National College of Ireland

# Configuration Manual

MSc Research Project
Data Analytics


## Bhumi Patel
Student ID: x18114865

School of Computing
National College of Ireland


Supervisor: Manuel Tova-Izquierdo

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**


| | |
|---|---|
| **Student Name:** | Bhumi Patel |
| **Student ID:** | x18114865 |
| **Programme:** | Data Analytics |
| **Year:** | 2018 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Manuel Tova-Izquierdo |
| **Submission Due Date:** | 20/12/2018 |
| **Project Title:** | Configuration Manual |
| **Word Count:** | XXX |
| **Page Count:** | 7 |

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:** | |
| **Date:** | 10th August 2019 |

**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

Bhumi Patel
x18114865

## 1 Introduction

Configuration manual clarify the system setup that needs to be built by making use of hardware and software. It includes screen shots that follow "Evaluate the effectiveness of AMP and non-AMP websites using machine learning algorithms" step-by-step execution.

## 2 R Studio

R 3.5.1 were downloaded for windows 64 bit system. Version 1.1.456 was installed with R language which is run on interface.

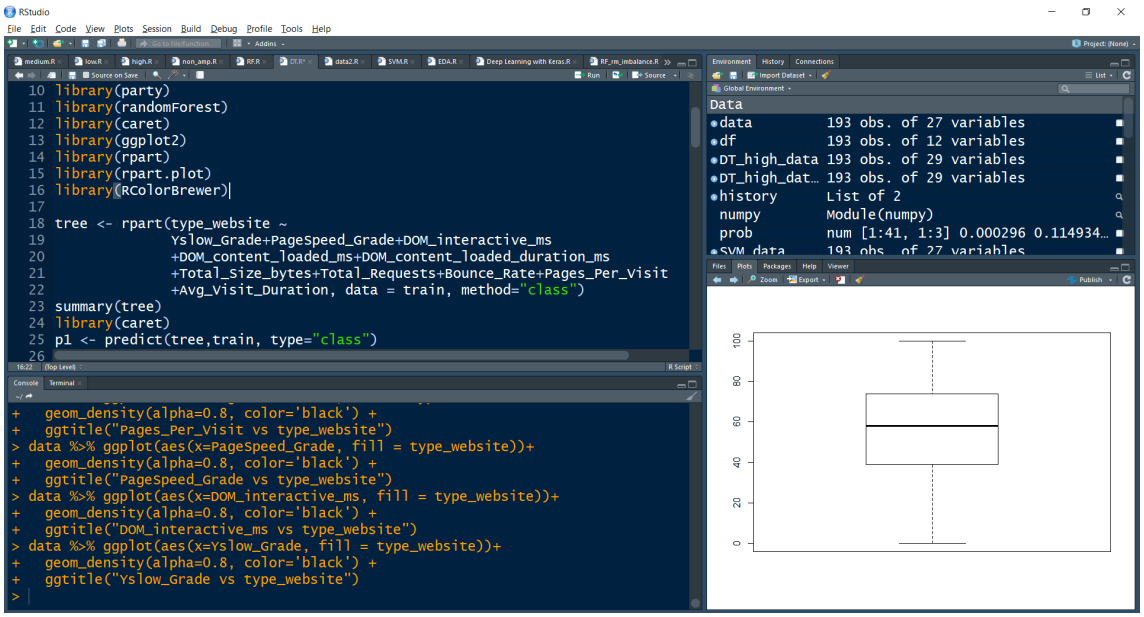



Figure 1: Process Flow Diagram

R studio was used for data pre-processing, performing model classification, and deep learning with keras. (Figure 1) shows the console of R studio while executing model. The following section represent the detail explanation of packages while implementing project.

## 2.1 R packages used for implementation

**pdftools :** Utilities based on 'libpoppler' for extract text, fonts, attachments and metadata from a pdf file (Ooms (2019))

**stringr :** The stringr package offers a cohesive set of features to make it as simple as possible to work with strings (Wickham (2019)).

**dplyr :** dplyr is an easier data handling and data manipulation package. It is designed to operate with information frames directly (Wickham et al. (2019)).

