Journal of Computer Science and Applications.
ISSN 2231-1270 Volume 8, Number 1 (2016), pp. 21-29

http://www.irphouse.com

Data Analytics and Performance of Mobile Apps Using R Language

J.Uma Mahesh *M.Tech(C.S.E),(Ph.D)*
P. Srinivas Reddy *M.Tech(C.S.E)*
G Vijay Kumar *Ph.D*
Bharat Institute Of Engineering & Technology, KL University, Andhra Pradesh, India.

Abstract

Using data mining techniques it is the pracsstice of examining large pre-existing databases in order to generate new information. Suppose to build any classification model for predicting the data we perform some analysis in order to predict or classify new data. It is continuous with many applications to find out the performance analysis for various data sets. Here the main abstract of our paper is to Build Classification Model for Predicting Usage of [5] Mobile Apps Using R Language. R is a open source programming language or a data mining tool used for Statistical Computing of any data sets
By building a Classification Model i.e. J48 and other data mining techniques SVM,PCA used for classification, Prediction and regression analysis over the mobile apps Using R. The customer's usage ranking or priority while downloading various apps and customer focusing advanced features of the downloading app. The R language which results deep analysis in the form of any statistical computing it may be decision tree structure or some of statistical graphs. By comparing this result the user can be able to know the most widely used app according to their ranking. With different apps this particular apps provides higher flexibility, functionality and user friendly than the other apps.

Keywords: Data mining, R, Data classification, J48, SVM, PCA

I. INTRODUCTION TO R & RATTLE PACKAGE

R: R is a open source programming language[4] and software environment designed for statistical computing and graphics. The R language is broadly used among statisticians and data miners for rising statistical software and data analysis. Polls and surveys of data miners are shows that R's fame has increased significantly in recent years.

R is an interpreted language; users naturally access it through a command-line interpreter

R is an combined suite of software facilities for data manipulation, calculation and graphical display. It include

1. an effective data handling and storage facility,
2. a suite of operators for calculations on arrays, in specific matrices,
3. a large, intelligible, integrated collection of intermediate tools for data analysis,
4. graphical services for data analysis and present either softcopy or hardcopy,
5. a well-developed, simple and operational programming language which includes conditionals, loops, user-defined recursive function and input and output services.

Rattle package:

Rattle GUI[1] is a free and open source software(GNU GPL v2) package given that a graphical user interface (GUI) for data mining using the R statistical programming language. Rattle is used in a diversity of situations. Currently 15 various government departments in Australia and around the world use rattle in their data mining actions and as a statistical package.

Rattle provides considerable data mining functionality by exposing the power of the R Statistical Software through a graphical user interface. Rattle is also used as a training facility to learn the R software Language. There is a Log file Code tab, which replicates the R code for any activity undertaken in the GUI, which can be duplicated and pasted. Rattle can be used for statistical study, or model generation. Rattle allows for the dataset to be partitioned into training, validation and verification. The dataset can be viewed and abbreviated. There is also an option for scoring an external data file.

Features of Rattle package:

- File Inputs = CSV, TXT, Excel, ARFF, ODBC, R Dataset, RData File, Library Packages Datasets, Corpus, and Scripts.
- Statistics = Min, Max, Mean, Missing, Medium, Sum, Variance.
- Statistical tests = Correlation, T-Test, F-Test, and.
- Clustering = KMean, Hierarchical, and BiCluster.
- Modeling = Decision Trees, Support Vector, Principle Component
- Evaluation = Confusion Matrix, Risk Charts, Cost Curve, Hand, Lift, ROC, Precision, Sensitivity.
- Charts = Box Plot, Histogram, Correlations, Principal Components, Bar Plot, Dot Plot, and Mosaic.

Rattle also uses two external graphical investigation / plotting tools. Latticist and GGobi are self-governing applications which provide highly dynamic and interactive graphic data visualization for exploratory data analysis.

II. DATA MINING & CLASSIFICATION:

Data Mining [6] is an analytic process considered to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the result by applying the detected patterns to new subsets of data. In the data mining and statistics, classification is difficulty of identifying to which of a set of categories (sub-populations) a latest observation belongs, on the basis of a training set of data set objects is known.

