>
   <property>
       <name>dfs.datanode.data.dir</name>
       <value>C:\hadoop-2.8.0\data\datanode</value>
   </property>
</configuration>
```

Step 6.4: Yarn-site.xml configuration

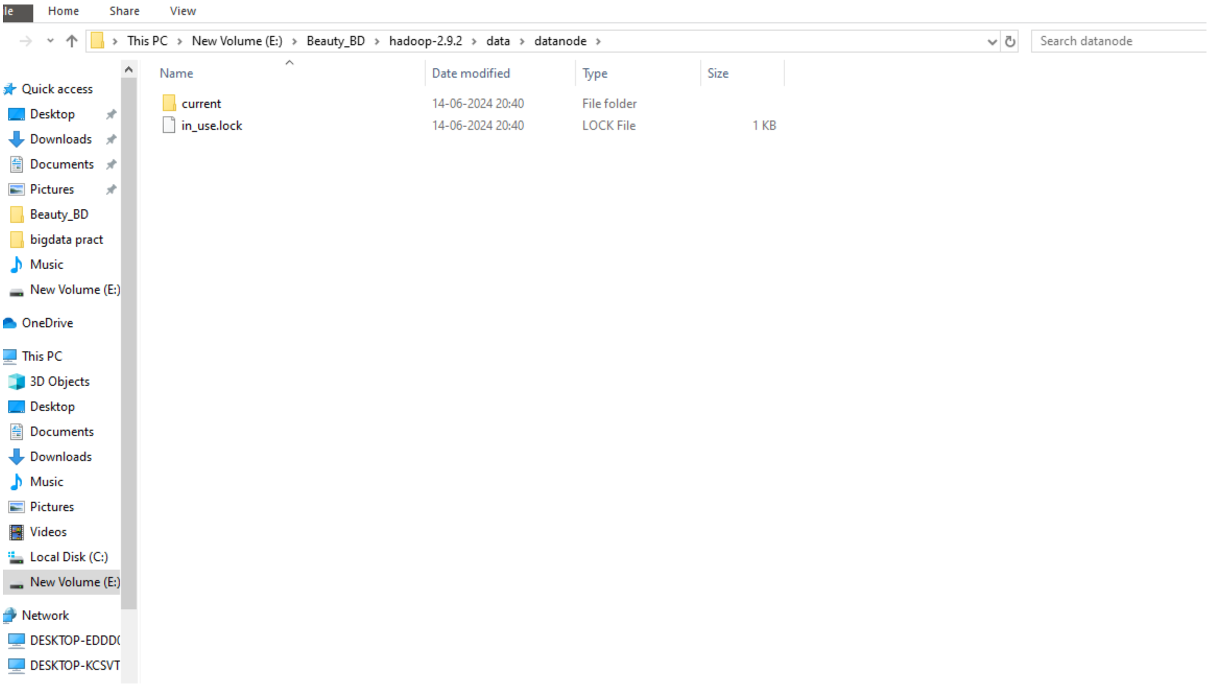
```
<configuration>
   <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
   </property>
   <property>

<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
   </property>
```

```
</configuration>
```

Step 6.5: Hadoop-env.cmd configuration

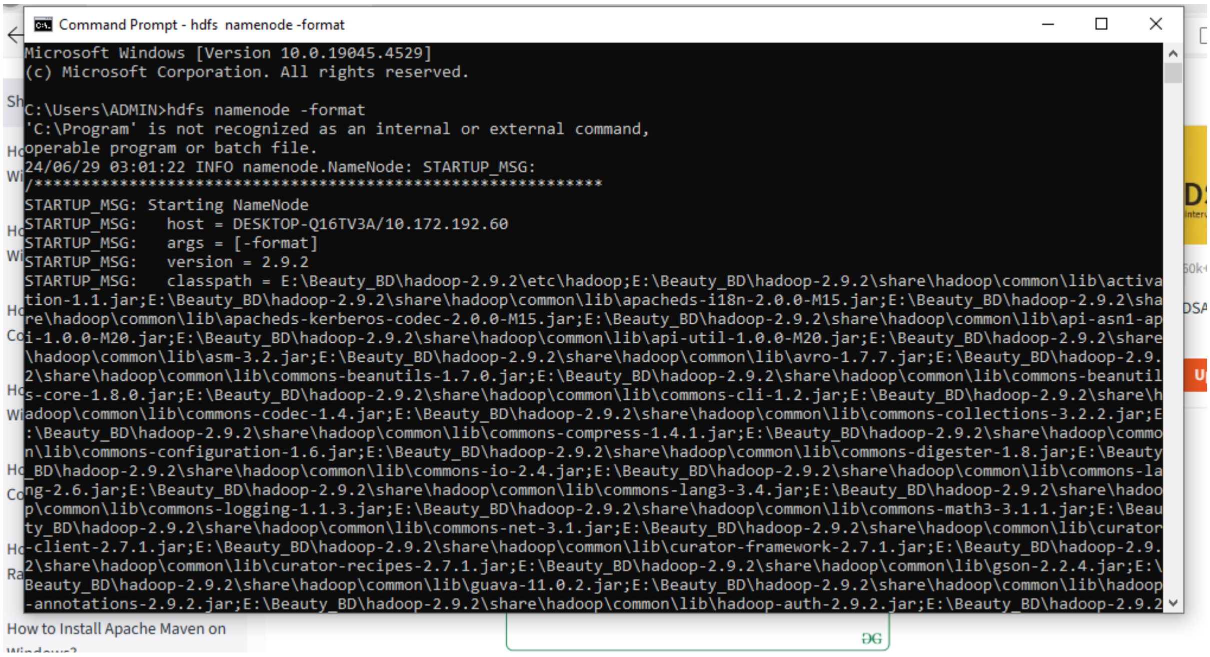
```
Set "JAVA_HOME=C:\Java" (On C:\java this is path to file jdk.18.0)
```

Step 6.6: Create datanode and namenode folders



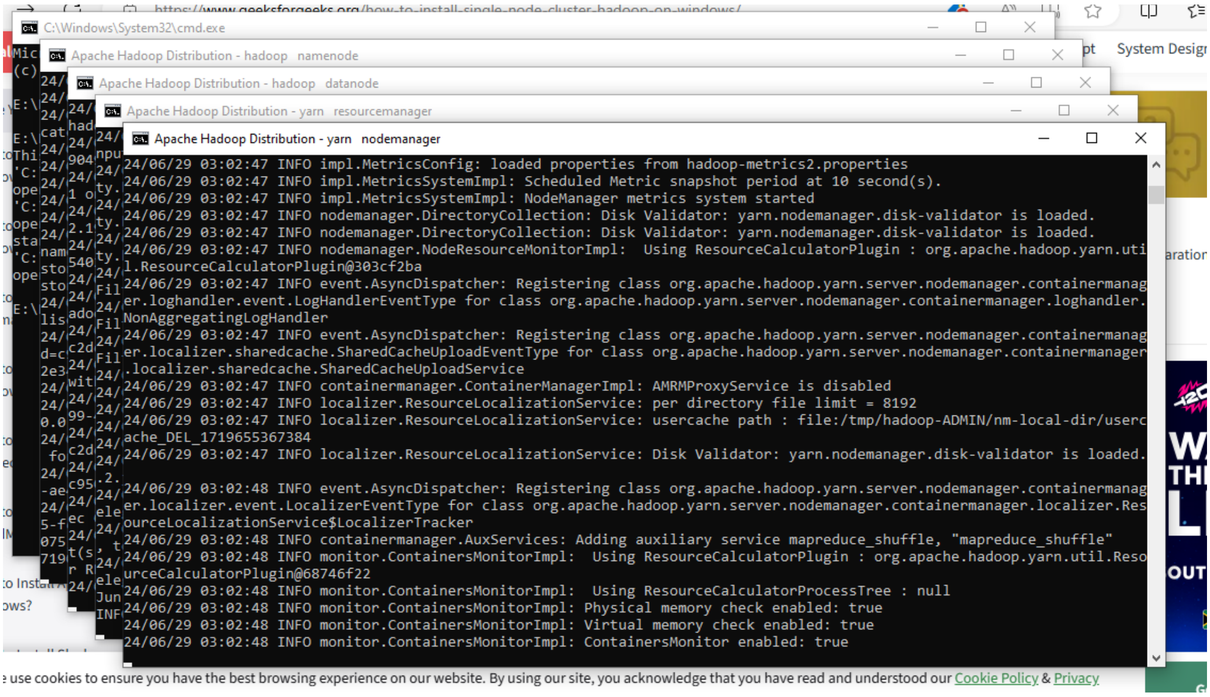
Name	Date modified	Type	Size
current	14-06-2024 20:40	File folder	
in_use.lock	14-06-2024 20:40	LOCK File	1 KB

Step 7: Format the namenode folder



```
Microsoft Windows [Version 10.0.19045.4529]
(c) Microsoft Corporation. All rights reserved.

C:\Users\ADMIN>hdfs namenode -format
'C:\Program' is not recognized as an internal or external command,
operable program or batch file.
24/06/29 03:01:22 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = DESKTOP-Q16TV3A/10.172.192.60
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.9.2
STARTUP_MSG:   classpath = E:\Beauty_BD\hadoop-2.9.2\etc\hadoop;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\activa
tion-1.1.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\apacheds-i18n-2.0.0-M15.jar;E:\Beauty_BD\hadoop-2.9.2\sha
re\hadoop\common\lib\apacheds-kerberos-codec-2.0.0-M15.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\api-asn1-ap
i-1.0.0-M20.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\api-util-1.0.0-M20.jar;E:\Beauty_BD\hadoop-2.9.2\share
\hadoop\common\lib\asm-3.2.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\avro-1.7.7.jar;E:\Beauty_BD\hadoop-2.9.
2\share\hadoop\common\lib\commons-beanutils-1.7.0.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\commons-beanutil
s-core-1.8.0.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\commons-cli-1.2.jar;E:\Beauty_BD\hadoop-2.9.2\share\h
adoop\common\lib\commons-codec-1.4.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\commons-collections-3.2.2.jar;E
:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\commons-compress-1.4.1.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\commo
n\lib\commons-configuration-1.6.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\commons-digester-1.8.jar;E:\Beauty
_BD\hadoop-2.9.2\share\hadoop\common\lib\commons-io-2.4.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\commons-la
ng-2.6.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\commons-lang3-3.4.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoo
p\common\lib\commons-logging-1.1.3.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\commons-math3-3.1.1.jar;E:\Beau
ty_BD\hadoop-2.9.2\share\hadoop\common\lib\commons-net-3.1.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\curator
-client-2.7.1.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\curator-framework-2.7.1.jar;E:\Beauty_BD\hadoop-2.9.
2\share\hadoop\common\lib\curator-recipes-2.7.1.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\gson-2.2.4.jar;E:\
Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\guava-11.0.2.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\hadoop
-annotations-2.9.2.jar;E:\Beauty_BD\hadoop-2.9.2\share\hadoop\common\lib\hadoop-auth-2.9.2.jar;E:\Beauty_BD\hadoop-2.9.2
```

Step 8: Testing the setup

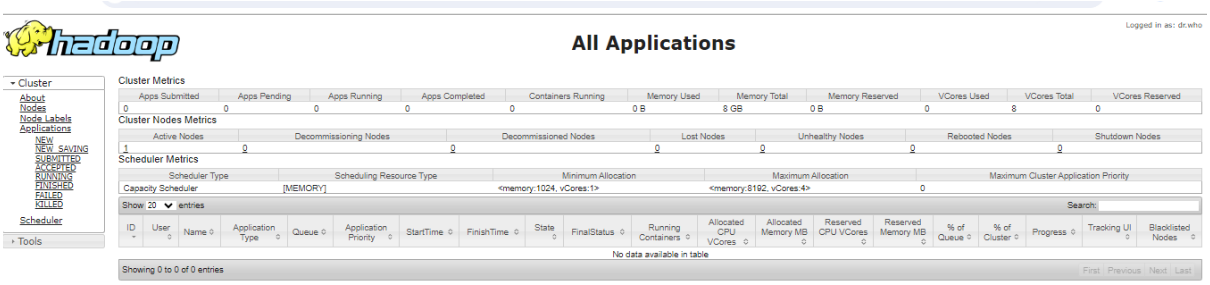

C:\Windows\System32\cmd.exe
Apache Hadoop Distribution - hadoop   namenode
Apache Hadoop Distribution - hadoop   datanode
Apache Hadoop Distribution - yarn   resourcemanager
Apache Hadoop Distribution - yarn   nodemanager
System Design

