Configuration Manual

Identification and Prediction of Factors Impact America Health Insurance Premium

Jun Jun Sun Student ID: X17162238

School of Computing National College of Ireland

Supervisor: Dr. Catherine Mulwa



National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Jun Jun Sun
Student ID:	X17162238
Programme:	Data Analytics Year:2020
Module:	MSc Data Analytics Research Project
Lecturer:	Dr. Catherine Mulwa
Programme:Data AnalyticsYear:2020Module:MSc Data Analytics Research ProjectYear:2020Lecturer:Dr. Catherine MulwaSubmissionDr. Catherine MulwaDue Date:17/08/2020Identification and Prediction of Factors Impact America Health Insurance Premium	
Project Title:	Identification and Prediction of Factors Impact America Health Insurance Premium
Word Count:	2107 Page Count: 32

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.

<u>ALL</u> 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:	Linduan
Date:	17/08/2020

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

Jun Jun Sun Student ID: x17162238

1 Introduction

The configuration manual determines from the begin setup stage till the result of the whole process. The purpose of this project is to define a best performance model to achieve the objective and address the research question. There are several models generated and conducted to identify the result. This configuration manual includes an explanation of hardware and software properties and installation used, with an implementation of each step of the process which includes data preparation, model code generated and results.

The structures of this configuration manual report are as follows:

Chapter 2: Discover the environment specification and configuration

Chapter 3: Explain the data preparation

Chapter 4: Discover model implementation and steps of each output generated

Chapter 5: Walk-through appendix

2 Environment Specification and Configuration

The environment specification and configuration deliver specifics of what systems are required to develop and implemented for this project, hardware and software are the key elements to implement this project. This chapter majority is to discover the integrated configuration environments that were used.

2.1 Hardware Configurations

This section will discuss the details of hardware, figure 1 shows the machine used for the implementation of this project. Windows 10 systems running on a Lenovo laptop named LAPTOP-3HR9A551 with 64-bit operating system, 2.7GHz processor and 8GM RAM was used.

Windows edition		
Windows 10 Home		
© 2019 Microsoft Corpora	tion. All rights reserved.	
-		
System		
Processor:	Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.70 GHz	
Installed memory (RAM):	8.00 GB (7.85 GB usable)	Lenovo
System type:	64-bit Operating System, x64-based processor	
Pen and Touch:	No Pen or Touch Input is available for this Display	Support Information
Computer name, domain, and	workgroup settings	
Computer name:	LAPTOP-3HR9A551	Change settings
Full computer name:	LAPTOP-3HR9A551	
Computer description:		
Workgroup:	WORKGROUP	
Windows activation		
Windows is activated Rea	d the Microsoft Software License Terms	
Product ID: 00325-95800-	0000-AAOEM	Change product key

Figure 1 Hardware Configuration

2.2 Software Configurations

This section will discuss the details of software implementation and installation.

2.2.1 RStudio

The RStudio was downloaded from here¹.

The RStudio used version 1.3.1056 to implement for this project shows in figure 2.



Figure 2 RStudio Version Used for Implementation

The RStudio properties (Figure 3) shown it has been created on 10/04/2018 from the following computer name LAPTOP-3HR9A551.

General	Shortcut Details	Com	patibility Versions
Security	D Ottalio	Frevious	v 81510/15
ed: Just now	Value		
File			
Name	RStudio.lnk		
Туре	Shortcut		
Folder path	C:\Users\junju\AppData	a\Roaming\Micro	soft\I
Size	948 bytes		
Date created	10/04/2018 23:24		
Date modified	11/03/2018 23:39		
Attributes	Α		
Owner	LAPTOP-3HR9A551\ju	nju	
Computer	LAPTOP-3HR9A551 (t	his PC)	
_	ties and Personal Inform	nation	

Figure 3 RStudio Properties

¹ https://rstudio.com/products/rstudio/download/

2.2.2 Tableau

The Tableau was downloaded from here².

The version of Tableau Desktop professional edition 2020 was used for this project shown in figure 4.

🍲 Tableau 2020.2 (20202.20.0721.	1350) Setup		×
Tableau Desktop	Tableau 2020.2.4		
	Progress		
	Installing: Tableau 2020.2 (20202.20.0721.1350)	Cancel	

Figure 4 Tableau Version

The Tableau was created on 22/10/2019, which is also installed on this computer name LAPTOP-3HR9A551 shown at left in Figure 5. But the error appeared when I tried to modify graphs it shown update request that the Tableau, therefore, the update version 2020 was installed for modification details as figure 5 right.

General	Shortcut	Compatibility	General	Shortcut	Compatibility
Security	Details	Previous Versions	Security	Details	Previous Versions
Property File	Value		Property File	Value	
Type Folder path Size	Shortcut C:\Users\junju\AppData\ 1.30 KB	Roaming\Microsoft\I	Type Folder path Size	Shortcut C:\ProgramData\Microso 1.30 KB	ft\Windows\Start Me
Date created Date modified	22/10/2019 17:50 22/10/2019 17:47		Date created Date modified	08/08/2020 14:54 08/08/2020 14:54	
Attributes Owner Computer	A LAPTOP-3HR9A551\jun LAPTOP-3HR9A551 (thi	ju is PC)	Attributes Owner Computer	A SYSTEM LAPTOP-3HR9A551 (thi	s PC)
emove Proper	rties and Personal Informa	ation	Remove Proper	ties and Personal Informa	ation

Figure 5 Tableau Properties

² https://www.tableau.com/products/desktop/download?signin=academic

2.2.3 IMB SPSS Statistics

IBM SPSS Statistics version 26 (Figure 6) was downloaded from here³.

SPSS was used to implement Statistic model.



Figure 6 SPSS Statistics Version

From the IBM SPSS statistics 26 properties (Figure 7) shown it was created on 02/10/2019, installed on my machine name PC LAPTOP-3HR9A551.

General Security	Shortcut Details	Com Previous	ipatibility Versions
Property File Name	Value	link	
Type	Shortcut		
Folder path	C:\Users\junju\AppData	\Roaming\Micro	soft\I
Size	2.27 KB	0	
Date created	02/10/2019 22:43		
Date modified	02/10/2019 22:37		
Attributes	A		
Owner	LAPTOP-3HR9A551\jur	nju	
Remove Proper	ties and Personal Inform	ation	

Figure 7 SPSS Statistics Properties

³ https://itsupport.ncirl.ie/hc/en-ie/articles/360014035839-How-do-I-install-SPSS-

3 Data Preparation

Data was extracted from Kaggle and Data.World website, datasets have been already in csv format, there were five csv datasets used. Each one was cleaned, formatted and prepared into one final dataset. The original five datasets can be located from these links⁴.