III. CLASSIFIER ALGORITHMS

1. J48 Algorithm:

J48 classifier[6] is a simple C4.5 decision tree for classification. It generates a binary tree. The decision tree approach is most useful in classification problem. With this procedure, a tree is constructed to model the classification process. If Once the tree is built, it is applied to each tuple in the database

Algorithm [1] J48:

```
INPUT:          D //Training data
OUTPUT:         T //Decision tree
                DTBUILD (*D)
  {
  T=//φ;
  T= generate root node and label with splitting attribute;
  T=//Add arc to root node for each split  predicate and label; For each arc do
    D=// Database generated  by applying splitting predicate to D;
    If stopping point reached for this path, then T'= generate leaf node and label with
  proper class;
  Else
    T'= //DTBUILD(D);
    T=// add T' to arc;
  }
```

While constructing a tree, J48 ignores the lost values i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic design is to divide the data into range based on the attribute values for that item that are found in the training sample. This allows classification via either decision trees or rules generated from them.

2. SVM algorithm:

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, SVM used for classification and prediction analysis. Given a set of training examples, each marked as belongs to one of two categories, SVM builds a model that assigns new examples into one category or the other, building it a non-probabilistic binary linear classifier.SVM model is a representation of the

examples as points in space, mapped so that the examples of the various categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

3. PCA algorithm:

Principle component analysis is among the most popular tools in data mining, statistics, and data analysis more usually.PCA is basis of many techniques in data mining and informational retrieval, including the talent semantic analysis of large data bases of text and HTML document as described in PCA. Principal component analysis (PCA) is a technique that is useful for the compression and classification of data. The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of objects that nonetheless retains most of the sample's information. By this we mean the variation present in the sample, given by the correlations between the unique variables. The new variable, called principal components (PCs), are uncorrelated, and are ordered by the fraction of the total information each retains.

Related work: Predicting Stock Market Returns (Refer Data Mining with R Learning with Case Studies)

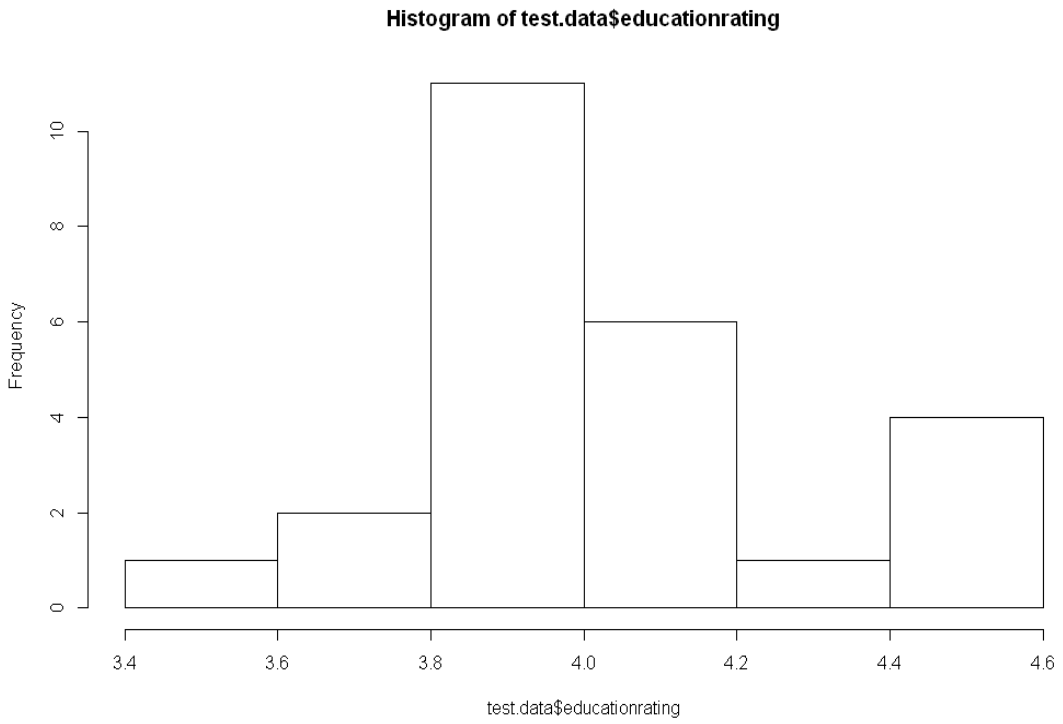
This second case study tries to move a bit further in terms of the use of data mining techniques. We will address some of the difficulties of incorporating data mining tools and techniques into a concrete business problem. The specific domain used to illustrate these problems is that of automatic stock trading systems. We will address the task of building a stock trading system based on prediction models obtained with daily stock quotes data. Several models will be tried with the goal of predicting the future returns of the S&P 500 market index. These predictions will be used together with a trading strategy to reach a decision regarding the market orders to generate. This chapter addresses several new data mining issues, among which are (1) how to use R to analyze data stored in a database, (2) how to handle prediction problems with a time ordering among data observations (also known as time series), and (3) an example of the difficulties of translating model predictions into decisions and actions in real-world applications. By this case study we got the idea to classify and Predicting the usage of Mobile Apps Using R Language through data mining techniques

