National College of Ireland

Configuration Manual

MSc Research Project
MSc in FinTech

Reena Pillay Rajagopalan
Student ID: 18186807

School of Computing
National College of Ireland

Supervisor: Victor del Rosal

National College of Ireland
MSc Project Submission Sheet
School of Computing

Student Name:	Reena Pillay Rajagopalan
Student ID:	18186807
Programme:	MSc in FinTech **Year:** 2019/2020
Module:	Research Project
Lecturer:	Victor del Rosal
Submission Due Date:	First submission – 17th August 2020 & Final submission – 28th September 2020
Project Title:	"A machine learning prediction-based analysis for the implementation of general practitioner E-health and Fintech services in Ireland."
Word Count:	938 **Page Count:** 4

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 required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ReenaPillay

Date: 28th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

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):	

Configuration Manual

Reena Pillay Rajagopalan
18186807

1 Introduction

This is the configuration manual to assist users to configure the research artefact titled *"A machine learning prediction-based analysis for the implementation of general practitioner E-health and Fintech services in Ireland"*. This manual will include the details on operating system, software and coding. The methodology used to complete this research project is CRISP-DM (Chapman, et.al, 2000).

2 Operating System

This research configuration was conducted on Windows 10 running on Intel(R) Core(TM) i7-4600U CPU @ 2.10 GHz 2.70 GHz processor. The installed memory (RAM) available for my DELL hardware is 8GB with 64-bit operating system processor.

3 RStudio

The programming software used to configure the research was RStudio. This software is an integrated development environment (IDE) that uses R programming language to conduct statistical, graphical and computing analyses. The RStudio version used is 1.2.5001 of the 2009-2019.

4 Data Collection & Coding

The data was obtained from Central Statistics Office (CSO/Census), Data.gov.ie and HSE. All the data narrows down to Dublin and County Meath and are for the year 2016 to suit the latest Census report.[1][2][3]

The packages used to conduct the machine learning techniques are as follows:

```
install.packages("dplyr","ggplot2","naivebayes","caret","psych")
library(dplyr)
library(ggplot2)
library(naivebayes)
library(caret)
library(psych)
```

[1] http://census2016.geohive.ie/datasets/population-by-general-health-sex-nuts-3-census-2016-theme-12-3-ireland-2016-cso-osi
[2] https://data.gov.ie/dataset/disability-and-carers-t12-ed
[3] https://www2.hse.ie/services/find-a-gp/

#----------LOADING & CHECKING DATA FILES----------#

In this section, the extracted data was saved on computer folder in a csv format and loaded on RStudio:

```
Population_CoDublinMeath<-read.csv
(file="CoDubMeathPopulation_Age_Tech_Labour_Census2016.csv",head=TRUE,sep=",",check.names=FALSE, na.strings = c("","#N/A"))

Disability_CoDublinMeath<-read.csv(file="Persons with a Disability_CoDCMH_Datagov2016.csv", head=TRUE,sep=",",check.names=FALSE, na.strings = c("","#N/A"))
```

Data was then checked for duplicated and missing values:

```
sum(duplicated(Population_CoDublinMeath))
sum(is.na(Population_CoDublinMeath))
sum(duplicated(Disability_CoDublinMeath))
sum(is.na(Disability_CoDublinMeath))
```

#----------DATA PREPARATION----------#

In this section, the following function was used to narrow down useful variables to this study and filter out those variables to put them into a new dataset. Some column headings were changed.

```
#Taking a glimpse into the data to help identify useful variables
glimpse(Population_CoDublinMeath)

#Summing up columns to croasscheck superfluous variables
#Columns associated to Smartgadgets
A<-colSums(Population_CoDublinMeath[,c("T6_8_OO","T6_8_TAA","T6_8_UHHH","T6_8_OVDD")])
sum(A)
#Sum of columns equal to total sum of SmartGadget column
sum(Population_CoDublinMeath$SmartGadget_Owners_T6_8_T)

#Summing columns associated with Broadband
B<-colSums(Population_CoDublinMeath[,c("T15_3_B","T15_3_OTH","T15_3_N","T15_3_NS")])
sum(B)
#Sum of columns equal to total sum of Broadband column
sum(Population_CoDublinMeath$Broadband_T15_3_T)

#Selecting Variables (based on personal hunch to test) and renaming columns header
Pop_Subset <- Population_CoDublinMeath %>% select(COUNTY, COUNTYNAME, EDNAME, T1_2T, SmartGadget_Owners_T6_8_T, Broadband_T15_3_T,STATISTIC, Lab_Age_Group, C02199V02655_Sex, LABOUR_PARTCP_STAT, LABVALUE_PERC, UNEMPLY_STAT, UNEMPLY_VALUEPERC,Age_classification_Gpvisit, gender_gpvisit, total_visit, area)

# Change column 3 and 4 names
colnames(Pop_Subset)[3:4] <- c("Town", "Population")
```