**corrplot :** A graphical representation of a matrix or general matrix for the correlation. It also includes some matrix reordering algorithms (Wei and Simko (2017)).

**randomForest :** RandomForest interfaces that can be applied in a magrittr-implemented pipeline (Liaw and Wiener (2002)).

**caret :** Its stands for Classification And REgression Training which contain the tools such as data splitting, pre-processing, feature selection, model tuning using resampling, and variable importance estimation (from Jed Wing et al. (2019)).

**ggplot2 :** Its create the graphic based on "The Grammar of Graphics. Its used for plotting graph to check the relationship between the variables (Wickham (2016)).

**rpart :** Its stands for Recursive Partitioning and Regression Trees. Its used to builds classification or regression models (Therneau and Atkinson (2018)).

**RColorBrewer :** Its provide A Z Rcolorbrewer Palette. It also used ready-to-use colour palette for graphics (Neuwirth (2014a)).

**party :**Its Laboratory for Recursive Partytioning (Neuwirth (2014b)).

**keras :** This package is used for deep learning (Allaire and Chollet (2019)).

**readr :** Its use it for quick and easy reading of rectangular information like csv, tsv, ets (Wickham et al. (2018)).

**reticulate :** The reticulate package offers an extensive set of interoperability tools between Python and R (Ushey et al. (2019)).

**tidyr :**This package is used to organize data (Wickham and Henry (2019)).

**naivebayes :** This package use for Naive Bayes Classifier. Its High Performance Implementation of the Naive Bayes Algorithm (Majka (2019)).

**Boruta :** This package is used to get confirmed attributes frm the data sets. Kursa and Rudnicki (2010)

# 3 Tools used in the project

## 3.1 WinRAR

WinRAR is used to unzip the zip file of AMP website URL as shown in Figure **??**.

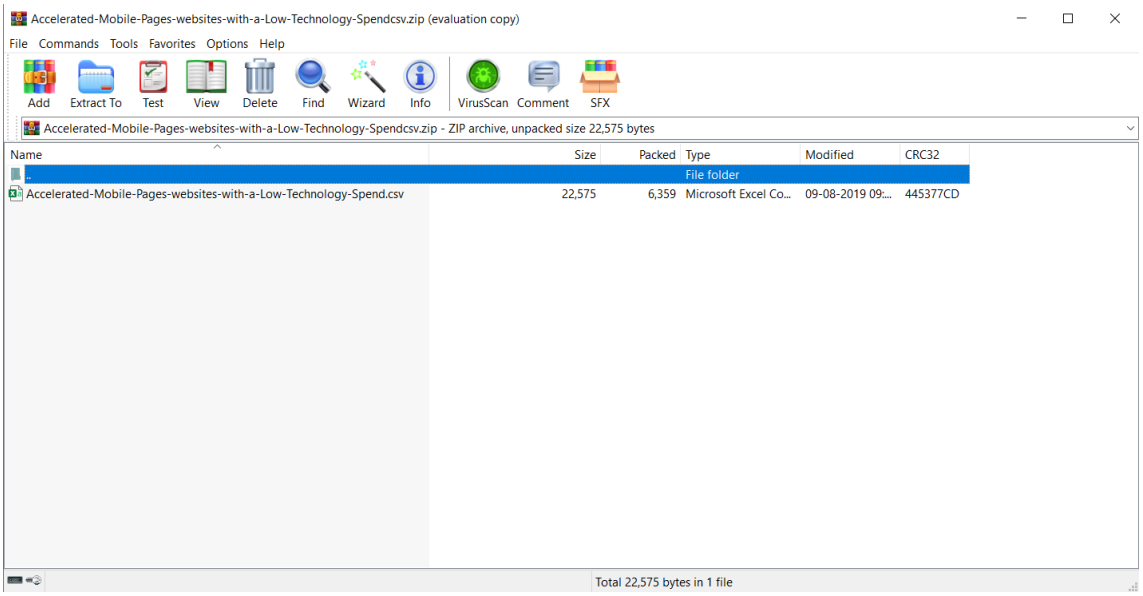
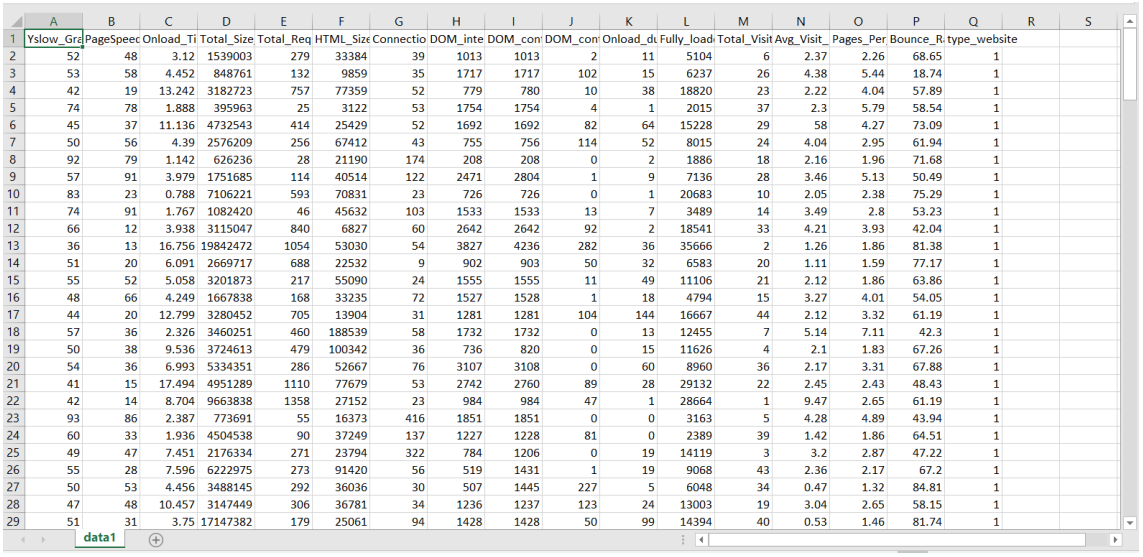



