National College of Ireland

Project Submission Sheet – 2016/2017

School of Computing

Student Name:	Dibyajyoti Bose		
Student ID:	15010732		
Programme:	M.Sc Data Analytics	**Year:**	2016-2017
Module:	Configuration Manual		
Lecturer:	Vikas Sahni		
Submission Due Date:	22/08/2016		
Project Title:	Classification of Seizure Disorder using Machine Learning		
Word Count:	816		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:

Date: 22/08/2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3 Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Chapter 1

Configuration Manual

1.1 SYSTEM SUMMARY

System summary section provides a general overview of the system been used. It also explains the hardware and software configuration used for performing the analysis.

1.1.1 System Configuration

All tools used for the analysis can be used on either a desktop or laptop devices. The below lists show the basic configuration used for this research.

Operating System: Windows 10
RAM: 8 GB
Hard Disc: 500 GB
Processor: i3 processor.

1.2 GETTING STARTED

Getting started section explains how to download and configure the tools which are used in this dissertation. A brief explanation on tool setup and workflow is given below.

1.2.1 TOOLS USED

Tools used section explains the number of tools used in this dissertation for the prediction analysis of medical data. The tools used are listed below.

- ✓ Microsoft Excel 2013
- ✓ R

1.3 Software overview

This research project consists of software tools and programming languages. R Version 3.3.1 statistical machine learning language was used to do the modelling and statistical analysis of the data. The basic R packages used in this project are neuralnet, randomForest, caret, class, ggplot2 which were installed from the CRAN library.

1.3.1 User Access Levels

Everyone can use the machine learning application since R is an open source tool. Users just need to install the mentioned libraries from CRAN in R studio and run the algorithms to get the results

.

1.3.2 Installation

The newest installation R version currently available can be downloaded from https://cran.r-project.org/bin/windows/base/ which should be installed on the device. For specific instruction on how to install application on specific device refer to CRAN installation guide.

1.3.3 System Menu

The original file received from NRS Medical Hospital, India was in excel format. The raw file was not usable to apply any machine learning models so feature engineering was done to extract meaningful insights from the data.

Age	Gender	Concomitant diseases	Urban/Rural	Onsetage	Impaired ADL	Frequency (/month)	Duration (minute)	last attack days back	Family history of Seizure	Febrile Seizure	Sleep deprivation	Drug/ alcohol abuse	EEG	CT Brain and MRI	Etiology of Sz	Type of Sz	Abnormal discharge
16	M	Mood Dis	Urban	14yrs 3 months	Y	0.75	0.75	25	N	N	Yes	N	focal Dysrhythmia	Rt focal isch change	HIE	Gen SZ	Slow Wave-40 Millisecond s
14	F	Headache	Urban	12yrs 5 months	N	1-2	15	170	N	N	No	N	Poly spike	Normal	Primary	Gen SZ	Multiple Spike- 35 to 45 MS
4	F	Nil	Urban	3yrs 2months	N	00-Jan	0.5	5	N	y	No	N	Gen Slow wave	Normal	Primary	Gen SZ	45 to 55 MS
27	M	Dep	Rural	24yrs 11 months	Y	0.25	15-20	11	N	N	Yes	N	Slowwave/spike	Normal	Primary	Gen Sz	Slow &Some Spike - 40 to 60 Ms
2	M	Nil	Rural	1yr 3 months	Y	2-3	**0.5-10**	7	N	y	No	N	sharp wave	Diffuse Ischae change	HIE	Gen Tonic Clonic	Gen sharp Wave-90 to 110 MS
30	M	Nil	Urban	28yrs 4 months	Y	0.5	10-15	40	N	N	No	N	Gen Slow wave	Normal	Primary	Gen Sz	Gen Slow Wave-30 to 55 MS
7	M	MCI	Rural	5yrs 11 months	Y	1-2	1-2	25	Y	Y	No	N	Poly Spike	Hypoxic Isch Change	HIE	Partial Sz(SPS) with Sec Gen	Multiple Spike- 30 to 45 MS

The attributes and medical terminologies used in the file has been consulted with domain experts (neurologists) and a short description is listed below for the user reference.