IV. STATISTICS/ RESULTS ON USAGE OF MOBILE APPS USING R AND RATTLE[2] :

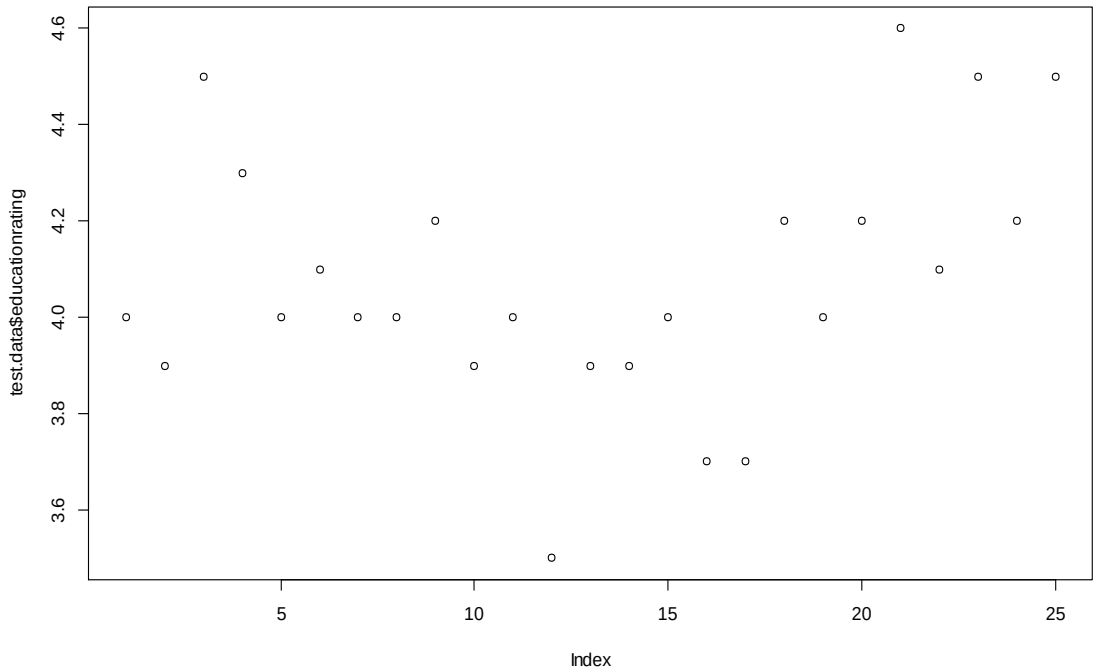
```
>educationnames<-c("Engineering math formulas","Engineering
apptitude","apptitude test and preparation","pocket
apptitude","apptitude cracker","ibps apptitude 2015","apptitude
interview questions","apptitude trianer","apptitude for
jobs","apptitude test preparation","iq test preparation","best iq
```

```
test","math iq","smart iq test","math iq challenge","logical iq
test","memory iq","general knowledge","general knowledge world
gk","general knowledge gk today","general knowledge quize","india
gk questions","crt 2015","campus friend","campus recruitment");

>educationdownload<-c("10 thousand","5 thousand","100
thousand","100 thousand","10 thousand","10 thousand","100
thousand","100 thousand","100 thousand","50 thousand","1
million","1 million","100 thousand","100 thousand","10
thousand","10 thousand","50 thousand","100 thousand","100
thousand","10 thousand","50 thousand","500 thousand","100","5
thousand","1 thousand");

>educationrating<-
c("4","3.9","4.5","4.3","4","4.1","4","4","4.2","3.9","4","3.5","
3.9","3.9","4","3.7","3.7","4.2","4","4.2","4.6","4.1","4.5","4.2
","4.5");

>educationsize<-
c("2.99mb","2.49mb","3.18mb","2.73mb","2.46mb","80kb","1.35mb","3
.20mb","1.80mb","2.99kb","1.34mb","2.85mb","1.05mb","4.01mb","1.8
4mb","6.71mb","0.94mb","4.02mb","6.69mb","4.42mb","3.14mb","1.39m
b","17.26mb","1.83mb","0.95mb");

>df1<data.frame(educationnames,educationdownload,educationrating,

educationsize)

>names(df1)<-c("educationnames char","educationdownload
char","educationrating int","educationsize char");

>write.csv(df1,"c:/mounika/mounika.csv”)

>df2<read.csv("c:/mounika/mounika.csv")

print(df2)

>test.data<- read.csv(file.choose())

test.data

>attach(test.data)
```