Figure 2: WinRAR

## 3.2 Microsoft Excel

This software is used to saved pre-processing of data into ċsv format as shown in Figure 3.



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Yslow_Gra | PageSpeec | Onload_Ti | Total_Size | Total_Req | HTML_Size | Connectio | DOM_inte | DOM_con | DOM_con | Onload_du | Fully_load | Total_Visit | Avg_Visit_ | Pages_Per | Bounce_R | type_website | | |
| 2 | 52 | 48 | 3.12 | 1539003 | 279 | 33384 | 39 | 1013 | 1013 | 2 | 11 | 5104 | 6 | 2.37 | 2.26 | 68.65 | 1 | | |
| 3 | 53 | 58 | 4.452 | 848761 | 132 | 9859 | 35 | 1717 | 1717 | 102 | 15 | 6237 | 26 | 4.38 | 5.44 | 18.74 | 1 | | |
| 4 | 42 | 19 | 13.242 | 3182723 | 757 | 77359 | 52 | 779 | 780 | 10 | 38 | 18820 | 23 | 2.22 | 4.04 | 57.89 | 1 | | |
| 5 | 74 | 78 | 1.888 | 395963 | 25 | 3122 | 53 | 1754 | 1754 | 4 | 1 | 2015 | 37 | 2.3 | 5.79 | 58.54 | 1 | | |
| 6 | 45 | 37 | 11.136 | 4732543 | 414 | 25429 | 52 | 1692 | 1692 | 82 | 64 | 15228 | 29 | 58 | 4.27 | 73.09 | 1 | | |
| 7 | 50 | 56 | 4.39 | 2576209 | 256 | 67412 | 43 | 755 | 756 | 114 | 52 | 8015 | 24 | 4.04 | 2.95 | 61.94 | 1 | | |
| 8 | 92 | 79 | 1.142 | 626236 | 28 | 21190 | 174 | 208 | 208 | 0 | 2 | 1886 | 18 | 2.16 | 1.96 | 71.68 | 1 | | |
| 9 | 57 | 91 | 3.979 | 1751685 | 114 | 40514 | 122 | 2471 | 2804 | 1 | 9 | 7136 | 28 | 3.46 | 5.13 | 50.49 | 1 | | |
| 10 | 83 | 23 | 0.788 | 7106221 | 593 | 70831 | 23 | 726 | 726 | 0 | 1 | 20683 | 10 | 2.05 | 2.38 | 75.29 | 1 | | |
| 11 | 74 | 91 | 1.767 | 1082420 | 46 | 45632 | 103 | 1533 | 1533 | 13 | 7 | 3489 | 14 | 3.49 | 2.8 | 53.23 | 1 | | |
| 12 | 66 | 12 | 3.938 | 3115047 | 840 | 6827 | 60 | 2642 | 2642 | 92 | 2 | 18541 | 33 | 4.21 | 3.93 | 42.04 | 1 | | |