3.1 Data Preparation used RStudio

3.1.1 Package Install in RStudio

All the below modelling (Table 1) functions used split data, training and testing. This was divided up into 70% training data and the remaining 30% for testing data. Both data were saved into csv files to be used on each model for reusability to save time, duplication and complexity.

Name	Model	Package
Data Preparation		library(dplyr)
Multiple Linear Regression	lm	library(ggplot2)
Random Forest	randomForest	library(randomForest)
Naïve Bayes	naiveBayes	library(e1071)
Logistic Regression	glm	N/A No packet needs for model
Support Vector Machine	svm	library(e1071)
Decision Tree	mp o at	library(party)
Decision free	ipart	library(rpart)
K-Nearest Neighbour	knn	library(class) Used for the model
Accuracy Check		library(caret)
		library(class)
		library(gmodels)

Table 1 RStudio Package Installed

3.1.2 Data Clean and Encode

```
#-Read the data set file
insurance1D5 <- read.csv("datasets_26475_38092_insurance2.csv", header=TRUE, stringsAsFactors=FALSE, fileEncoding="latin1")
insurance1 <- select(insurance1DS, age, sex, bmi, children, smoker, region, charges)
insurance1$smoker <- ifelse(insurance1DS$smoker=="1", "yes", "no")
insurance1$sex <- ifelse(insurance1DS$sex=="1", "Male", "Female")</pre>
#-Read the data set file
insurance2D5 <- read.csv("datasets 26475_38092_insurance3r2.csv", header=TRUE, stringsAsFactors=FALSE, fileEncoding="latin1")
insurance2 <- select(insurance2DS, age, sex, bmi ,children, smoker, region, charges)
insurance2$smoker <- ifelse(insurance2D5$smoker=="1", "yes", "no")
insurance2$sex <- ifelse(insurance2DS$sex=="1", "Male", "Female")</pre>
```

Figure 8 Clean First Two Datasets

There are five csv datasets were selected for this project, figure 8 shows the reading of first and second csv files, selecting 7 variables, converting both smoker and sex data values from numeric to factor value.

4 https://www.kaggle.com/easonlai/sample-insurance-claim-prediction-dataset https://www.kaggle.com/hiralpandhi/healthcaredataset?select=test_2v.csv

https://data.world/healthdatany/gaf8-ac33

```
#-Merged Data sets
mergedInsurance <- rbind(insurance1, insurance2)
#-Convert numerical values to description values
for(ns in 1:nrow(mergedInsurance)) {
    if(mergedInsurance$region[ns] == 0){
        mergedInsurance$region[ns] <- "northeast"
    }else if(mergedInsurance$region[ns] == 1){
        mergedInsurance$region[ns] <- "northwest"
    }else if(mergedInsurance$region[ns] == 2){
        mergedInsurance$region[ns] <- "southeast"
    }else (mergedInsurance$region[ns] <- "southwest")
}</pre>
```

Figure 9 Merge First Two Datasets

Figure 9 shows the merging of the first and second csv file into one dataset (mergeInsurance) after the merged has been completed the process loops through each mergeInsurance row and assign each region value to a descriptive value.

```
#-Read the data set file
train_2v <- read.csv("train_2v.csv", header=TRUE, stringsAsFactors=FALSE, fileEncoding="latin1")</pre>
newTrain <- select(train_2v, age, gender, bmi, smoking_status)</pre>
newTrain[newTrain==""]<-NA
newTrain <- na.omit(newTrain)</pre>
newTrain$smoking_status <- ifelse(newTrain$smoking_status=="never smoked", "no", "yes")</pre>
newTrain <- subset(newTrain, age >= 18)
newTrain <- subset(newTrain, bmi >= 40 & bmi <= 70)</pre>
#-Read the data set file
test_2v <- read.csv("test_2v.csv", header=TRUE, stringsAsFactors=FALSE, fileEncoding="latin1")</pre>
newTest <- select(test_2v, age, gender, bmi, smoking_status)</pre>
newTest[newTest==""]<-NA
newTest <- na.omit(newTest)</pre>
newTest$smoking_status <- ifelse(newTest$smoking_status=="never smoked", "no", "yes")</pre>
newTest <- subset(newTest, age >= 18)
newTest <- subset(newTest, bmi >= 40 & bmi <= 70)
```

Figure 10 Clean Other Two Datasets

The above figure 10 reads third and fourth dataset, both datasets were select 4 variables and assign NA into empty value row to allow them to be removed. And anything else into smoker value. The selection was also done by age group which over or equal to18 and bmi group between 40 to 70.

```
#-Merged Data sets
mergedTestTrain <- rbind(newTest, newTrain)
#-Naming the columns
colnames(mergedTestTrain)=c("age","sex", "bmi", "smoker")
#-Check for NA values in the merged Data sets
any(is.na(mergedTestTrain))</pre>
```

Figure 11 Merged Other Two Datasets

The figure 11 shows the merging of the third and the fourth dataset into one dataset called mergeTestTrain, then modifies all the four column names to match to first two datasets, put these variables into mergeTestTrain data, and check any NA value.

```
#-Read the data set file
chargesData <- read.csv("Inpatient_Prospective_Payment_System_IPPS_Provider_Summary_for_the_Top_100_Diagnosis-Related_Groups_DRG_-_FY2011.csv",
                         header=TRUE, stringsAsFactors=FALSE, fileEncoding="latin1")
chargeData <- select(chargesData, Average.Total.Payments)
chargeData1 <- subset(chargeData, Average.Total.Payments >= 40000 & Average.Total.Payments <= 70000)
chargeData1 <- unique(chargeData1)
#-Add one region values to each mergedTestTrain Data sets row
regions <- c("northeast", "northwest", "southeast", "southwest")</p>
mergedTestTrain$region <- sample(regions, size = nrow(mergedTestTrain), replace = TRUE)
#-Add number of children to each row in mergedTestTrain Data sets
children <- c(0,1,2,3,4,5)
mergedTestTrain$children <- sample(children, size = nrow(mergedTestTrain), replace = TRUE)</pre>
#-Add charges to each row in mergedTestTrain Data sets
mergedTestTrain$charges <- sample(chargeData1$Average.Total.Payments, size = nrow(mergedTestTrain), replace = TRUE)
#-Merged both data sets into insuranceData
insuranceData <- rbind(mergedInsurance, mergedTestTrain)
#-Check for NA values in insuranceData set
any(is.na(insuranceData))
```



This figure 12 reads the fifth csv file, and then select charges value between 40000 to 70000 with unique value. Create two new columns region and children, then added 4 different regions and number of children (0-5) to its own variable to be added into MergeTestTrain dataset. All five datasets were added into master dataset called insuranceData.

3.1.3 Presenting Data



Figure 13 Data Presentation of age and bmi Distribution

Figure 13 shows the data presentation using boxplot generate age and bmi, showing the outlier of the data in red.