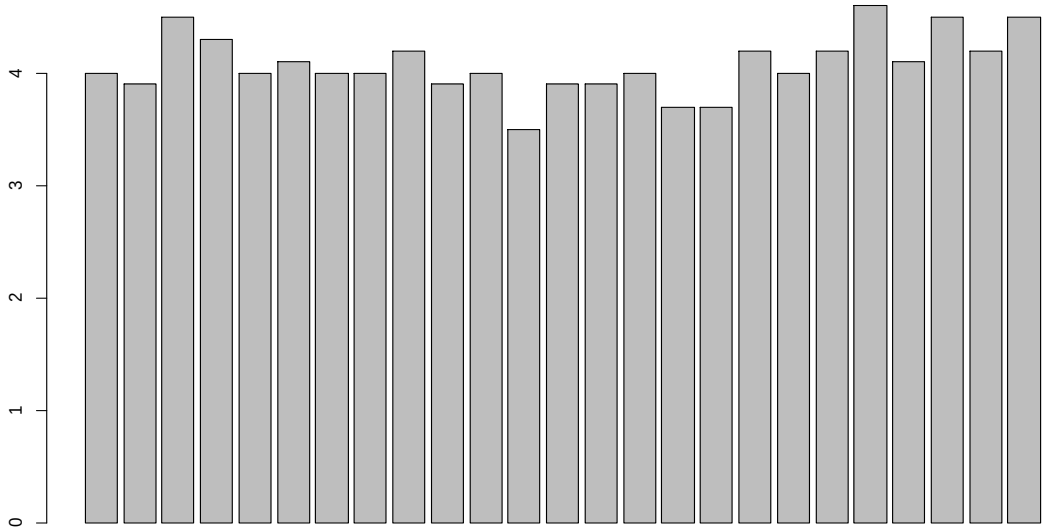
>hist(test.data$educationrating)



>plot(test.data$educationrating)