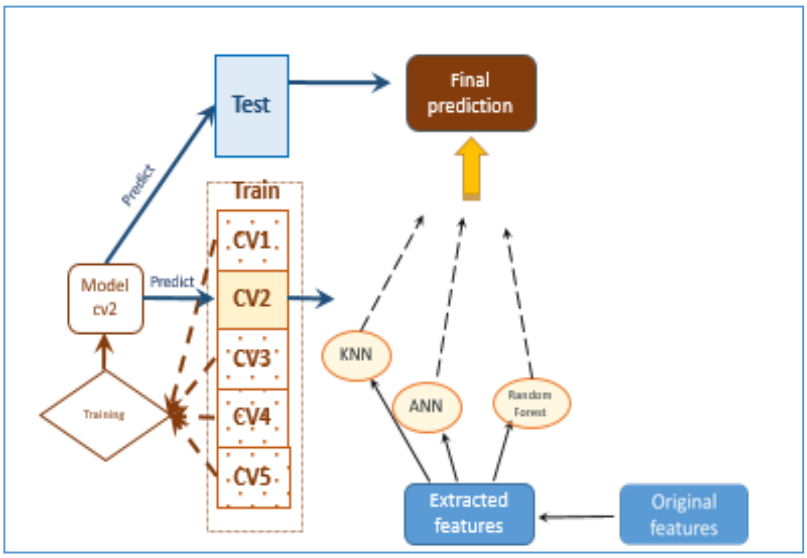
1. Abnormal Discharge in the Brain is detected by EEG(Electro Encephalography)
2. Sharp Wave - Transient, clearly distinguishable from background activity with pointed peak at conventional paper speeds and a duration of 70 milliseconds to 200 milliseconds (MS).Spike - Same as sharp wave but with duration of 20 MS to less than 70 MS. Poly Spike means- Multiple Spike discharge. Slow Wave - Classically the Slow wave is same as spike but of higher amplitude than the spike discharge.
3. A concomitant disease is a 2nd illness occurring at the same time as a primary illness. Associated Disease along with the Present Disease -eg Epilepsy along with mood disorder or Behavioural Disorder or Migraine etc.
4. Impaired ADL - Impairment of daily activity of any Individual
5. The onset of epilepsy is most common during early childhood and after age 60, but the condition can occur at any age
6. Family history- If you have a family history of epilepsy, you may be at an increased risk of developing a seizure disorder
7. A febrile seizure, also known as a fever fit or febrile convulsion, is a seizure associated with a high body temperature but without any serious underlying health issue. They most commonly occur in children between the ages of 6 months and 5 years

Seizure Classification - 1) Generalised Seizure involving whole of the Brain.
2) Partial (focal) Seizure - involving part of Brain only

The below feature engineering process was applied on the raw data.

Is_Mood	Is_Headache	Is_Depression	Is_Bipolar	Is_ADHD	Is_HypoThyroid	Is_MCI	Is_Urban	Onsetage	Impaired A	Frequenc	Min_Duration	Max_Duration	last attack days bac	Family hi	Febrile Si	Sleep deprivation	Alcohol a	Is_MRI_Normal	Is_Primar	Is_Slow_1	Is_Sharp	Is_Spike	Is_Discharge	Type of Sz
1	0	0	0	0	0	0	1	14	1	1	0	0.75	25	0	0	1	0	0	0	1	0	0	0	Gen
0	1	0	0	0	0	0	1	12	0	2	0	15	170	0	0	0	0	1	1	0	0	1	0	Gen
0	0	0	0	0	0	0	1	3	0	1	0	0.5	5	0	1	0	0	1	1	0	0	0	1	Gen
0	0	1	0	0	0	0	0	25	1	0	15	20	11	0	0	1	0	1	1	1	1	0	0	Gen
0	0	0	0	0	0	0	0	1	1	3	0.5	10	7	0	1	0	0	0	0	0	1	0	0	Gen
0	0	0	0	0	0	0	1	28	1	1	10	15	40	0	0	0	0	1	1	1	0	0	0	Gen
0	0	0	0	0	0	1	0	6	1	2	1	2	25	1	1	0	0	0	0	0	0	1	0	Partial
1	0	0	0	0	0	0	1	38.5	1	1	3	4	14	1	0	1	1	1	1	1	0	0	0	Gen
0	0	0	0	0	0	0	0	9	1	3	0	10	3	0	1	0	0	1	1	1	0	0	0	Gen
0	0	0	0	0	0	0	0	19	0	0	3	4	60	1	0	0	0	1	1	1	1	0	0	Gen
0	0	0	0	0	0	0	0	28	0	2	4	5	20	0	0	0	0	0	0	1	0	0	0	Partial
0	0	0	0	1	0	0	1	9	0	1	7	8	60	0	1	1	0	1	1	0	0	0	1	Gen
0	0	0	0	0	0	0	0	16	0	2	3	4	10	0	0	0	0	0	0	1	0	0	0	Partial
0	1	0	0	0	0	0	0	18	0	1	7	8	50	0	0	1	0	0	0	0	0	1	0	Partial
0	0	0	0	0	0	0	1	11	0	0	4	5	8	0	0	0	0	1	1	0	0	0	0	Gen
1	0	0	0	0	0	0	1	39	0	0	5	7	90	0	0	1	0	0	0	1	0	0	0	Gen
0	0	0	0	0	0	0	0	27	1	0	10	12	30	0	0	0	0	0	0	0	0	0	0	Partial
0	0	1	0	0	0	0	0	36	0	3	2	3	110	1	0	1	1	0	0	0	0	0	0	Partial
0	0	0	0	0	0	0	1	6	1	3	8	10	10	1	1	1	0	0	0	0	0	0	0	Partial

1.4 Design Workflow



USING THE MACHINE LEARNING ALGORITHMS

This section provides a detailed description of how the machine learning algorithm work and applied on the dataset.