A density plot is done to understand the spread of the following variables to identifying the county with higher population, smart gadget users and broadband users.

```
#Density plot to understand the 3 variables based on County
Pop_Subset %>% ggplot(aes(x= Population, fill = COUNTYNAME)) + stat_density(alpha=2, color= 'black') + ggtitle("Density Population Plot")
```

```
sum(Pop_Subset$Population)

Pop_Subset %>% ggplot(aes(x=SmartGadget_Owners_T6_8_T, fill = COUNTYNAME)) + stat_density(alpha=0.8, color= 'black') + ggtitle("Density Smart Gadget Plot")
sum(Pop_Subset$SmartGadget_Owners_T6_8_T)

Pop_Subset %>% ggplot(aes(x=Broadband_T15_3_T, fill = COUNTYNAME)) + stat_density(alpha=0.8, color= 'black') + ggtitle("Density Broadband Plot")
sum(Pop_Subset$Broadband_T15_3_T)
```

#----------CORRELATION----------#

To understand the relationship between variables, the following correlation function was used. This provides insights on the significance level of the variable along with the impact on each other if any changes occurred (James, et.al, 2013)

```
# To identify the relation between Population & GP Visitation
# Spearman Method
cor.test(Pop_Subset$Population, Pop_Subset$total_visit, method='spearman')

# Pearson Method between Population and Unemployment
cor.test(Pop_Subset$Population, Pop_Subset$UNEMPLY_VALUEPERC, method='pearson')

# Identify a relation between Age and GP visit
# Spearman Method
str(Pop_Subset) #to obtain the classification of each variable whether it is factor or numerical or etc.
Agevisit<-Pop_Subset$Age_classification_Gpvisit
#to change from factor to numerical
Pop_Subset$Age_classification_Gpvisit<-as. numeric(Pop_Subset$Age_classification_Gpvisit)
cor.test(Pop_Subset$Age_classification_Gpvisit, Pop_Subset$total_visit,method= "spearman", exact = FALSE)

# Change back the numeric to factor for Age
Pop_Subset$Age_classification_Gpvisit<-Agevisit
str(Pop_Subset)

# Correlation Visualisation Summary based on the above
pairs.panels(Pop_Subset[-1]) #plots out all the correlation values for all columns in a combined chart
```

#----------NAIVE BAYES----------#

In this section, classification Naïve Bayes model was used to make categorical analysis and obtain a predictive accuracy of the dataset (Dietrich, 2015).

```
# Data Partition splitting data to 70% training data and 30% testing data
set.seed(1234)
index <- sample(1:nrow(Pop_Subset),size=nrow(Pop_Subset)*0.70, replace=FALSE)
train_popsub1<- Pop_Subset[index,]
test_popsub1 <- Pop_Subset[-index,]


# Naive Bayes Model of population techsavvyness based on county
model<-naive_bayes(COUNTYNAME~Population+SmartGadget_Owners_T6_8_T+ Broadband_T15_3_T, data = train_popsub1, usekernel = T)
model #gives the accuracy and statistical attributes results
plot(model) #graphically presents the results

#Repeat the function for other selected variables
# Naive Bayes Model of unemployed age group based on county
```

```
modelgp <- naive_bayes(COUNTYNAME~Age_classification_Gpvisit+UNEMPLY_VALUEPERC, data = train_popsub1, usekernel = T)
modelgp
plot(modelgp)

# Naive Bayes Model of same gender class from gp visits owning smart gadget based on county
modelgendtech <- naive_bayes(COUNTYNAME~gender_gpvisit+SmartGadget_Owners_T6_8_T, data = train_popsub1, usekernel = T)
modelgendtech
plot(modelgendtech)
```

Check performance of the naïve bayes model to show the accuracy of the results and Kappa value by using Confusion Matrix function.

```
p1<-predict(modelgp,test_popsub1)
(tab1<-table(p1,test_popsub1$COUNTYNAME))
# Based on the unemployed age group model, 230 were predicted to the wrong county
1-sum(diag(tab1))/sum(tab1)
confusionMatrix(p1, reference = test_popsub1$COUNTYNAME)
```

#----------**DATA VISUALISATION**----------#

GGplot package was used here to visually present the data based on guidance by James (2013).

```
# Visualisation on total people who have disability in both counties
ggplot(Disability_CoDublinMeath, aes(x= C03367V04052, y= VALUE)) + geom_bar(aes(fill= C03367V04052), stat="identity", colour="black",position=position_dodge())

#Total visitation of people to GP in both counties
ggplot(Pop_Subset, aes(x= COUNTY, y= total_visit)) + geom_bar(aes(fill= COUNTY), stat="identity", colour="black",position=position_dodge())

#Scatter plot on the population in both counties
ggplot(Pop_Subset, aes(COUNTY, Population)) +geom_point() + geom_point(data = Pop_Subset, aes(y = Population), colour = 'red', size = 3)
```

#----------#

References

Chapman, P., Clinton, J. M., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. R. H. and Wirth, R. (2000). ‘Crisp-dm 1.0: Step-by-step data mining guide’.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) ‘An introduction to statistical learning’, *New York: springer*,112, p. 18.

Dietrich, D. (2015) ‘Data science and big data analytics: Discovering, analyzing, visualizing and presenting data’, *John Wiley & Sons*.