```
24/06/29 03:02:47 INFO impl.MetricsConfig: loaded properties from hadoop-metrics2.properties
24/06/29 03:02:47 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
24/06/29 03:02:47 INFO impl.MetricsSystemImpl: NodeManager metrics system started
24/06/29 03:02:47 INFO nodemanager.DirectoryCollection: Disk Validator: yarn.nodemanager.disk-validator is loaded.
24/06/29 03:02:47 INFO nodemanager.DirectoryCollection: Disk Validator: yarn.nodemanager.disk-validator is loaded.
24/06/29 03:02:47 INFO nodemanager.NodeResourceMonitorImpl:  Using ResourceCalculatorPlugin : org.apache.hadoop.yarn.uti
l.ResourceCalculatorPlugin@303cf2ba
24/06/29 03:02:47 INFO event.AsyncDispatcher: Registering class org.apache.hadoop.yarn.server.nodemanager.containermanag
er.loghandler.event.LogHandlerEventType for class org.apache.hadoop.yarn.server.nodemanager.containermanager.loghandler.
NonAggregatingLogHandler
24/06/29 03:02:47 INFO event.AsyncDispatcher: Registering class org.apache.hadoop.yarn.server.nodemanager.containermanag
er.localizer.sharedcache.SharedCacheUploadEventType for class org.apache.hadoop.yarn.server.nodemanager.containermanager
.localizer.sharedcache.SharedCacheUploadService
24/06/29 03:02:47 INFO containermanager.ContainerManagerImpl: AMRMProxyService is disabled
24/06/29 03:02:47 INFO localizer.ResourceLocalizationService: per directory file limit = 8192
24/06/29 03:02:47 INFO localizer.ResourceLocalizationService: usercache path : file:/tmp/hadoop-ADMIN/nm-local-dir/userc
ache_DEL_1719655367384
24/06/29 03:02:47 INFO localizer.ResourceLocalizationService: Disk Validator: yarn.nodemanager.disk-validator is loaded.

24/06/29 03:02:48 INFO event.AsyncDispatcher: Registering class org.apache.hadoop.yarn.server.nodemanager.containermanag
er.localizer.event.LocalizerEventType for class org.apache.hadoop.yarn.server.nodemanager.containermanager.localizer.Res
ourceLocalizationService$LocalizerTracker
24/06/29 03:02:48 INFO containermanager.AuxServices: Adding auxiliary service mapreduce_shuffle, "mapreduce_shuffle"
24/06/29 03:02:48 INFO monitor.ContainersMonitorImpl:  Using ResourceCalculatorPlugin : org.apache.hadoop.yarn.util.Reso
urceCalculatorPlugin@68746f22
24/06/29 03:02:48 INFO monitor.ContainersMonitorImpl:  Using ResourceCalculatorProcessTree : null
24/06/29 03:02:48 INFO monitor.ContainersMonitorImpl: Physical memory check enabled: true
24/06/29 03:02:48 INFO monitor.ContainersMonitorImpl: Virtual memory check enabled: true
24/06/29 03:02:48 INFO monitor.ContainersMonitorImpl: ContainersMonitor enabled: true
```

e use cookies to ensure you have the best browsing experience on our website. By using our site, you acknowledge that you have read and understood our Cookie Policy & Privacy

Step 8.1: Testing the setup:

Step 9: Open: http://localhost:8088



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	8 GB	0 B	0	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	0	0

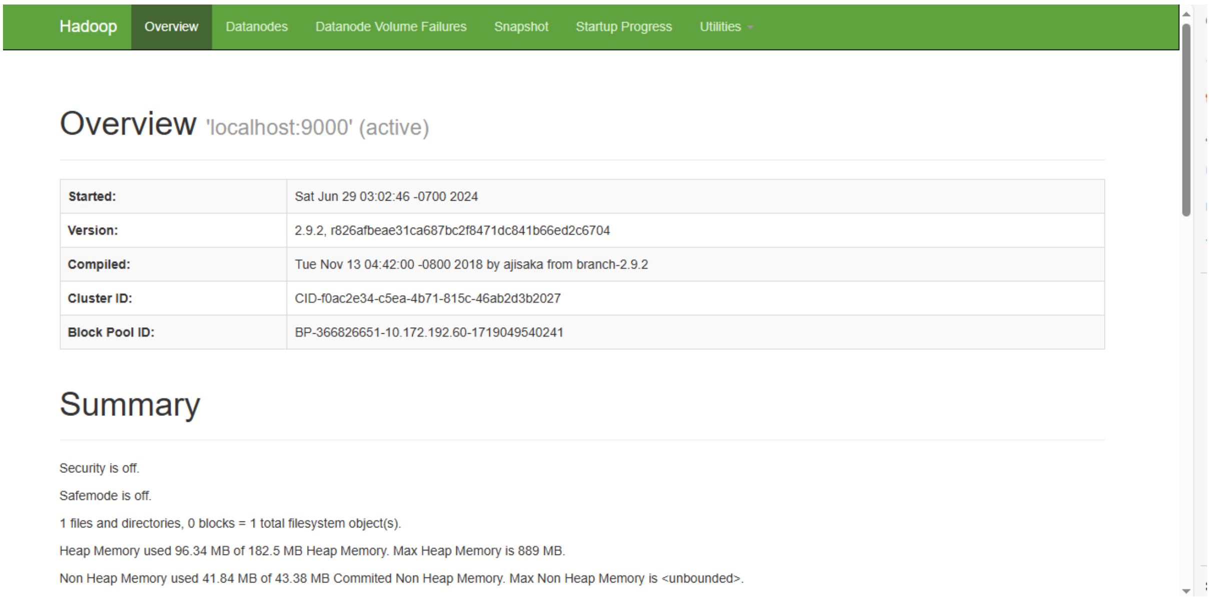
Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
No data available in table																			

Showing 0 to 0 of 0 entries

Step 10: http://localhost:50070



Overview 'localhost:9000' (active)

Started:	Sat Jun 29 03:02:46 -0700 2024
Version:	2.9.2, r826afbeae31ca687bc2f8471dc841b66ed2c6704
Compiled:	Tue Nov 13 04:42:00 -0800 2018 by ajisaka from branch-2.9.2
Cluster ID:	CID-f0ac2e34-c5ea-4b71-815c-46ab2d3b2027
Block Pool ID:	BP-366826651-10.172.192.60-1719049540241

Summary

Security is off.

Safemode is off.

1 files and directories, 0 blocks = 1 total filesystem object(s).

Heap Memory used 96.34 MB of 182.5 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 41.84 MB of 43.38 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

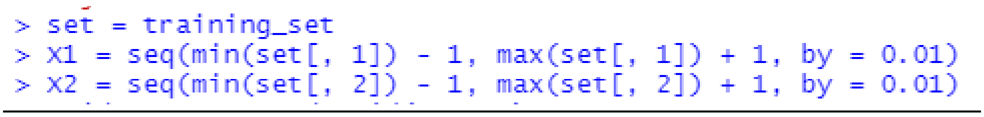
Practical 2:

Aim: Classification using SVM

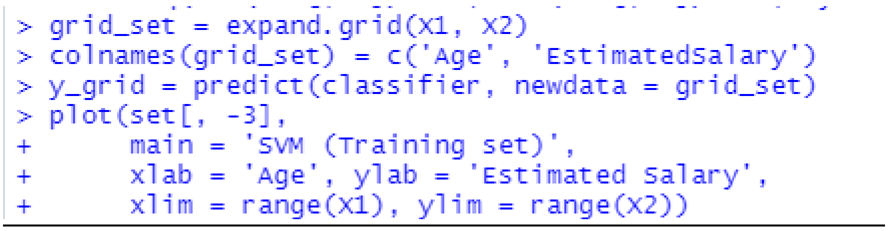
Requirement:
R tool

Code:

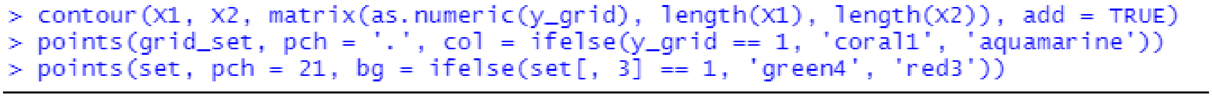
```
getwd()

read.csv()

ds=read.csv("E:/Rajdeep/bigdata pract/dataset/social.csv",TRUE,",")

ds

ds=ds[3:5]

ds

install("catools")

library(caTools)

set.seed(123)

split=sample.split(ds$Purchased, SplitRatio=0.75)

training_set=(subset(ds, split == TRUE))

test_set =(subset(ds, split == FALSE))

ds

test_set[-3]=scale(test_set[-3])

training_set[-3]=scale(training_set[-3])

test_set[-3]

training_set[-3]

install.packages('e1071')

library('e1071')

classifier=svm(formula=Purchased ~ ., data= training_set , type='C-classification',kernal='linear')

classifier

y_pred=predict(classifier, newdata=test_set[-3])

y_pred

cm=table(test_set[, 3],y_pred)

cm
```



```
> set = training_set
> X1 = seq(min(set[, 1]) - 1, max(set[, 1]) + 1, by = 0.01)
> X2 = seq(min(set[, 2]) - 1, max(set[, 2]) + 1, by = 0.01)
```

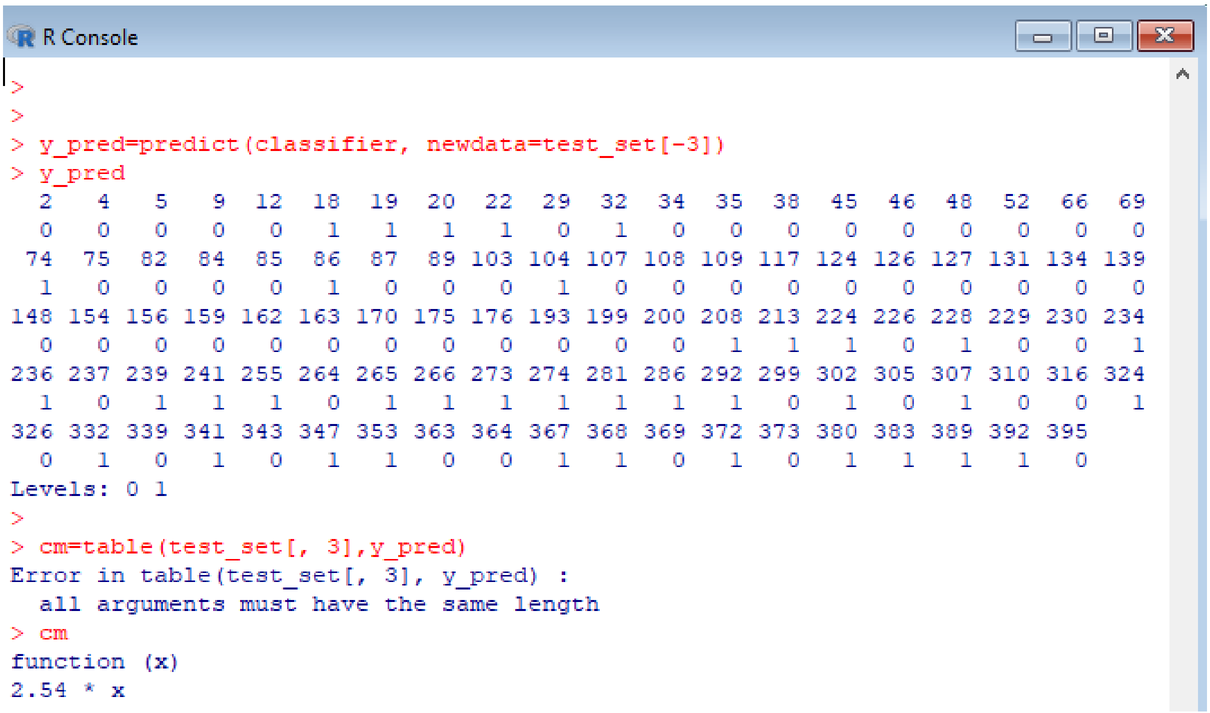