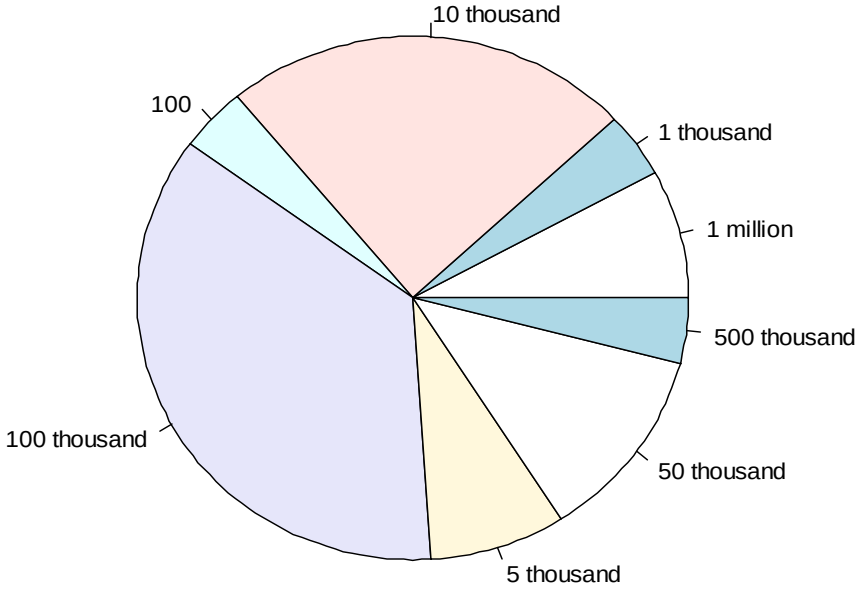


>barplot(test.data$educationrating)



>table(test.data$educationdownload)

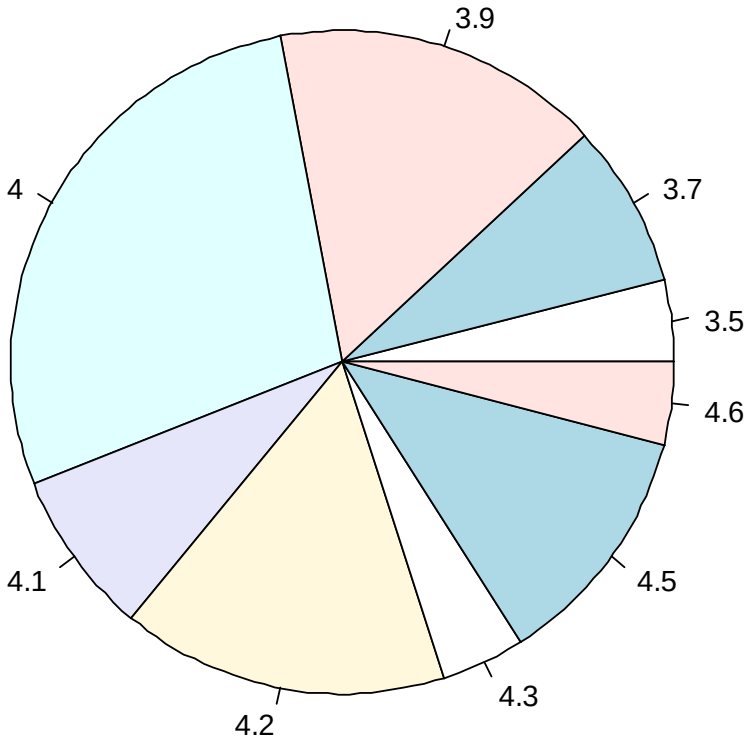
```
1million 1thousand 10thousand
     2       1          6

100   100thousand   5thousand
 1         9            2

 50thousand 500thousand
    3            1
```