Figure 14 Data Presentation for age, bmi, charges Distribution

The figure 14 presents data distribution from generated histogram, used age, bmi and charges variables shows the count and the number of values related each group.



Figure 15 Data Presentation children, sex, region, smoker

The figure 15 using bar charts generate the data presentation, which demonstrating four aspect groups for children, sex, smoker and region distribution.

3.1.4 Split into Training and Testing Dataset

```
#-Set data types
insuranceData$charges <- as.numeric(insuranceData$charges)</pre>
insuranceData$age <- as.integer(insuranceData$age)</pre>
insuranceData$sex <- as.factor(insuranceData$sex)</pre>
insuranceData$children <- as.integer(insuranceData$children)</pre>
insuranceData$smoker <- as.factor(insuranceData$smoker)</pre>
insuranceData$region <- as.factor(insuranceData$region)</pre>
str(insuranceData)
#-Create insurance Data file
write.csv(insuranceData, "insuranceData.csv", row.names=FALSE)
#Split the data 70/30
percentData <- round(0.7 * nrow(insuranceData))</pre>
sampleData <- sample(1:nrow(insuranceData), percentData)</pre>
trainInsuranceData <- insuranceData[sampleData, ]</pre>
testInsuranceData <- insuranceData[-sampleData, ]</pre>
#-Create a Train and Test files
write.csv(trainInsuranceData, "trainInsuranceData.csv", row.names=FALSE)
write.csv(testInsuranceData, "testInsuranceData.csv", row.names=FALSE)
#-Read the data set files and store in variables
train <- read.csv("trainInsuranceData.csv", stringsAsFactors = TRUE, header=TRUE, fileEncoding="latin1")</pre>
test <- read.csv("testInsuranceData.csv", stringsAsFactors = TRUE , header=TRUE, fileEncoding="latin1")</pre>
#-Check for NA values in train & test data sets
any(is.na(train))
any(is.na(test))
> #-Set data types
> insuranceData$charges <- as.numeric(insuranceData$charges)</pre>
> insuranceData$age <- as.integer(insuranceData$age)</pre>
> insuranceData$sex <- as.factor(insuranceData$sex)</pre>
> insuranceData$children <- as.integer(insuranceData$children)</pre>
> insuranceData$smoker <- as.factor(insuranceData$smoker)</pre>
> insuranceData$region <- as.factor(insuranceData$region)</pre>
> str(insuranceData)
'data.frame': 6406 obs. of 7 variables:
$ age : int 19 18 28 33 32 31 46 37 37 60 ...
         : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 1 1 1 2 1 ...
 $ sex
          : num 27.9 33.8 33 22.7 28.9 ...
 $ bmi
 $ children: int 0130001320...
 $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 ...
 $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges : num 16885 1726 4449 21984 3867 ...
```

Figure 16 Split to Training and Testing Data

The code in figure 16 shows the setting to data types to meet the model requirement and save the insuranceData file, the data is then split into 70% training and 30% testing and saved both datasets to be kept secure. Both train and test files are read and store in its own variable to be used in each model.

3.2 Data Preparation uses SPSS

The first step was imported csv file into SPSS software (Figure 17).

🝓 Untitled2 [DataSet1] - IBM	A SPSS Statistics D	Data Edit	or												-	o ×
<u>Eile E</u> dit <u>V</u> iew Dat	a <u>T</u> ransform	Analyz	e <u>G</u> raphs	<u>U</u> tilities	Extensions	Window	<u>H</u> elp	p								
New		•	醫士	=	H H											
<u>O</u> pen		•						1								
Import <u>D</u> ata		•	Data <u>b</u> ase		•				¥	K	Y	W.	1	W.	Visible: 0	of 0 Variables
💼 <u>C</u> lose	Ctrl+F4		Excel			va	r	var	var	var	var	var	var	var	var	var
Save	Ctrl+S		<u>C</u> SV Data				_									-
S <u>a</u> ve As			Text Data													
🔚 Save All Data			<u>S</u> AS													
Export		•	Stata													
Mark File Read Only			d <u>B</u> ase													
Revert to Saved File			Lotus													
🛒 Rena <u>m</u> e Dataset			SYLK													
Display Data File Inform	nation	•	Cognos T <u>M</u> 1													_
🙍 Cac <u>h</u> e Data			Cognos Bus	iness Intell	igence		_									
Stop Processor	Ctrl+Pe	riod				_	-									
🙀 Switch Server																
Repository		•														
Rrint Pre <u>v</u> iew																
Print	Ctrl+P					_										
Welcome Dialog						_										
Recently Used Data		•				_										
Recently Used Files		•					_									
Exit																-
4			-													•
Data View Variable View																
COV Data											IDM ODO	2 Otobiotico Dr		utu - 11	nice de: Ohl	
CSV Data											IBM SPS	s statistics Pro	ocessor is rea	idy U	nicode:ON	

Figure 17 SPSS Data Import

The second step was encoded data used to transform and selected record into same variables (Figure 18).

e <u>E</u> dit	<u>View</u> <u>D</u> ata	Transform Analyze G	araphs <u>U</u> tilit	es Ext	ension	ns <u>W</u> indow	Help								
28		Compute Variable			h.		14 0 0								
		Count Values within C	ases			ELECTRON							V	isible: 7 of 7	Varia
	1 A	Shift Values				0	R alarma	1000	Later	1.00.0	1000		1		Turiu
21	of age €0.1	Recode into Same Var	iables		er	region	Charges 12228 946950	Vär	Var	Val	Var	Var	Var	Vdf	
20	20 1	Recode into Different V	/ariables			1	4140 726000								+
22	10 2	Automatic Recode				4	4143.730000								+
23	24.1	Visual Binning	Visual <u>B</u> inning				27701 976900								+
24	34 1	🚰 Optįmal Binning				2	57701.8766000								-
25	50 1					2	14001 133800								+
20	62 1	Prepare Data for Mode	aling	,		1	14461 935150								+
21	65 1	Rank Cases				2	12268 632260								+
20	22.2					2	2775 102150								+
20	23 2	Create Time Series	Create Time Series Replace Missing Values				22713.132130								+
31	22.2	Replace Missing Value					35585 576000							_	+
32	18 1	😪 Random Number Gen	erators			4	2198 189850								+
32	19 1	Run Pandino Transfor		3r1+G		4	4687 797000								+
34	63 2	Run Pending Transforms Ctri+G			2	13770 097900								+	
35	28.2	36 400	12			4	51194 559140								+
36	19.2	20.425	0 1			2	1625 433750								+
37	62 1	32 965	31			2	15612 193350								+
38	26.2	20,800	0 1			4	2302 300000								+
39	35.2	36 670	12			1	39774 276300							-	+
40	60 2	39,900	0.2			4	48173.361000								+
41	24 1	26.600	0 1			1	3046.062000							-	+
	1													_	

Figure 18 Data Encode

Old Value	New Value	
∑alue:	Value:	
	System-missing	
System-missing		
) System- or user-missing	Ol <u>d</u> > New:	
Bange:	'Female'> '1'	
grid <u>rige</u> .	'Male'> '2'	
li and a second	Add	
inrougn	Change	
	Remove	
Range, LOWEST through value:		
Range, value through HIGHEST:		
All other values		

The third step below was modified sex variable into a numeric value (Figure 19).