```
> grid_set = expand.grid(X1, X2)
> colnames(grid_set) = c('Age', 'EstimatedSalary')
> y_grid = predict(classifier, newdata = grid_set)
> plot(set[, -3],
+      main = 'SVM (Training set)',
+      xlab = 'Age', ylab = 'Estimated Salary',
+      xlim = range(X1), ylim = range(X2))
```



```
> contour(X1, X2, matrix(as.numeric(y_grid), length(X1), length(X2)), add = TRUE)
> points(grid_set, pch = '.', col = ifelse(y_grid == 1, 'coral1', 'aquamarine'))
> points(set, pch = 21, bg = ifelse(set[, 3] == 1, 'green4', 'red3'))
```

Output:



```
>
>
> y_pred=predict(classifier, newdata=test_set[-3])
> y_pred
  2   4   5   9  12  18  19  20  22  29  32  34  35  38  45  46  48  52  66  69
  0   0   0   0   0   1   1   1   1   0   1   0   0   0   0   0   0   0   0   0
 74  75  82  84  85  86  87  89 103 104 107 108 109 117 124 126 127 131 134 139
  1   0   0   0   0   1   0   0   0   1   0   0   0   0   0   0   0   0   0   0
148 154 156 159 162 163 170 175 176 193 199 200 208 213 224 226 228 229 230 234
  0   0   0   0   0   0   0   0   0   0   0   0   1   1   1   0   1   0   0   1
236 237 239 241 255 264 265 266 273 274 281 286 292 299 302 305 307 310 316 324
  1   0   1   1   1   0   1   1   1   1   1   1   1   0   1   0   1   0   0   1
326 332 339 341 343 347 353 363 364 367 368 369 372 373 380 383 389 392 395
  0   1   0   1   0   1   1   0   0   1   1   0   1   0   1   1   1   1   0
Levels: 0 1
>
> cm=table(test_set[, 3],y_pred)
Error in table(test_set[, 3], y_pred) :
  all arguments must have the same length
> cm
function (x)
2.54 * x
```

Practical 3:

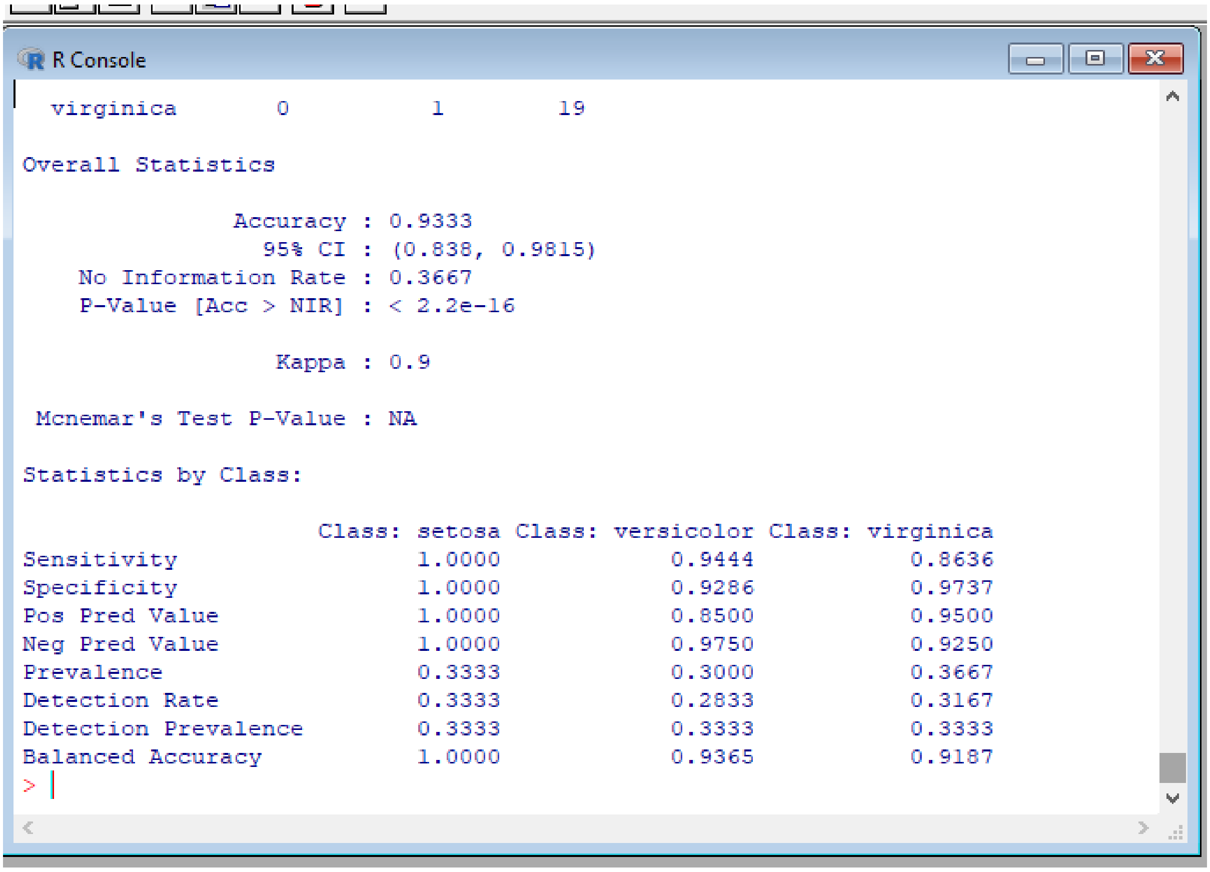
Aim: write program in R of Naive baye's theorem

Requirement:
R tool

Code:

```
data(iris)

str(iris)


install packages("e1071")

install packages("caTools")

install packages("caret")


library(e1071)

library(caTools)

library(caret)


split <- sample.split(iris,SplitRatio=0.7)

train_c1 <-subset(iris,split=="TRUE")

test_c1 <- subset(iris,split == "FALSE")

train_scale <- scale(train_c1[, 1:4])

test_scale <- scale(test_c1[,1:4])


set.seed(120)

classifier_c1 <- naiveBayes(Species ~ ., data = train_c1)

classifier_c1


y_pred <- predict(classifier_c1, newdata= test_c1)

cm <- table(test_c1$Species, y_pred)

cm


confusionMatrix(cm)
```

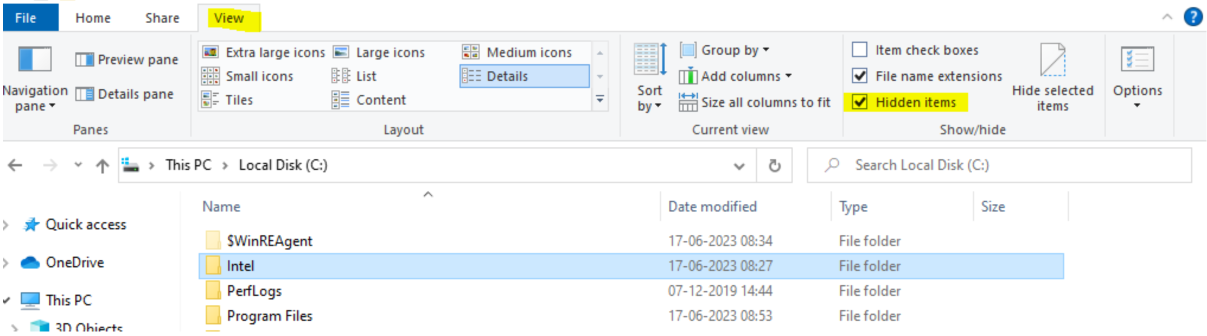
Output:



```
  virginica       0          1         19

Overall Statistics

               Accuracy : 0.9333
                 95% CI : (0.838, 0.9815)
    No Information Rate : 0.3667
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.9444           0.8636
Specificity                 1.0000            0.9286           0.9737
Pos Pred Value              1.0000            0.8500           0.9500
Neg Pred Value              1.0000            0.9750           0.9250
Prevalence                  0.3333            0.3000           0.3667
Detection Rate              0.3333            0.2833           0.3167
Detection Prevalence        0.3333            0.3333           0.3333
Balanced Accuracy           1.0000            0.9365           0.9187
> |
```

Install python package:

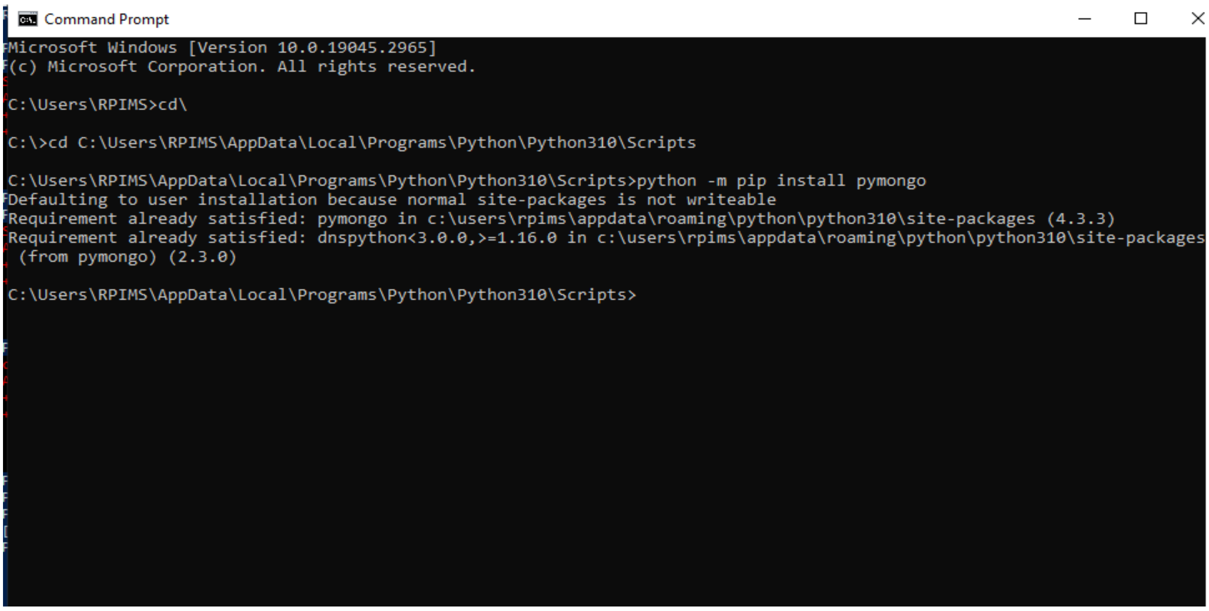
1. You will need to make the hidden folder visible: go to “C:” drive on top click on tab “view”
2. Select “hidden Items” option:



☐ Item check boxes
☑ File name extensions
☑ Hidden items

Name	Date modified	Type	Size
$WinREAgent	17-06-2023 08:34	File folder	
Intel	17-06-2023 08:27	File folder	
PerfLogs	07-12-2019 14:44	File folder	
Program Files	17-06-2023 08:53	File folder	

3. Go to the below path: C:\Users\Your Name\AppData\Local\Programs\Python\Python36-32\Scripts
4. Set the below path in command prompt and then use the below command: python -m pip install pymongo



```
Microsoft Windows [Version 10.0.19045.2965]
(c) Microsoft Corporation. All rights reserved.