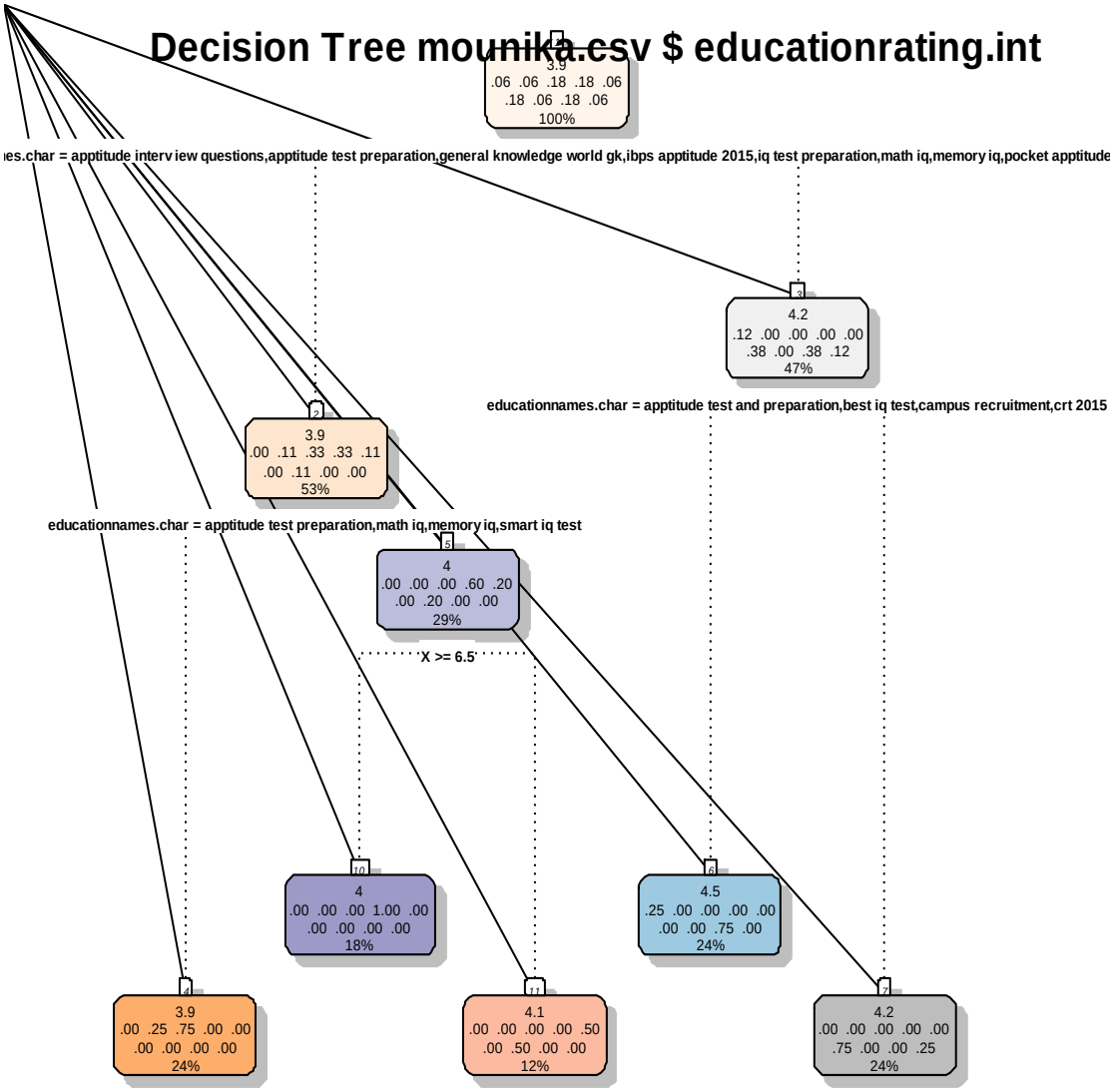
>table(test.data$educationrating)

3.5	3.7	3.9	4	4.1	4.2	4.3	4.5	4.6
1	2	4	7	2	4	1	3	1

>pie(table(test.data$educationrating))



Decision Tree :



Rattle 2015-Feb-28 11:09:51 UmaMahesh

REFERENCES

[1] URL http://CRAN.R-project.org/package=rattle

[2] Cook, D., Swayne, D. F., 2007. Interactive Dynamic Graphics for Data Analysis R. Springer-Verlag, New York.

[3] www.rdatamining.com

[4] www.r-project.org/

[5] en.wikipedia.org/wiki/Mobile_app

[6] Data Mining: Concepts and Techniques. Second Edition. Jiawei Hanand. Micheline Kamber.