Figure 19 Change Variable sex to Number

This step was same as previous step change variable smoker to a numeric value (Figure 20).

🙀 Recode into Same Variables: Old and New Values	×
Old Value	New Value Value: System-missing Old> New: 'no'> '1' 'yes'> '2' Add Change Remove
	Cancel Help

Figure 20 Change Variable Smoker to Number

Figure 21 below is shown the Encode result.

```
RECODE sex ('Female'='l') ('Male'='2').
EXECUTE.
RECODE smoker ('no'='l') ('yes'='2').
EXECUTE.
```



Name	Туре	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
age	Numeric	2	0		None	None	8	端 Right	🛷 Scale	🔪 Input
sex	String	6	0		None	None	6	📑 Left	💦 Nominal	🔪 Input
bmi	Numeric	6	3		None	None	8	I Right	Scale	> Input
children	Numeric	1	0		None	None	8	■ Right	💦 Nominal	> Input
smoker	String	3	0		None	None	10	≣ Left	💦 Nominal	> Input
region	Numeric	1	0		None	None	8	≣ Right	💦 Nominal	> Input
charges	Numeric	12	6		None	None	14	I Right	Scale	🔪 Input
-										
Name	Туре	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
age	Numeric	2	0		None	None	8	遍 Right	I Scale	🔪 Input
sex	Numeric	6	0		None	None	6	遍 Right	🙈 Nominal	🔪 Input
bmi	Numeric	6	3		None	None	8	🗃 Right	I Scale	🔪 Input
children	Numeric	1	0		None	None	8	🖷 Right	🚓 Nominal	🔪 Input
smoker	Numeric	3	0		None	None	10	Right	🙈 Nominal	🔪 Input
region	Numeric	1	0		None	None	8	🖷 Right	🚓 Nominal	🔪 Input
charges	Numeric	12	6		None	None	14	■ Right	scale 🎸	ゝ Input
				Variable Type Varia	ation ency umeric (integer eric type honors iever uses digit	with leading zero the digit groupin grouping.	Decimal <u>f</u> is) is setting, whil	Width: 8 2laces: 2 ie the Restricted		

After encoding two variables, final step changed string variable to a numeric value (Figure 22).

Figure 22 Change Data Type from String to Numeric

3.3 Tableau Data Imported

Figure 23 shows Tableau imported csv dataset for analyses and used visualisation for data presentation.

💮 Tableau - Book2 File Data Server Help						– 0 ×
*						
Connect Search for Data Tableau Server To a File Microsoft Excel Text file JSON file Microsoft Access PDF file Spatial file Statistical file More	Open	Tableau 5 graphs	Project Add this Graph	Betriko Gesender of Bible Serverse Ermission_Sum Oorte Research of c 	Open a Workbook	Discover Iraining View all 87 training videos Composed Resources Get Tableau Prep Blog - Applications are open! Why you should apply today to speek at Tableau's Virtual Forums
Microsoft SQL Server MySQL Oracle Amazon Redshift	Sample Workboo	Tableau 5 graphs			More Samples	
More >		236				Update to 2020.2.4 now

Figure 23 Tableau Imported Data

Figure 24 shows data were successfully imported into Tableau, then the change data type from string variable to a numeric variable, this is for analyses required.

🕸 Tableau - Book1 File Data Server Window Help						– a ×
$\label{eq:constraint} \Leftrightarrow \leftarrow \rightarrow \hbox{\mbox{${\rm m}$}$} \bigcirc $	^{₀.} insu	ranceDa	ata			Connection Filters
Connections Add						
insuranceData Text file	insurance)ata.csv				
Files P						
Use Data Interpreter						
Data Interpreter might be able to clean your Text file workbook.						
# datasets_26475insurance2.csv						
≣ datasets_26475nsurance3r2.csv	🔳 🗏 Sort fiel	ds Data sou	rce order 🔻			✓ Show aliases Show hidden fields 1,000 → rows
Inpatient_ProspGFY2011.csv	# Abs	#	🔹 👻 Abc	+	+	
	insuran insura	ice insuranc	Number (decimal)	insuranceD	insuranceDat	
m test_2v.csv	age sex	bmi	Number (whole)	region	charges	
testInsuranceData.csv	26 Fema	le 28.7850	Date & Time	1	3,385.40	
train_zv.csv train_sv.csv	18 Male	33.7700	String	3	1,725.55	
iiii traininsuranceData.csv	28 Male	33.0000	Boolean	3	4,449.46	
Rew Union	33 Male	22,7050	✓ Default	2	21.984.47	
	22 Malo	20 0000	Geographic Role 🔸	2	2 966 96	
	32 Widie	20.0000	0 110	2	3,800.80	
	31 Fema	le 25.7400	U no	3	3,/50.02	
	46 Fema	le 33.4400	1 no	3	8,240.59	
	37 Fema	le 27.7400	3 no	2	7,281.51	
Data Source Sheet 1 Sheet 2 Image: Advance 4	884 UL					
						((く)) 田 田 田

Figure 24 Dataset Preparation in Tableau

4 Implementation and Result Generated Steps

The implementation steps delivered into feature selection and machine learning model generation. There were 7 features selected and applied into 7 machine learning models in order to achieve the project objective. The features selected used RStudio generated Correlation Coefficient function and used SPSS generate a correlation table, both correlation result represent and evaluated the features was expected to conduct the implementation. After feature selection, there were several suitable libraries were installed in RStudio for generating the models, machine learning models used Multiple Linear Regression, Random Forest (Regression and Classification), Support Vector Machine (Regression and Classification), Naïve Bayes, Decision Tree, Logistic Regression and K-Nearest Neighbor to implemented, the output of models implementation used for evaluated by R-square (Regression model generated) and accuracy(Classification model generated).

4.1 RStudio – Models Generated

This project uses random report from code generated, the result of data might be varied as each time the output result from code generated was machine random selected value.

4.1.1 Multiple Linear Regression (Regression)



Figure 25 Multiple Linear Regression Model

Multiple Linear Regression model generated shows in figure 25, the result shows the R-square value and P-value, and the linear ggplot shows the model visualisation.

4.1.2 Random Forest (Regression)



Figure 26 Random Forest Regression Model

Figure 26 shows the Random Forest Regression Model, which used charge as dependent variable generated R-square value for comparison with other models. Source code for accuracy generated was from DataTechNotes (Website, 2018), and source code for plot chart was by RStudio-pubs-static (Website, n.d.).