1) K-nearest neighbour (KNN) method: the KNN model is performed in R with the class library. KNN requires the input data be normalized in range of values. The values of the 24 extracted features from the original data are scaled into the range between 0 and 1. The scaled data are then used for KNN modelling. Preliminary tests are performed to tune the model and find the optimal value of k=9 for this study, and the same k value is used for every run of KNN modelling for all the 5 folds cross validation.

2) Random forest (RF): RF is performed using the random forest package in R, with mtry=10 and ntrees=200 selected from preliminary tests. The featured engineer attributes are used as the input for this model. The variable importance from RF is plotted.

3) Artificial neural network (ANN): ANN is performed with the neural net package in R. It is good practice to normalize the data before training a neural network because depending on dataset some times the algorithm will not converge before the number of maximum iterations allowed. With the activation function as “sigmoid”, and one hidden layer with 5 neurons, learning rate 0.1, the ANN is trained on the normalized data.

R-Code

Artificial Neural Network (ANN) R code

```
install.packages("class")
library(ggplot2)
library(lubridate)
library(neuralnet)
library(caret)
```

```
library(class)
train <- read.csv("C:/Users/Dibyajyoti/Desktop/Thesis/Seizure_New.csv")
Org <- read.csv("C:/Users/Dibyajyoti/Desktop/Thesis/Seizure_New.csv")
m<-with(train, model.matrix(~ Type.of.Sz + 0))

set.seed(1)
apply(train,2,function(x) sum(is.na(x)))
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x))) }
train_nor <- as.data.frame(lapply(train[1:24], normalize))
train<-cbind(train_nor,m)
k<-5
i<-1
# K-Fold Cross Validation
n=floor(nrow(train)/k)
for(i in 1:k){
  s1=((i-1)*n+1)
  s2=(i*n)
  subset=s1:s2
  cv.train <- train[-subset,]
  cv.test <- train[subset,]
  cv.Org.train <- Org[-subset,]
  cv.Org.test <- Org[subset,]

 fit <-
neuralnet(Type.of.SzGen+Type.of.SzPartial~Is_Mood_Disorder+Is_Headache+Is_Depressio
n+Is_Bipolar+Is_ADHD+Is_HypoThyroid+Is_MCI+Is_Urban+Onsetage+Impaired.ADL+Freq
uency...month.+Min_Duration+Max_Duration+last.attack.days.back+Family.history.of.Seizur
e+Febrile.Seizure+Sleep.deprivation+Alcohol.abuse+Is_MRI_Normal+Is_Primary+Is_Slow_
Wave+Is_Sharp_Wave+Is_Spike_Wave+Is_Discharge,data=cv.train,hidden =
1,learningrate=0.1,lifesign="full")

}

mypredict <- compute(fit, cv.test[,1:24])$net.result
maxidx <- function(arr) {
  return(which(arr == max(arr)))
}
idx <- apply(mypredict, c(1), maxidx)
prediction <- c('Gen', 'Partial', 'Unclassified')[idx]
table(prediction,cv.Org.test$Type.of.Sz)

# Confusion matrix
confusionMatrix(prediction,cv.Org.test$Type.of.Sz)
```

K-Nearest Neighbour (kNN) R code

```
install.packages("class")
library(ggplot2)
library(lubridate)
library(caret)
library(class)
train <- read.csv("C:/Users/Dibyajyoti/Desktop/Thesis/Seizure_New.csv")
```

```
set.seed(1)
apply(train,2,function(x) sum(is.na(x)))
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x))) }
train_nor <- as.data.frame(lapply(train[1:24], normalize))
Seizure_Type<-train[,25]
train<-cbind(train_nor,Seizure_Type)
k<-5
i<-1
# K-Fold cross validation
n=floor(nrow(train)/k)
for(i in 1:k){
  s1=((i-1)*n+1)
  s2=(i*n)
  subset=s1:s2
  cv.train <- train[-subset,]
  cv.test <- train[subset,]
  prc_test_pred <- knn(train = cv.train[,1:24], test = cv.test[,1:24],cl = cv.train[,25], k=9)

}

# Confusion matrix
confusionMatrix(prc_test_pred, cv.test[,25])
}
```

Random Forest R code

```
install.packages("ROCR")
library(caret)
library(randomForest)
train <- read.csv("C:/Users/Dibyajyoti/Desktop/Thesis/Seizure_New.csv")
set.seed(1)
apply(train,2,function(x) sum(is.na(x)))


k.folds <- function(k) {
  folds <- createFolds(train, k = k, list = TRUE, returnTrain = TRUE)
  for (i in 1:k) {

    model <- randomForest(Type.of.Sz~.,
                          data = train[folds[[i]],], method = "class",savePredictions=TRUE)
    predictions <- predict(object = model, newdata = train[-folds[[i]],], type = "class")
    accuracies.dt <- c(accuracies.dt,
                       confusionMatrix(predictions, train[-folds[[i]], ]$Type.of.Sz))

    }
  accuracies.dt

}
set.seed(567)
accuracies.dt <- c()
accuracies.dt <- k.folds(5)
accuracies.dt
varImpPlot(model)
```