C:\Users\RPIMS>cd\

C:\>cd C:\Users\RPIMS\AppData\Local\Programs\Python\Python310\Scripts

C:\Users\RPIMS\AppData\Local\Programs\Python\Python310\Scripts>python -m pip install pymongo
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pymongo in c:\users\rpims\appdata\roaming\python\python310\site-packages (4.3.3)
Requirement already satisfied: dnspython<3.0.0,>=1.16.0 in c:\users\rpims\appdata\roaming\python\python310\site-packages
 (from pymongo) (2.3.0)

C:\Users\RPIMS\AppData\Local\Programs\Python\Python310\Scripts>
```

Practical :4

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

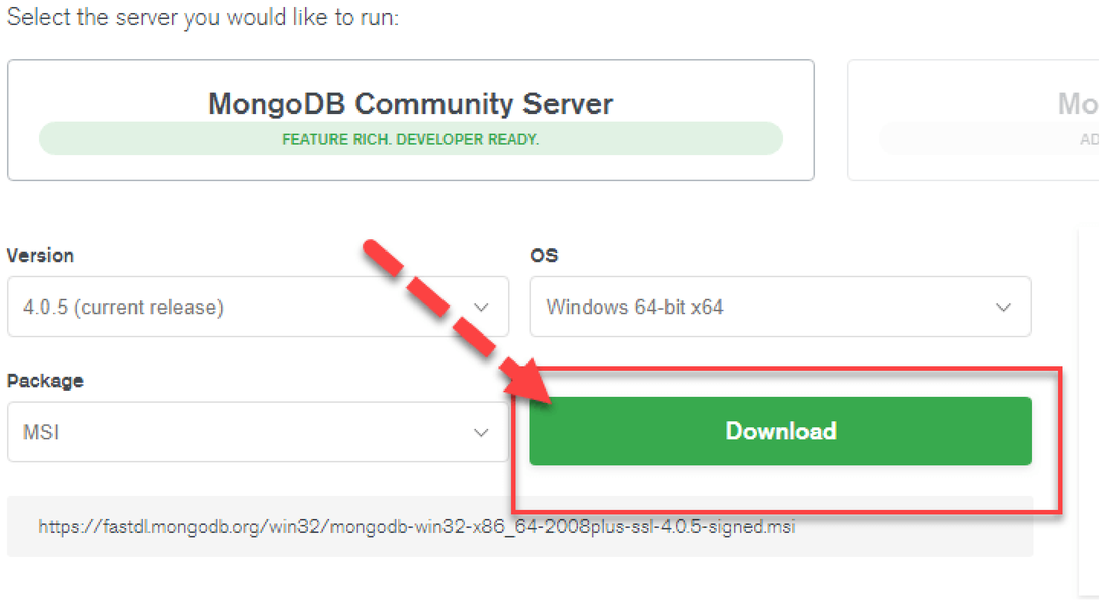
Requirement:

a. Python Package: PyMongo

b. Mongo Database

Step A: Install Mongo database

Step 1) Go to (https://www.mongodb.com/download-center/community) and Download MongoDB Community Server. We will install the 64-bit version for Windows.



Select the server you would like to run:

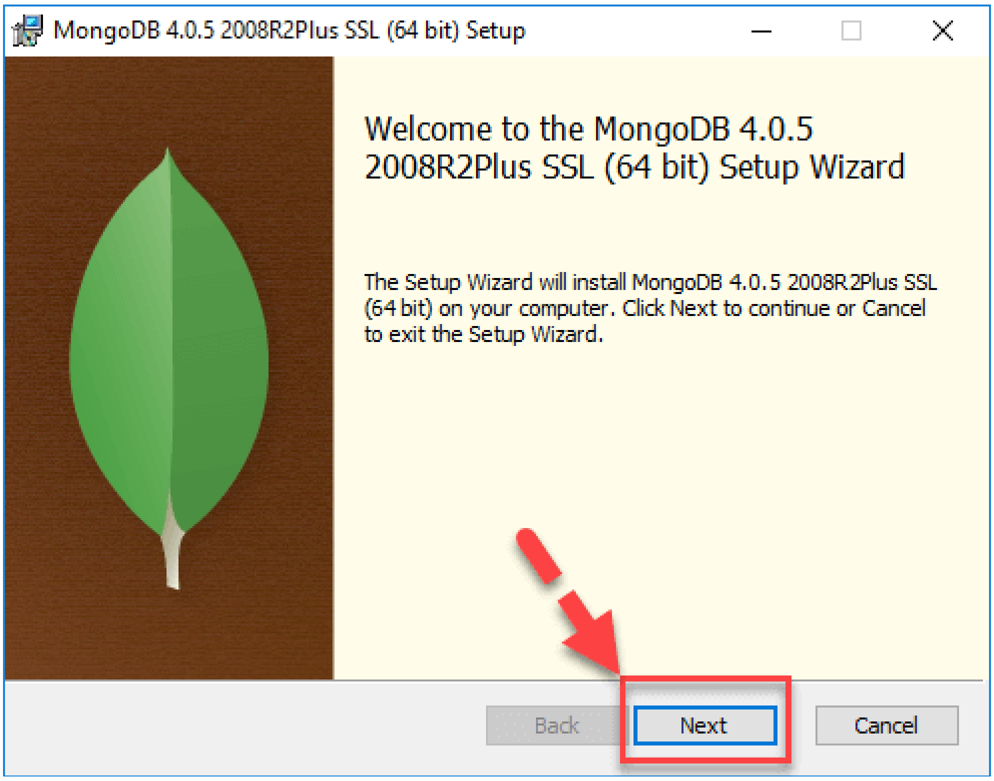
Version: 4.0.5 (current release)

OS: Windows 64-bit x64

Package: MSI

https://fastdl.mongodb.org/win32/mongodb-win32-x86_64-2008plus-ssl-4.0.5-signed.msi

Step 2) Once download is complete open the msi file. Click Next in the start up screen

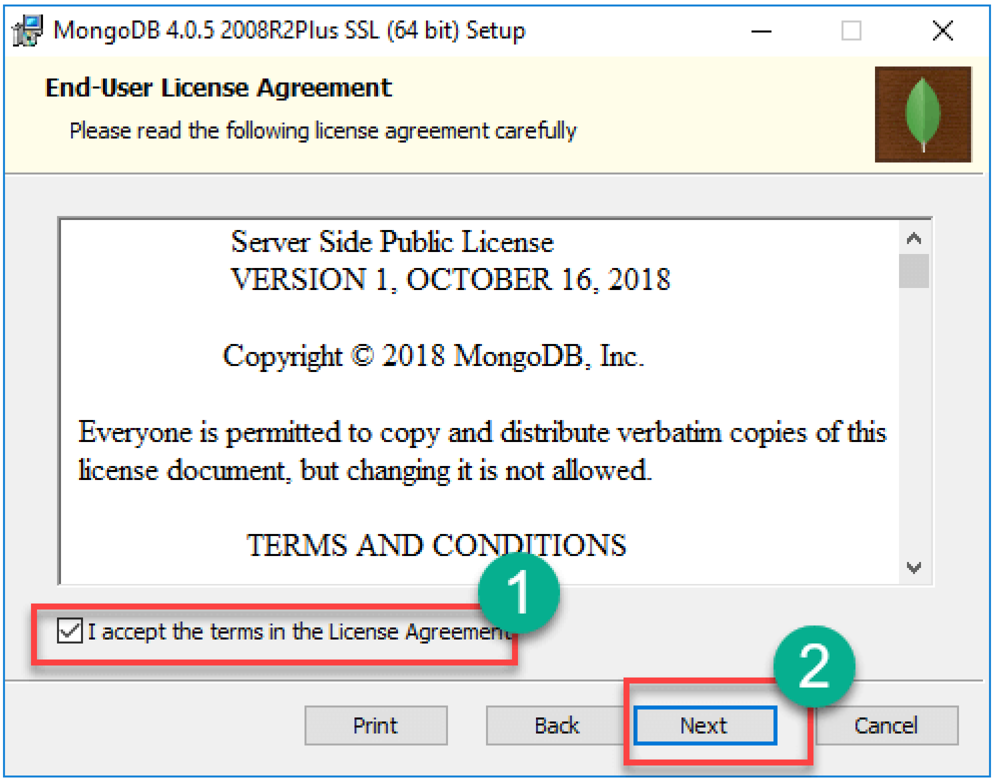


MongoDB 4.0.5 2008R2Plus SSL (64 bit) Setup

Welcome to the MongoDB 4.0.5 2008R2Plus SSL (64 bit) Setup Wizard

The Setup Wizard will install MongoDB 4.0.5 2008R2Plus SSL (64 bit) on your computer. Click Next to continue or Cancel to exit the Setup Wizard.

Step 3)
1. Accept the End-User License Agreement
2. Click Next



MongoDB 4.0.5 2008R2Plus SSL (64 bit) Setup

End-User License Agreement

Please read the following license agreement carefully

Server Side Public License
VERSION 1, OCTOBER 16, 2018

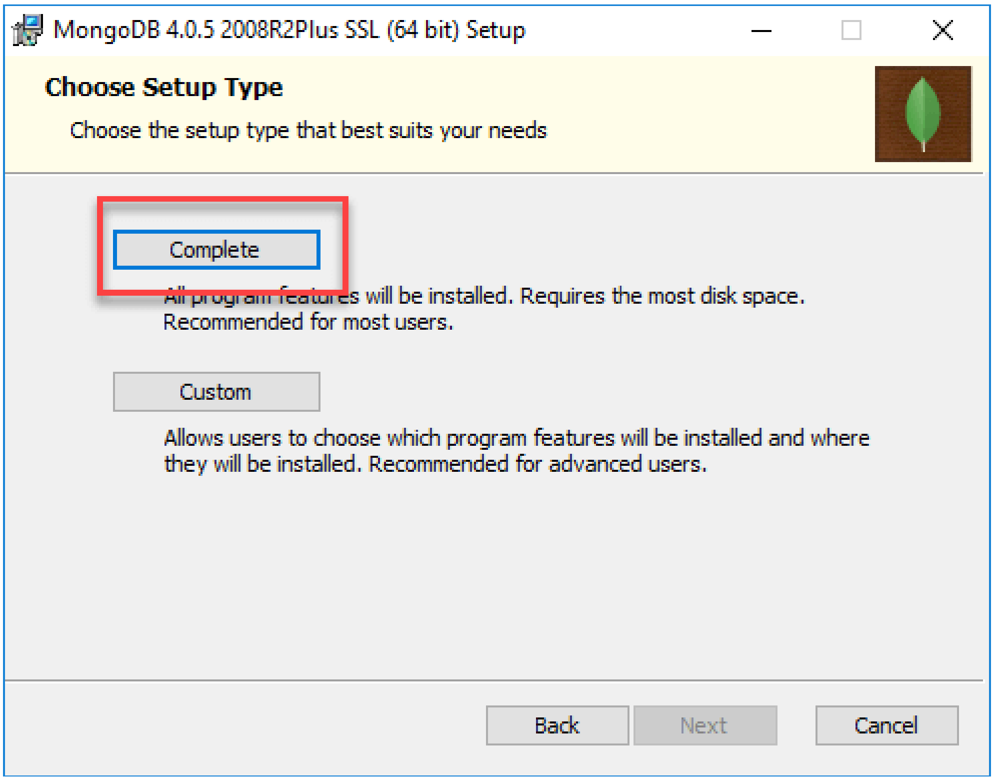
Copyright © 2018 MongoDB, Inc.

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

TERMS AND CONDITIONS

☑ I accept the terms in the License Agreement