4.1.3 Support Vector Machine (Regression)



Figure 27 Support Vector Machine Regression Model

Support Vector Machine Regression Model generated in figure 27, it is generated R-square result for comparison with other models. The source code for accuracy generated was from DataTechNotes (Website, 2018), the source code for plot generated was by DataTechNotes (Website, 2019).

4.1.4 Naïve Bayes (Classification)

```
#-----Naive Bayes-----
#-Create NB Model
model_nb <- naiveBayes(smoker ~ ., data = train)</pre>
#-Create the prediction from the NB Model
pred1 <- predict(model_nb, test, type="class")</pre>
#Create Confusion Matrix
confusionMatrix(pred1, test$smoker)
             > model_nb <- naiveBayes(smoker ~ ., data = train)</pre>
             > #-Create the prediction from the NB Model
             > pred1 <- predict(model_nb, test, type="class")</pre>
             > #Create Confusion Matrix
             > confusionMatrix(pred1, test$smoker)
             Confusion Matrix and Statistics
                        Reference
             Prediction no yes
                     no 654 164
                     yes 543 561
                             Accuracy : 0.6322
                               95% CI : (0.6101, 0.6538)
                  No Information Rate : 0.6228
                  P-Value [Acc > NIR] : 0.2053
                                Kappa : 0.2902
               Mcnemar's Test P-Value : <2e-16
                          Sensitivity : 0.5464
                          Specificity : 0.7738
                       Pos Pred Value : 0.7995
                       Neg Pred Value : 0.5082
                           Prevalence : 0.6228
                       Detection Rate : 0.3403
                Detection Prevalence : 0.4256
                    Balanced Accuracy : 0.6601
                     'Positive' Class : no
```

Figure 28 Naive Bayes Classification Model

Figure 28 shows the Naïve Bayes model generated and result represented accuracy and P-value compared other models.

4.1.5 Decision Tree (Classification)

```
#-----Decision Tree-----
#-Create DT Model
model_dt <- rpart(smoker ~ .,data=train,method = "class",parms = list(prior = c(0.3, 0.7)))</pre>
#-Create the prediction from the DT model
pred <- predict(model_dt,test,type="class")</pre>
#Create Confusion Matrix
confusionMatrix(pred,test$smoker)
#-Create Conditional Inference Tree
output.tree <- ctree(smoker ~ ., data = train)</pre>
plot(output.tree, main="Conditional Inference Tree with Insurance Data", gp = gpar(fontsize = 6))
     > model_dt <- rpart(smoker ~ .,data=train,method = "class",parms = list(prior = c(0.</pre>
     3. 0.7)))
     > #-Create the prediction from the DT model
     > pred <- predict(model_dt,test,type="class")</pre>
     > #Create Confusion Matrix
     > confusionMatrix(pred,test$smoker)
     Confusion Matrix and Statistics
               Reference
     Prediction no yes
            no 574 0
            yes 623 725
                    Accuracy : 0.6759
                      95% CI : (0.6544, 0.6968)
         No Information Rate : 0.6228
         P-Value [Acc > NIR] : 7.016e-07
                        Kappa : 0.4101
      Mcnemar's Test P-Value : < 2.2e-16
                 Sensitivity : 0.4795
                 Specificity : 1.0000
              Pos Pred Value : 1.0000
              Neg Pred Value : 0.5378
                  Prevalence : 0.6228
              Detection Rate : 0.2986
        Detection Prevalence : 0.2986
           Balanced Accuracy : 0.7398
             'Positive' Class : no
```

Figure 29 Decision Tree Classification Model

Figure 29 shows the Decision Tree model generation, which includes accuracy and P-value. Next figure 30 is the tree graphs represent the model visualisation.



Figure 30 Decision Tree Graph

Decision Tree graph (Figure 30) shows smoker behaviour insurers paid higher health insurance premium than non-smoker behaviour insurers.

4.1.6 Logistic Regression (Classification)

```
#-----Logistic Regression-----
#-Create General Linear Model
glmModel <- train(smoker ~ .,data = train, method="glm",family = "binomial")
summary(glmModel)
confusionMatrix(glmModel)
#-Create the prediction from the GLM
pred <- predict(glmModel,test)</pre>
#Create Confusion Matrix
confusionMatrix(pred, test$smoker)
> glmModel <- train(smoker ~ .,data = train, method="glm",family = "binomial") > confusionMatrix(glmModel)
                                                                             Bootstrapped (25 reps) Confusion Matrix
> summary(glmModel)
                                                                              (entries are percentual average cell counts across resamples)
Call:
NULL
                                                                                         Reference
                                                                             Prediction no yes
no 51.2 20.7
Deviance Residuals:
Min 1Q Median 3Q Max
-2.1662 -0.9302 -0.4427 1.0914 1.8904
                                                                                      yes 12.9 15.3
                                                                              Accuracy (average) : 0.6645
Coefficients:
                                                                             > #-Create the prediction from the GLM
                Estimate Std. Error z value Pr(>|z|)
                                                                             > pred <- predict(glmModel,test)</pre>
(Intercept)
             -1.322e+00 2.142e-01 -6.171 6.79e-10 ***
                                                                             > #Create Confusion Matrix
              -8.470e-03 2.405e-03 -3.522 0.000428 ***
                                                                              > confusionMatrix(pred, test$smoker)
age
sexMale
              5.050e-01 7.383e-02 6.840 7.93e-12 ***
                                                                             Confusion Matrix and Statistics

        bmi
        -4.523e-02
        5.879e-03
        -7.695
        1.42e-14 ***

        children
        -2.284e-02
        2.179e-02
        -1.048
        0.294682

        regionnorthwest
        9.796e-02
        9.956e-02
        0.984
        0.325178

                                                                                         Reference
                                                                             Prediction no yes
no 959 408
regionsoutheast 1.706e-01 9.783e-02 1.744 0.081118 .
regionsouthwest -1.233e-02 9.944e-02 -0.124 0.901289
                                                                                     yes 238 317
                                                                                               Accuracy : 0.6639
95% CI : (0.6423, 0.685)
               7.396e-05 3.169e-06 23.338 < 2e-16 ***
charges
---
                                                                                  No Information Rate : 0.6228
P-Value [Acc > NIR] : 9.912e-05
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
                                                                                                   карра : 0.25
   Null deviance: 5851.5 on 4483 degrees of freedom
                                                                              Mcnemar's Test P-Value : 2.947e-11
Residual deviance: 4882.8 on 4475 degrees of freedom
                                                                                            Sensitivity : 0.8012
AIC: 4900.8
                                                                                            Specificity : 0.4372
                                                                                       Pos Pred Value : 0.7015
Neg Pred Value : 0.5712
Number of Fisher Scoring iterations: 4
                                                                                             Prevalence : 0.6228
                                                                                        Detection Rate : 0.4990
                                                                                 Detection Prevalence : 0.7112
Balanced Accuracy : 0.6192
                                                                                      'Positive' Class : no
```

Figure 31 Logistic Regression Classification Model

Figure 31 shows the Logistic Regression model generated used confusion matrix and summary to illustrate the model. The accuracy and P-value were used for compression result.

4.1.7 K-Nearest Neighbour (Classification)

```
#----
      -----K-Nearest Neighbour-----
#-Store Smoker label Value
train_label <- train[,5]</pre>
test_label <- test[,5]</pre>
#-Set Factors data types to Nummeric types
train$smoker<-as.numeric(train$smoker)</pre>
test$smoker<-as.numeric(test$smoker)</pre>
train$region<-as.numeric(train$region)</pre>
test$region<-as.numeric(test$region)</pre>
train$sex<-as.numeric(train$sex)</pre>
test$sex<-as.numeric(test$sex)</pre>
#https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-programming/
#-Create KNN Model
model_knn <- knn(train = train, test = test,cl = train_label, k = 5, prob = TRUE)</pre>
#-check the accuracy of the predicted values
CrossTable(x = test_label, y = model_knn)
#-Create table count of Prediction VS Actual
table <- table(model_knn, test_label,dnn=c("Prediction","Actual"))</pre>
#-Get table accuracy
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}</pre>
accuracy(table)
 > model_knn <- knn(train = train, test = test,cl = train_label, k = 5, prob = TRUE)
> #-check the accuracy of the predicted values
> CrossTable(x = test_label, y = model_knn)
    cell Contents
   Chi-square contribution
             N / ROW TOTAL
N / COl TOTAL
N / TABLE TOTAL
 Total Observations in Table: 1922
                  model_knn
                  no |
                              9 | yes | Row Total
   test label
                                 301 |
45.746 |
0.251 |
0.423 |
0.157 |
                          896
                                                    1197
            no
                      26.918
0.749
0.740
                                                    0.623
                        0.466
                                             -
                                  411
                          314
           yes
                               | 411
| 75.529
| 0.567
| 0.577
| 0.214
                      44.443
                                                    0.377
                        0.260
                        0.163
                        ----i ·
                                      712 |
                       1210 | 712 | 1922
0.630 | 0.370 |
                                               1922
 Column Total
 > #-Create table count of Prediction VS Actual
> table <- table(model_knn, test_label,dnn=c("Prediction","Actual"))
> #-Get table accuracy
> accuracy <- function(x){sum(diag(x)/(sum(rowsums(x)))) * 100}
> accuracy(table)
fil se oppoe
 > accuracy(t
[1] 68.00208
```

Figure 32 K-Nearest Neighbor Classification Model

Figure 32 shows the K-Nearest Neighbour model generated using cross table to get accuracy for comparison. Source code was from a website (Choudhury, 2015).

4.1.8 Support Vector Machine (Classification)

```
#-----Support Vector Machine Classification-----
#-Set Nummeric data types to Factors types
train$smoker<-as.factor(train$smoker)</pre>
test$smoker<-as.factor(test$smoker)</pre>
train$region<-as.factor(train$region)</pre>
test$region<-as.factor(test$region)</pre>
train$sex<-as.factor(train$sex)</pre>
test$sex<-as.factor(test$sex)</pre>
#-Create SVM Classification Model
model_svmc <- svm(smoker ~ . , train)</pre>
#-Create the prediction from the SVMC
pred <- predict(model_svmc, test, type="class")</pre>
#Create Confusion Matrix
confusionMatrix(pred, test$smoker)
                   > model_svmc <- svm(smoker ~ . , train)</pre>
                   > #-Create the prediction from the SVMC
                   > pred <- predict(model_svmc, test, type="class")</pre>
                   > #Create Confusion Matrix
                   > confusionMatrix(pred, test$smoker)
                   Confusion Matrix and Statistics
                             Reference
                   Prediction
                            on 1 2
1 1005 353
2 192 372
                                  Accuracy : 0.7164
                                    95% CI : (0.6957, 0.7365)
                       No Information Rate : 0.6228
                       P-Value [Acc > NIR] : < 2.2e-16
                                     Kappa : 0.3689
                    Mcnemar's Test P-Value : 7.199e-12
                               Sensitivity : 0.8396
                               Specificity : 0.5131
                            Pos Pred Value : 0.7401
                            Neg Pred Value : 0.6596
                                Prevalence : 0.6228
                            Detection Rate : 0.5229
                      Detection Prevalence : 0.7066
                         Balanced Accuracy : 0.6764
                          'Positive' Class : 1
```

Figure 33 Support Vector Machine Classification Model

Figure 33 shows the Support Vector Machine model generated used confusion matrix to get output from accuracy and P-value for comparison.

4.1.9 Random Forest (Classification)

```
#------Random Forest Classification------Random Forest Classification
#-Create RF Classification Model
model_rfc <- randomForest(smoker ~ ., data = train)</pre>
#-Create the prediction from the RFC
pred = predict(model_rfc, newdata=test[-5])
#Create Confusion Matrix
confusionMatrix(pred, test$smoker)
             > model_rfc <- randomForest(smoker ~ ., data = train)</pre>
             > #-Create the prediction from the RFC
             > pred = predict(model_rfc, newdata=test[-5])
             > #Create Confusion Matrix
             > confusionMatrix(pred, test$smoker)
             Confusion Matrix and Statistics
                       Reference
             Prediction 1 2
                      1 950 314
                      2 247 411
                            Accuracy : 0.7081
                              95% CI : (0.6872, 0.7284)
                 No Information Rate : 0.6228
                 P-Value [Acc > NIR] : 2.389e-15
                               Kappa : 0.3672
              Mcnemar's Test P-Value : 0.005328
                         Sensitivity : 0.7937
                         Specificity : 0.5669
                      Pos Pred Value : 0.7516
                      Neg Pred Value : 0.6246
                          Prevalence : 0.6228
                      Detection Rate : 0.4943
                Detection Prevalence : 0.6576
                   Balanced Accuracy : 0.6803
                    'Positive' Class : 1
```

Figure 34 Random Forest Classification Model

Figure 34 shows Random Forest model generation, it used confusion matrix generated accuracy and P-value for comparison.

4.2 SPSS – Analysis & Models Generated

4.2.1 Correlation Table Generated

Click SPSS-Analyze-Correlate-Bivariate (Figure 35).