Step 4) Click on the "complete" button to install all of the components. The custom option can be used to install selective components or if you want to change the location of the installation.



MongoDB 4.0.5 2008R2Plus SSL (64 bit) Setup

Choose Setup Type

Choose the setup type that best suits your needs

Complete

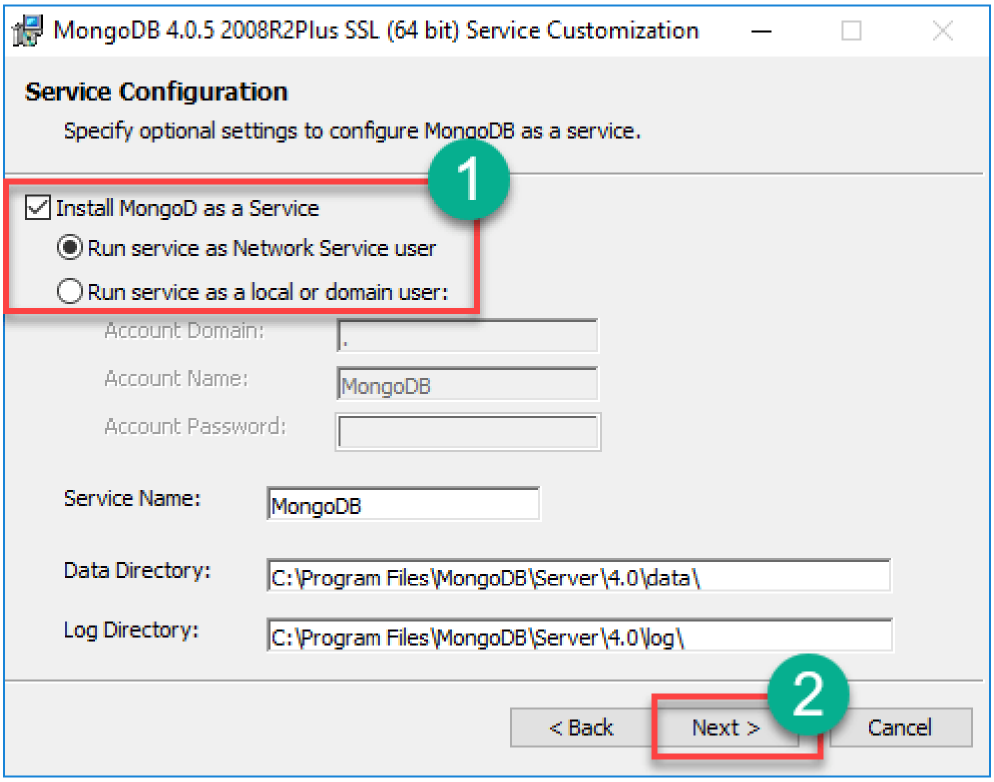
All program features will be installed. Requires the most disk space. Recommended for most users.

Custom

Allows users to choose which program features will be installed and where they will be installed. Recommended for advanced users.

Step 5)

1. Select “Run service as Network Service user”. make a note of the data directory, we’ll need this later.
2. Click Next



MongoDB 4.0.5 2008R2Plus SSL (64 bit) Service Customization

Service Configuration

Specify optional settings to configure MongoDB as a service.

☑ Install MongoD as a Service

● Run service as Network Service user

○ Run service as a local or domain user:

Account Domain: .

Account Name: MongoDB

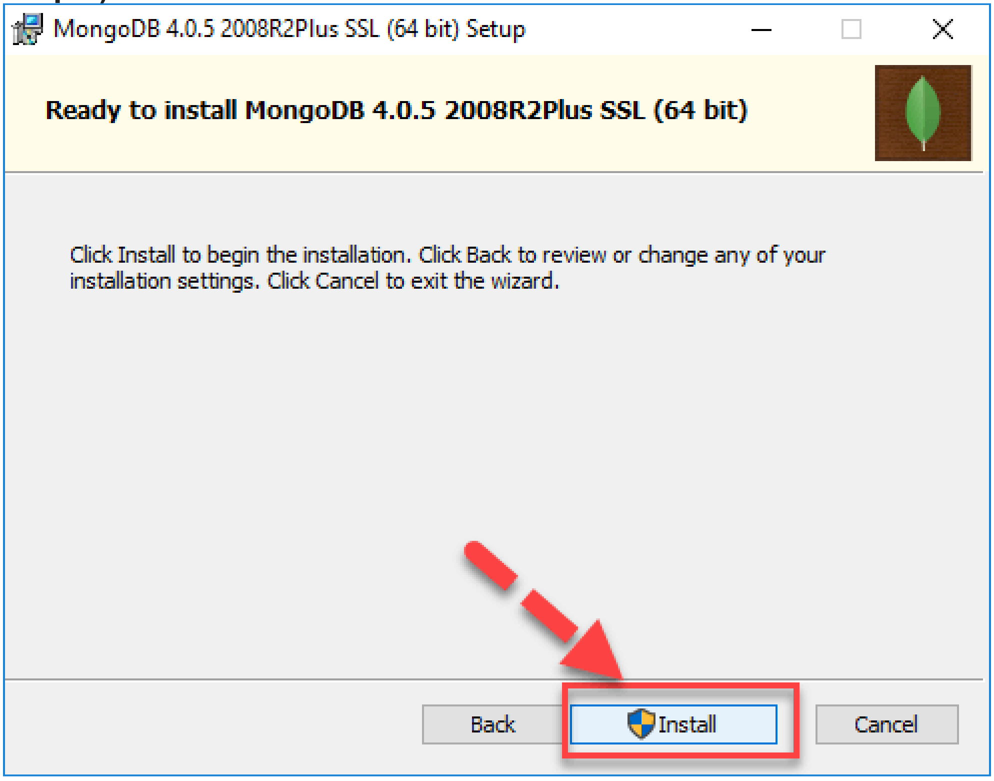
Account Password:

Service Name: MongoDB

Data Directory: C:\Program Files\MongoDB\Server\4.0\data\

Log Directory: C:\Program Files\MongoDB\Server\4.0\log\

Step 6) Click on the Install button to start the installation.

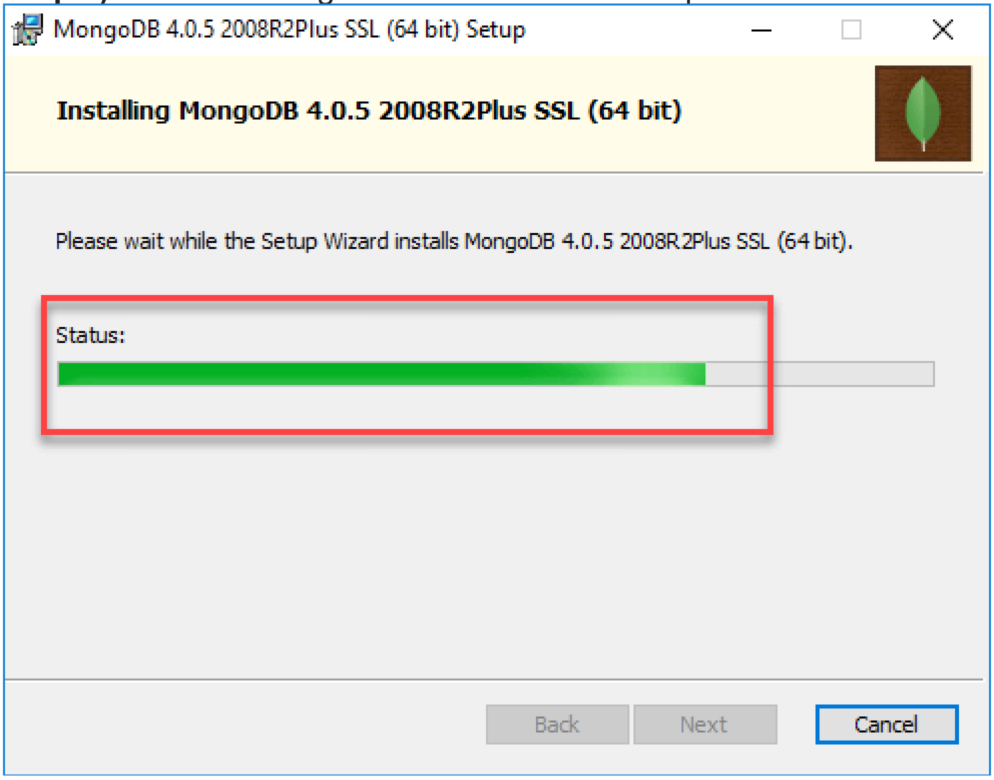


MongoDB 4.0.5 2008R2Plus SSL (64 bit) Setup

Ready to install MongoDB 4.0.5 2008R2Plus SSL (64 bit)

Click Install to begin the installation. Click Back to review or change any of your installation settings. Click Cancel to exit the wizard.

Step 7) Installation begins. Click Next once completed.



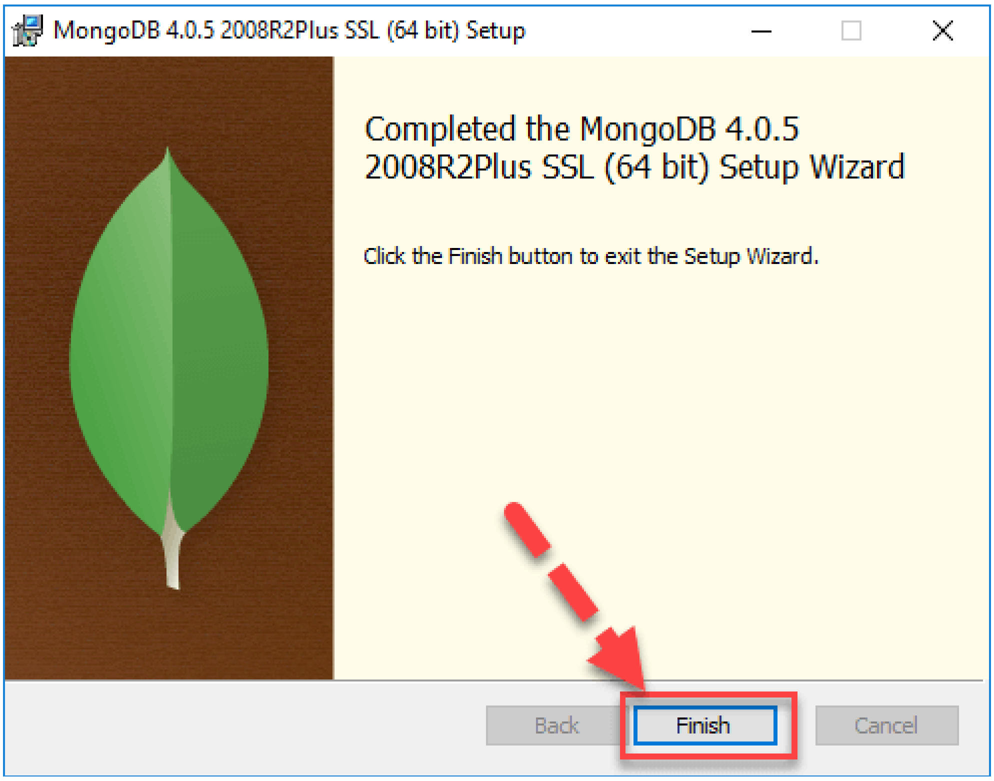
MongoDB 4.0.5 2008R2Plus SSL (64 bit) Setup

Installing MongoDB 4.0.5 2008R2Plus SSL (64 bit)

Please wait while the Setup Wizard installs MongoDB 4.0.5 2008R2Plus SSL (64 bit).

Status:

Step 8) Click on the Finish button to complete the installation



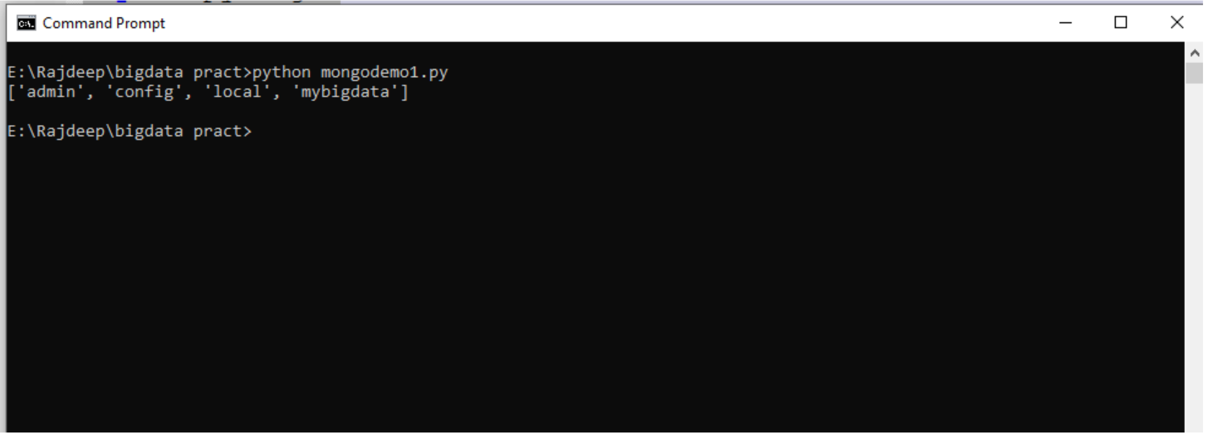
Completed the MongoDB 4.0.5 2008R2Plus SSL (64 bit) Setup Wizard

Click the Finish button to exit the Setup Wizard.

Program 1: Displaying the database name:

```
import pymongo
myclient = pymongo.MongoClient("mongodb://localhost:27017/")
mydb = myclient["mybigdata"]
print(myclient.list_database_names())
```

Output:

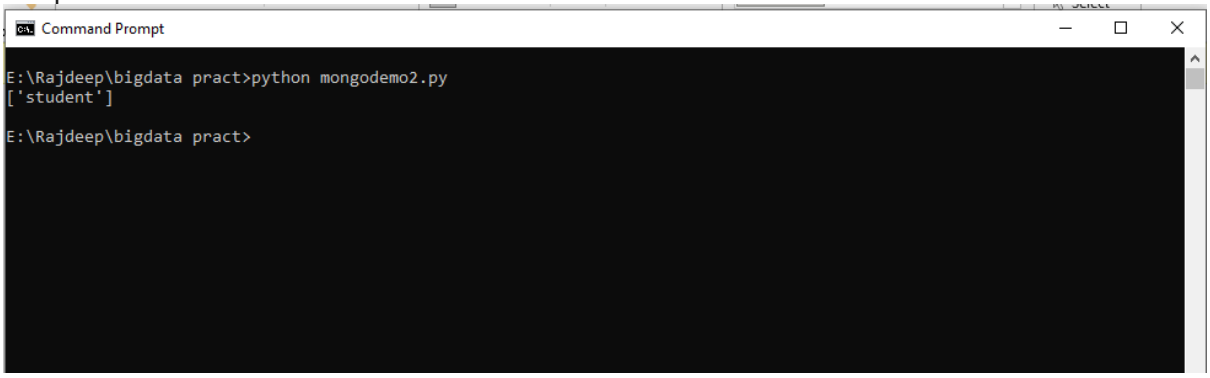


```
E:\Rajdeep\bigdata pract>python mongodemo1.py
['admin', 'config', 'local', 'mybigdata']

E:\Rajdeep\bigdata pract>
```

Program 2: Creating collection:

```
import pymongo
myclient = pymongo.MongoClient("mongodb://localhost:27017/")
mydb = myclient["mybigdata"]
mycol=mydb["student"]
print(mydb.list_collection_names())
```

Output:

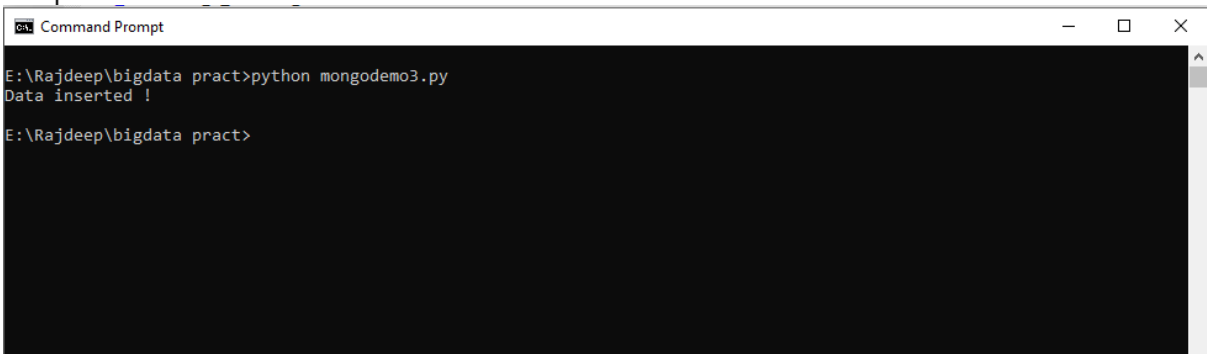


```
E:\Rajdeep\bigdata pract>python mongodemo2.py
['student']

E:\Rajdeep\bigdata pract>
```

Program 3: Inserting Data

```
import pymongo
myclient = pymongo.MongoClient("mongodb://localhost:27017/")
mydb = myclient["mybigdata"]
mycol=mydb["student"]
mydict={"name":"vai", "address":"bhy"}
x=mycol.insert_one(mydict)
print("Data inserted !")
```

Output:

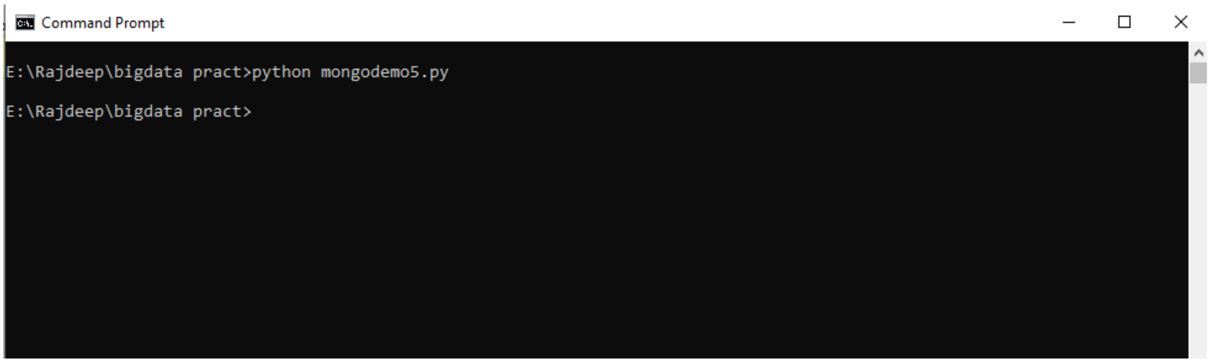


```
E:\Rajdeep\bigdata pract>python mongodemo3.py
Data inserted !

E:\Rajdeep\bigdata pract>
```

Program 4: Insert Multiple data into Collection

```
import pymongo
myclient = pymongo.MongoClient("mongodb://localhost:27017/")
mydb = myclient["mybigdata"]
mycol=mydb["student"]
mylist=[{"name":"Ganesh", "address":"Mumbai"}, {"name":"Varun",
"address":"Mumbai"},
{"name":"Prasoon", "address":"Pune"}, {"name":"Satish", "address":"Pune"},]
x=mycol.insert_many(mylist)
print("Data inserted !")
```

Output:

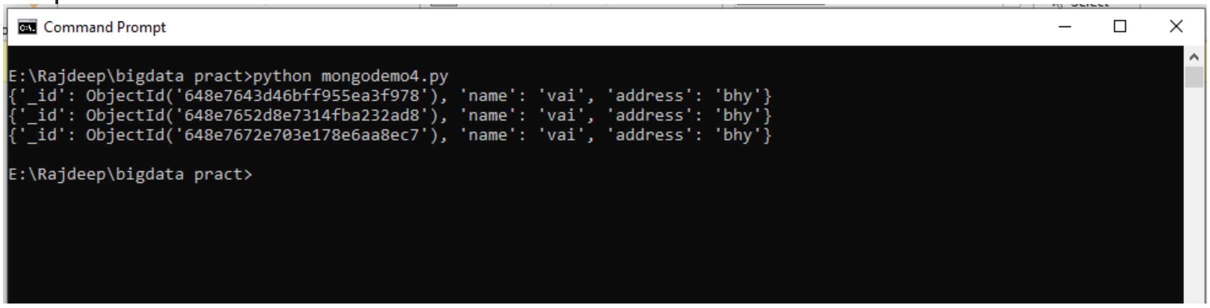


```
E:\Rajdeep\bigdata pract>python mongodemo5.py

E:\Rajdeep\bigdata pract>
```

Program 5: Displaying the collection data:

```
import pymongo
myclient = pymongo.MongoClient("mongodb://localhost:27017/")
mydb = myclient["mybigdata"]
mycol = mydb["student"]

myquery = { "name": "Vai" }

mydoc = mycol.find(myquery)

for x in mydoc:
  print(x)
```

Output:



```
E:\Rajdeep\bigdata pract>python mongodemo4.py
{'_id': ObjectId('648e7643d46bff955ea3f978'), 'name': 'vai', 'address': 'bhy'}
{'_id': ObjectId('648e7652d8e7314fba232ad8'), 'name': 'vai', 'address': 'bhy'}
{'_id': ObjectId('648e7672e703e178e6aa8ec7'), 'name': 'vai', 'address': 'bhy'}

E:\Rajdeep\bigdata pract>
```

Practical 5:

K means clustering.

Aim: Read a datafile grades_km_input.csv and apply k-means clustering.

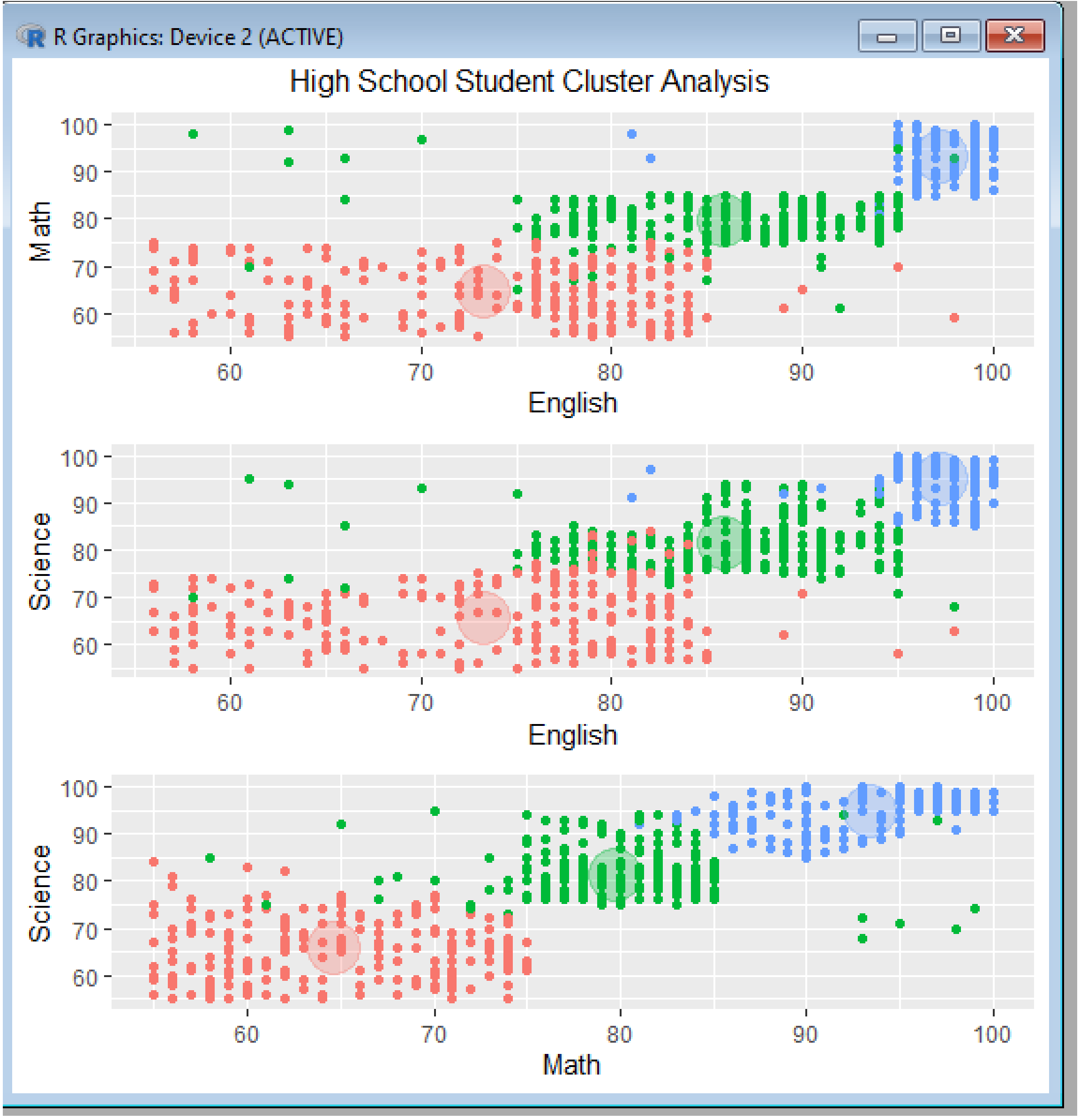
Requirement:

R tool

Code:

```
install.packages("plyr")
install.packages("ggplot2")
install.packages("cluster")
install.packages("lattice")
install.packages("grid")
install.packages("gridExtra")

library(plyr)
library(ggplot2)
library(cluster)
library(lattice)
library(grid)
library(gridExtra)

grade_input=as.data.frame(read.csv("E:/Rajdeep/bigdata
pract/dataset/grades_km_input.csv"))

kmdata_orig=as.matrix(grade_input[, c ("Student","English","Math","Science")])
kmdata=kmdata_orig[,2:4]
kmdata[1:10,]
wss=numeric(15)

for(k in 1:15)wss[k]=sum(kmeans(kmdata,centers=k,nstart=25)$withinss)
plot(1:15,wss,type="b",xlab="Number of Clusters",ylab="Within sum of square")
km = kmeans(kmdata,3,nstart=25)
km

c( wss[3] , sum(km$withinss))
df=as.data.frame(kmdata_orig[,2:4])
df$cluster=factor(km$cluster)
centers=as.data.frame(km$centers)

g1=ggplot(data=df, aes(x=English, y=Math, color=cluster )) +
geom_point() + theme(legend.position="right") +
geom_point(data=centers,aes(x=English,y=Math, color=as.factor(c(1,2,3))),size=10,
alpha=.3, show.legend =FALSE)

g2=ggplot(data=df, aes(x=English, y=Science, color=cluster )) +
geom_point () +geom_point(data=centers,aes(x=English,y=Science,
color=as.factor(c(1,2,3))),size=10, alpha=.3, show.legend=FALSE)

g3 = ggplot(data=df, aes(x=Math, y=Science, color=cluster )) +
geom_point () + geom_point(data=centers,aes(x=Math,y=Science,
color=as.factor(c(1,2,3))),size=10,  alpha=.3, show.legend=FALSE)

tmp=ggplot_gtable(ggplot_build(g1))
grid.arrange(arrangeGrob(g1  + theme(legend.position="none"),g2 +
theme(legend.position="none"),g3 + theme(legend.position="none"),top ="High
School Student Cluster Analysis" ,ncol=1))
```

Output:



Practical 6:

a. Simple Linear regression

Aim: Create your own data for years of experience and salary in lakhs and apply linear regression model to predict the salary

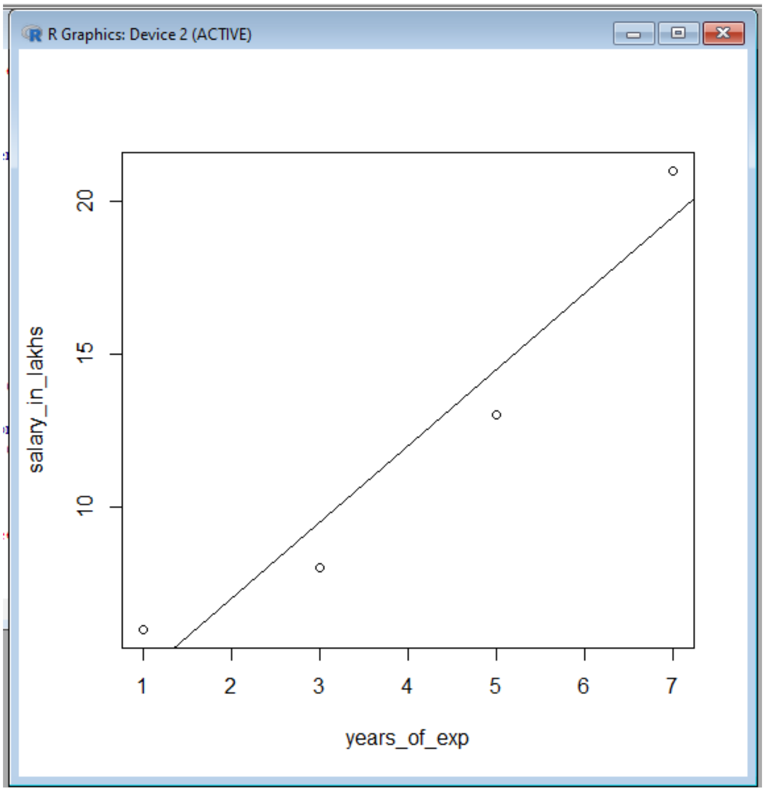
Requirement:

R tool

Code:

```
years_of_exp = c(7,5,1,3)
salary_in_lakhs = c(21,13,6,8)
employee.data = data.frame(years_of_exp, salary_in_lakhs)
employee.data

model <- lm(salary_in_lakhs ~ years_of_exp, data = employee.data)
summary(model)

plot(salary_in_lakhs ~ years_of_exp, data = employee.data)
abline(model)
```

Output:

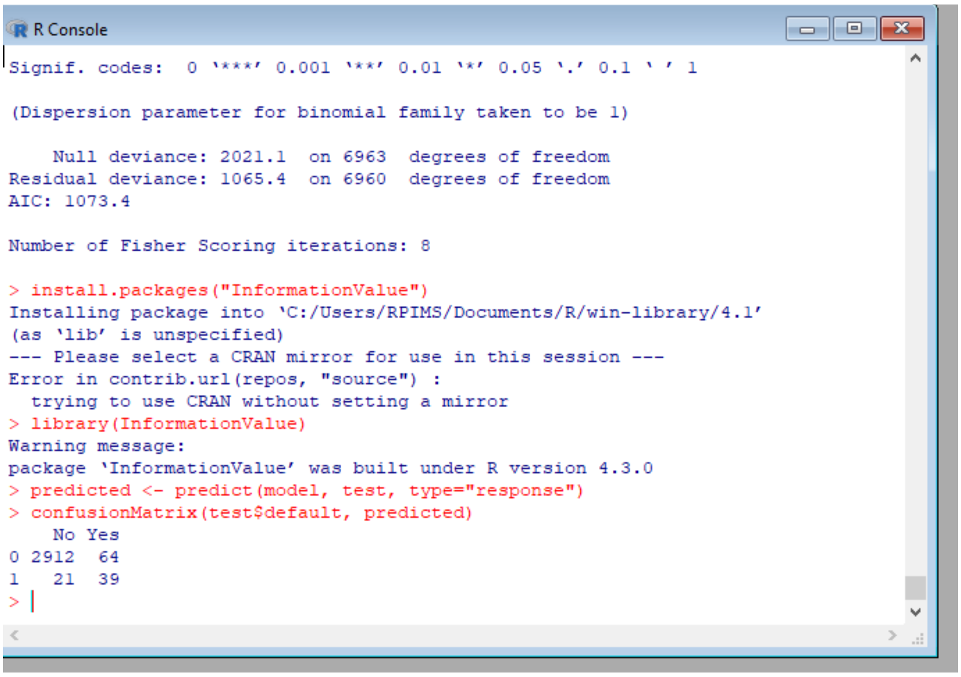


b.: Logistic regression:
Aim: Take the in-built data from ISLR package and apply generalized logistic regression to find whether a person would be defaulter or not; considering input as student, income and balance.

Code:

```
install.packages("ISLR")
library(ISLR)
data <- ISLR::Default
print (head(ISLR::Default))
summary(data)
nrow(data)
set.seed(1)
sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.7,0.3))
print (sample)
train <- data[sample, ]
test <- data[!sample, ]
nrow(train)
nrow(test)
model <- glm(default~student+balance+income, family="binomial", data=train)
summary(model)
install.packages("InformationValue")
library(InformationValue)
predicted <- predict(model, test, type="response")
confusionMatrix(test$default, predicted)
```

Output:



```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2021.1  on 6963  degrees of freedom
Residual deviance: 1065.4  on 6960  degrees of freedom
AIC: 1073.4

Number of Fisher Scoring iterations: 8

> install.packages("InformationValue")
Installing package into 'C:/Users/RPIMS/Documents/R/win-library/4.1'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
Error in contrib.url(repos, "source") :
  trying to use CRAN without setting a mirror
> library(InformationValue)
Warning message:
package 'InformationValue' was built under R version 4.3.0
> predicted <- predict(model, test, type="response")
> confusionMatrix(test$default, predicted)
    No Yes
0 2912  64
1   21  39
> 
```

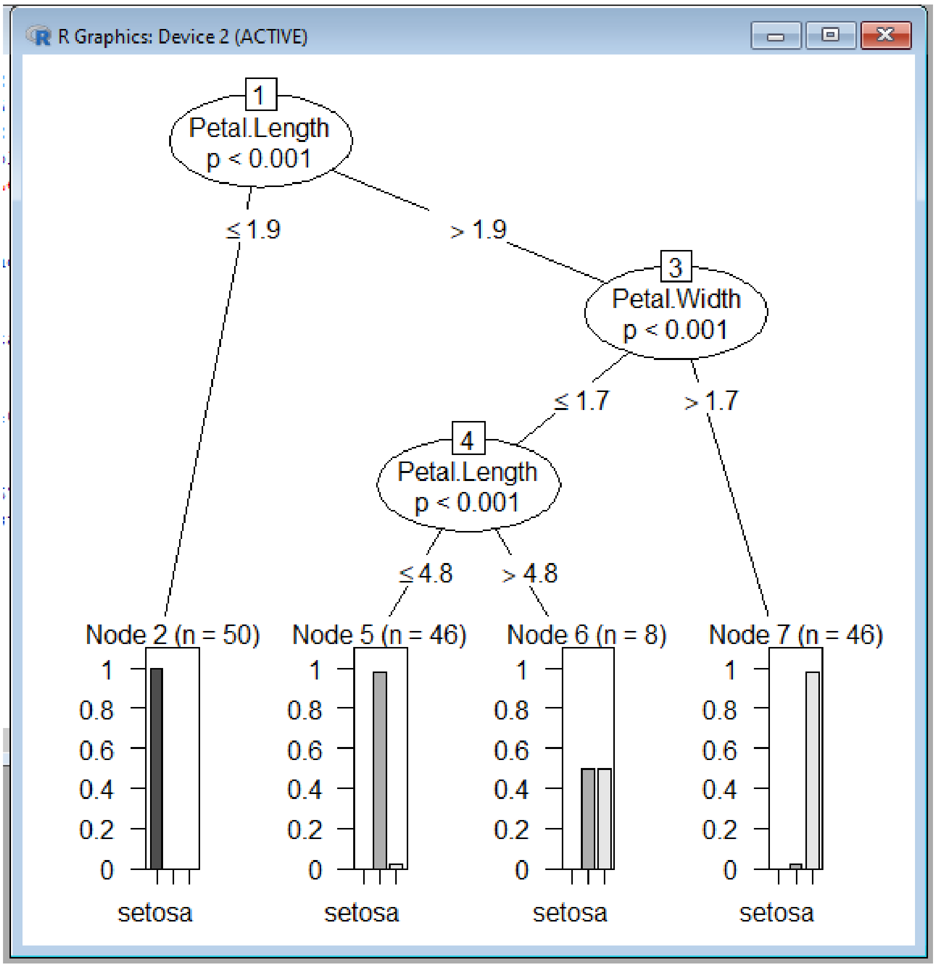
Practical 7:
Aim: Implement Decision tree classification techniques

Requirement:
R tool

Code:

```
library("party")
print(head(readingSkills))

str(iris)
iris_ctree <- ctree(Species ~ Sepal.Width + Sepal.Length + Petal.Length +
Petal.Width, data=iris

print (iris_ctree)
plot(iris_ctree)
```

Output:



<u>Practical 8</u>:

Apriori algorithm

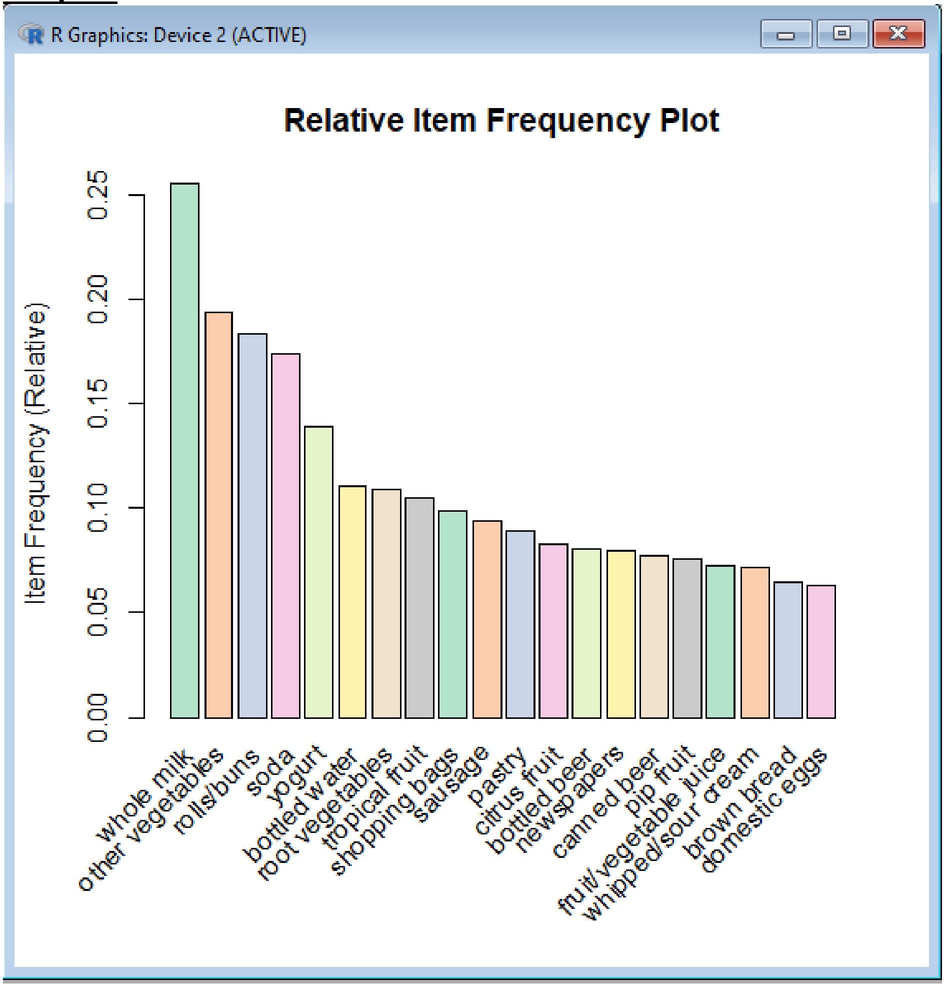
Aim: Perform Apriori algorithm using Groceries dataset from the R arules package.

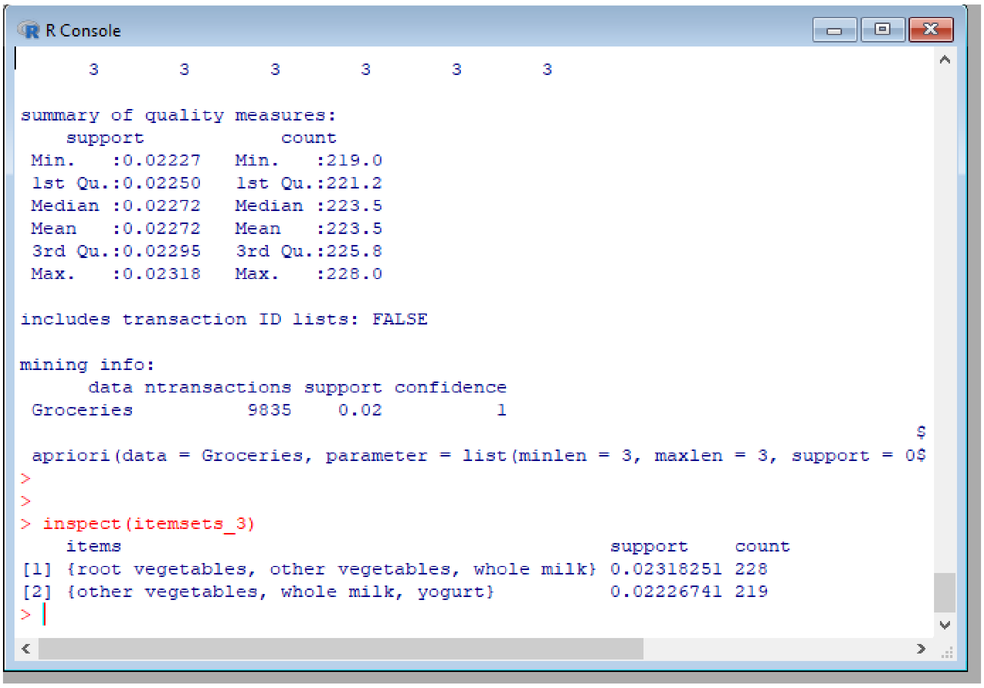
<u>Requirement:</u>
R tool

Code:

```
library(arules)
library(arulesViz)
library(RColorBrewer)

data(Groceries)
Groceries

summary(Groceries)
class(Groceries)

rules = apriori(Groceries, parameter = list(supp = 0.02, conf = 0.2))
summary (rules)

inspect(rules[1:10])

arules::itemFrequencyPlot(Groceries, topN = 20,
col = brewer.pal(8, 'Pastel2'),
main = 'Relative Item Frequency Plot',
type = "relative",
ylab = "Item Frequency (Relative)")

itemsets = apriori(Groceries, parameter = list(minlen=2, maxlen=2,support=0.02,
target="frequent itemsets"))
summary(itemsets)
inspect(itemsets)
itemsets_3 = apriori(Groceries, parameter = list(minlen=3, maxlen=3,support=0.02,
target="frequent itemsets"))
summary(itemsets_3)


inspect(itemsets_3)
```

Output:





```
         3         3         3         3         3         3

summary of quality measures:
    support            count
 Min.   :0.02227   Min.   :219.0
 1st Qu.:0.02250   1st Qu.:221.2
 Median :0.02272   Median :223.5
 Mean   :0.02272   Mean   :223.5
 3rd Qu.:0.02295   3rd Qu.:225.8
 Max.   :0.02318   Max.   :228.0

includes transaction ID lists: FALSE

mining info:
       data ntransactions support confidence
 Groceries           9835    0.02          1
                                                                                 $
 apriori(data = Groceries, parameter = list(minlen = 3, maxlen = 3, support = 0$
>
>
> inspect(itemsets_3)
    items                                        support    count
[1] {root vegetables, other vegetables, whole milk} 0.02318251 228
[2] {other vegetables, whole milk, yogurt}          0.02226741 219
> 
```