🝓 *Untitled	3 [DataSet2]	- IBM SI	PSS Statistics D	ata Editor											-	٥	×
<u>E</u> ile <u>E</u> di	t <u>V</u> iew	<u>D</u> ata	Transform	Analyze	<u>G</u> raphs	Utilities	Extensions	Window	Help								
			5	Re <u>p</u> o D <u>e</u> so	orts criptive Stati	stics	•			•							
5 : smoker		1		Baye	sian Statist	ics	*								Visible	7 of 7 Varia	ables
	🧳 ag	ge	💦 sex 🛛 🤞	Ta <u>b</u> le	es		*	💫 region	🔗 charges	var	var	var	var	var	var	var	
1		19	1	Com	pare Means	8	+	4	16884.92	000		1					-
2		18	2	Gene	eral Linear I	lodel		3	1725.55	300							
3		28	2	Gene	eralized Lin	ear Models		3	4449.46	000							
4		33	2	Mixe	d Models		+	2	21984.47	610							
5		32	2	Corre	elate		- F	Rivariate	3866.85	200		7					
6	_	31	1	Regr	ression			Dortiol	3756.62	600							
7	_	46	1	Logli	inear			E Fajual	8240.58	600		1					
8	_	37	1	Neur	ral Networks	8		Distanc	es 7281.50	600							_
9		37	2	Clas	sifv		,	1	6406.41	700							
10	_	60	1	Dime	ension Red	uction		2	28923.13	920							
11		25	2	Scale		Generi		1	2721.32	800							
12	_	62	1	Non	o noromatrio 1	Tanta		3	27808.72	100							
13		23	2	E area	parametric	16515		4	1826.84	000							
14	_	56	1	Fore	casjing		ľ.	3	11090.71	800							
15		27	2	Survi	ival			3	39611.75	700							
16		19	2	Multi	ple Respon	se	×.	4	1837.23	000							
17		52	1	Missi	ing Value Ar	nal <u>v</u> sis		1	10797.33	200							
18		23	2	Mulți	ple Imputati	on	*	1	2395.17	550							
19		56	2	Com	plex Sampl	es	*	4	10602.38	000							
20		30	2	📳 Sįmu	lation			4	36837.46	000							
21		60	1	Qual	lity Control			1	13228.84	950							-
	4			Spat	ial and Tem	poral Mode	ling 🕨										Þ
Data View	Variable V	iew		Direc	ct Mar <u>k</u> eting		+										
Bivariate											IBM	SPSS Statisti	cs Processor i	s ready	Unicode:ON		1

Figure 35 Generated Correlation

After the previous step, this step selected all variables into Pearson for correlation analyses (Figure 36).

tai Bivariate Correlations		\times
	Variables:	Options Style Bootstrap
Correlation Coefficients	au-b 🥅 <u>S</u> pearman	
Test of Significance <u>Two-tailed</u> One-tailed	d	
Elag significant correlatio	ns ste <u>R</u> eset Cancel Help	

Figure 36 Select Variables

		200	5 O.Y	h mi	childron	emeker	ragion	charges
		age	sex		children	smoker	region	charges
age	Pearson Correlation	1	062	.235	.132	.082	001	.333
	Sig. (2-tailed)		.000	.000	.000	.000	.908	.000
	Ν	6406	6406	6406	6406	6406	6406	6406
sex	Pearson Correlation	062**	1	165**	080""	.029	015	169**
	Sig. (2-tailed)	.000		.000	.000	.022	.224	.000
	Ν	6406	6406	6406	6406	6406	6406	6406
bmi	Pearson Correlation	.235**	165**	1	.333""	.219**	.031	.730**
	Sig. (2-tailed)	.000	.000		.000	.000	.014	.000
	Ν	6406	6406	6406	6406	6406	6406	6406
children	Pearson Correlation	.132**	080**	.333	1	.110**	.001	.369**
	Sig. (2-tailed)	.000	.000	.000		.000	.907	.000
	Ν	6406	6406	6406	6406	6406	6406	6406
smoker	Pearson Correlation	.082**	.029	.219**	.110**	1	.008	.412**
	Sig. (2-tailed)	.000	.022	.000	.000		.504	.000
	Ν	6406	6406	6406	6406	6406	6406	6406
region	Pearson Correlation	001	015	.031	.001	.008	1	005
	Sig. (2-tailed)	.908	.224	.014	.907	.504		.674
	N	6406	6406	6406	6406	6406	6406	6406
charges	Pearson Correlation	.333	169**	.730**	.369**	.412**	005	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.674	
	Ν	6406	6406	6406	6406	6406	6406	6406

The correlation table generated show figure 37.

Figure 37 Correlation Table

4.2.2 ANOVE Model Generated

SPSS-Analyze-Regression-Linear (Figure 38).



Figure 38 ANOVA Generated

🝓 *Untitled3	[DataSet2] - IBM	SPSS Statistic	s Data Editor												_	đ	×
<u>File</u> Edit	<u>V</u> iew <u>D</u> ata	Transform	n <u>A</u> nalyze	<u>G</u> raphs	<u>U</u> tilities	Extensions	<u>W</u> indo	w <u>F</u>	<u>H</u> elp								
📄 H			∽ 📱		L L												
5 : smoker	1														Visible	7 of 7 Va	ariables
	🛷 age	💑 sex	🛷 bmi	💑 childre	ta Linear	r Rearession		И		N.	×	var	var	var	var	var	
1	19	1	27.900														
2	18	2	33.770					-	Dependent:	(Statistics						
3	28	2	33.000		🔗 age)		7	🖋 charges		Plots						
4	33	2	22.705		sex 🖉 bm	1	L E	Block 1	of 1		Sano						
5	32	2	28.880		💰 chil	dren		Prev	ious Ne	ext	3 <u>a</u> ve	L					
6	31	1	25.740		💰 sm	oker			Independent(s):		Options	L					
7	46	1	33.440		💰 reg	ion			🛷 age		Style	L					
8	37	1	27.740					•	💑 sex		<u>B</u> ootstrap	-					
9	37	2	29.830						S Dmi	*						_	
10	60	1	25.840						Method: Enter	-	Cinea Linea	r Regressio	on: Statistics			×	
11	25	2	26.220								- Reare	ssion Coe	efficients n	Model fit			
12	62	1	26.290					•	Selection Variable:	Rula	Es	timates		R squared d	hange		
13	23	2	34.400					_		110	Co	nfidence i	intervals	Descriptives			
14	00	1	39.020					•	<u>C</u> ase Labels:		Lev	/el(%): 95	;	Part and par	tial correlations		-11
10	10	2	42.130									variance r	matrix	Collinearity of	liagnostics		
17	F2	2	24.000					4	WLS Welgiji.		Resid	uals					
18	23	2	23 845			0						ushie W-t					
19	56	2	40 300			l	OK	Past	e <u>R</u> eset Cancel Help	p	D	urbin-wats	ingeneration				-1
20	30	2	35.300		0	2		4	36837 467000			outliers of	iteide:	a stand	ard deviations		
21	60	1	36.005		0	- 1		1	13228.846950				naide.	3 atanu	al a deviarons		— Ļ
21	1		50.000					-				gii cases					
Data View	Variable View												ntinue) Ca	ancel Help			
											IBM SPS	S Statistics	s Processor	is ready	Unicode:ON		

This step input all the variables in dependent and independent block (Figure 39).

Figure 39 Select Variables

ANOVA model generated with model summary (Figure 40). R-Square value and sig (P-value) are used to identify the model performance for this project.

		Мо	del Summary	0		
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin- Watson	
1	.801 ^a	.642	.641	11950.46168	1.724	 }
		Sum of	ANOVA			
Model		Sum of Square	ANOVA ^a f s df	Mean Square	F	Sig.
Model 1	Regression	Sum of Square: 1.636E	ANOVA ^a f df +12 6	Mean Square 2.726E+11	F 1909.060	Sig. .000 ^b
Model 1	Regression	Sum of Square 1.636E 9.139E	ANOVA ^a f df +12 6 +11 6399	Mean Square 2.726E+11 142813534.4	F 1909.060	Sig. .000 ^b
Model 1	Regression Residual Total	Sum of Square 1.636E 9.139E 2.550E	ANOVA ^a f df +12 6 +11 6399 +12 6405	Mean Square 2.726E+11 142813534.4	F 1909.060	Sig. .000 ^b

Figure 40 ANOVA Model

4.3 Tableau Visualisation Generated

Figure 41 selected variables in column and rows, this graph used Sex, Smoker and Charges variables demonstrate the female user charges more insurance premium than male user. Also smoker users paid higher health insurance premium than non-smoker users.



Figure 41 Charge Distribution with Smoker, Sex

The second visualisation (Figure 42) used Children and Charges variables to show how users with different amount children would impact health insurance charge.



Figure 42 Charges Distribution with children

5 Appendix

There are some appendix works tried through this project to achieve the research and subresearch objectives.

Neutral Network generated from RStudio and SPSS.

5.1 Neutral Network Generated Used SPSS

Used SPSS to generate Neutral Network model (Figure 43)

ile <u>E</u> dit	<u>V</u> iew <u>D</u> ata	Iransform	Analyze	Graphs	Utilities	Extensions	Window	Help							
<u>></u> H			Rego Desc	rts riptive Stati	stics	;	***		•						
smoker	1		Bayes	sian Statist	ics									Visible	e: 7 of 7 Vari
	🛷 age	뤚 sex	Table	s			💫 region	🛷 charges	var	var	var	var	var	var	var
1	19	1	Com	are Means			4	16884.924	000						
2	18	2	Gene	ral Linear I	lodel		3	1725.553	2300						
3	28	2	Gene	ralized Line	ear Models		3	4449.462	2000						
4	33	2	Mixed	Models			2	21984.470	0610						
5	32	2	Corre	late		,	2	3866.855	5200						
6	31	1	Rear	assion			3	3756.62	1600						
7	46	1	Logic	haar			3	8240.589	9600						
8	37	1	blour	al blobuorke			-	7004 50	600						
9	37	2	Class	a reetworks	•		Multilaye	r Perceptron	00						
10	60	1	Class	HIY			R Radial B	asis Function	20						
11	25	2	Dime	nsion Red	uction		1	2721.320	0800						
12	62	1	Scale			1	3	27808.725	5100						
13	23	2	Nonp	arametric	rests		4	1826.843	8000						
14	56	1	Fored	asting		*	3	11090.717	7800						
15	27	2	Survi	ral		•	3	39611.757	700						
16	19	2	Multip	le Respon	se		4	1837.23	7000						
17	52	1	Missie Missie	ng Value Ar	nalysis		1	10797.336	6200						
18	23	2	Mulţip	le Imputati	on		1	2395.17	1550						
19	56	2	Com	ex Sampl	es		4	10602.385	5000						
20	30	2	Simul	ation			4	36837.467	000						
21	60	1	Quali	ty Control			1	13228.846	950						
	4		Spati	al and Tem	poral Model	ing +	a second s						_		
ata View V	ariable View		Direc	Marketing		,									

Figure 43 Neutral Network Model Generated

Figure 44 shows the Neutral Network Information of the model. Which including the output layer with six dependent variables and 1 input layer.

	Network Inf	ormation				
Input Layer	Factors	1	charges			
	Number of Units ^a		2600			
Hidden Layer(s)	Number of Hidden Lay	ers	1			
	Number of Units in Hid	1				
	Activation Function	Hyperbolic tangent				
Output Layer	Dependent Variables	1	age			
		2	sex			
		3	bmi			
		4	children			
		5	smoker			
		6	region			
	Number of Units		16			
	Rescaling Method for S	Scale Dependents	Standardized			
1	Activation Function		Identity			
-	Error Function	Error Function				

Figure 44 Network Information

Figure 45 was the Neutral Network model summary and overall result.

Training	Sum of Squares Error		10036.234
	Average Overall Relative Error		1.000
	Percent Incorrect Predictions for Categorical Dependents	sex	38.5%
		children	72.1%
		smoker	36.7%
		region	73.3%
	Relative Error for Scale Dependents	age	1.000
		bmi	1.000
	Stopping Rule Used		1 consecutive step(s) with no decrease in error ^a
	Training Time		0:00:04.14
Testing	Sum of Squares Error		3087.418
	Average Overall Relative Error		1.001
	Percent Incorrect Predictions for Categorical Dependents	sex	38.2%
		children	72.6%
		smoker	36.7%
		region	72.3%
	Relative Error for Scale Dependents	age	1.000
		hmi	1 000

a. Error computations are based on the testing sample.

Overall Percent Correct

Overall Percent	
Correct	
44.9%	
45.0%	

Figure 45 Model Summary and Result





Figure 46 Neutral Network Model Graph

5.2 BMI value compared charge

The figure 47 used the bmi value to compare the charges, the health insurance cost related to high bmi value from the visualization represented.



Figure 47 bmi with Charges

5.3 Regions compared Charges

Figure 48 shown the four regions health insurance charge distribution is quite consistent.





References

Choudhury, P. R., 2015. *analyticsvidhya*. [Online] Available at: <u>https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-programming/</u>

Website, 2018. *DataTechNotes*. [Online] Available at: <u>https://www.datatechnotes.com/2018/10/lasso-least-absolute-shrinkage-and.html</u>

Website, 2019. [Online] Available at: <u>https://www.datatechnotes.com/2019/09/support-vector-regression-example-with.html</u>

Website, n.d. *Random Forest : Walk Through*. [Online] Available at: <u>https://rstudio-pubs-</u> <u>static.s3.amazonaws.com/22159_77fe3f8ed4cb4ab7b3647a8d5c628ca4.html</u>