| 13 | 36 | 13 | 16.756 | 19842472 | 1054 | 53030 | 54 | 3827 | 4236 | 282 | 36 | 35666 | 2 | 1.26 | 1.86 | 81.38 | 1 | | |
| 14 | 51 | 20 | 6.091 | 2669717 | 688 | 22532 | 9 | 902 | 903 | 50 | 32 | 6583 | 20 | 1.11 | 1.59 | 77.17 | 1 | | |
| 15 | 55 | 52 | 5.058 | 3201873 | 217 | 55090 | 24 | 1555 | 1555 | 11 | 49 | 11106 | 21 | 2.12 | 1.86 | 63.86 | 1 | | |
| 16 | 48 | 66 | 4.249 | 1667838 | 168 | 33235 | 72 | 1527 | 1528 | 1 | 18 | 4794 | 15 | 3.27 | 4.01 | 54.05 | 1 | | |
| 17 | 44 | 20 | 12.799 | 3280452 | 705 | 13904 | 31 | 1281 | 1281 | 104 | 144 | 16667 | 44 | 2.12 | 3.32 | 61.19 | 1 | | |
| 18 | 57 | 36 | 2.326 | 3460251 | 460 | 188539 | 58 | 1732 | 1732 | 0 | 13 | 12455 | 7 | 5.14 | 7.11 | 42.3 | 1 | | |
| 19 | 50 | 38 | 9.536 | 3724613 | 479 | 100342 | 36 | 736 | 820 | 0 | 15 | 11626 | 4 | 2.1 | 1.83 | 67.26 | 1 | | |
| 20 | 54 | 36 | 6.993 | 5334351 | 286 | 52667 | 76 | 3107 | 3108 | 0 | 60 | 8960 | 36 | 2.17 | 3.31 | 67.88 | 1 | | |
| 21 | 41 | 15 | 17.494 | 4951289 | 1110 | 77679 | 53 | 2742 | 2760 | 89 | 28 | 29132 | 22 | 2.45 | 2.43 | 48.43 | 1 | | |
| 22 | 42 | 14 | 8.704 | 9663838 | 1358 | 27152 | 23 | 984 | 984 | 47 | 1 | 28664 | 1 | 9.47 | 2.65 | 61.19 | 1 | | |
| 23 | 93 | 86 | 2.387 | 773691 | 55 | 16373 | 416 | 1851 | 1851 | 0 | 0 | 3163 | 5 | 4.28 | 4.89 | 43.94 | 1 | | |
| 24 | 60 | 33 | 1.936 | 4504538 | 90 | 37249 | 137 | 1227 | 1228 | 81 | 0 | 2389 | 39 | 1.42 | 1.86 | 64.51 | 1 | | |
| 25 | 49 | 47 | 7.451 | 2176334 | 271 | 23794 | 322 | 784 | 1206 | 0 | 19 | 14119 | 3 | 3.2 | 2.87 | 47.22 | 1 | | |
| 26 | 55 | 28 | 7.596 | 6222975 | 273 | 91420 | 56 | 519 | 1431 | 1 | 19 | 9068 | 43 | 2.36 | 2.17 | 67.2 | 1 | | |
| 27 | 50 | 53 | 4.456 | 3488145 | 292 | 36036 | 30 | 507 | 1445 | 227 | 5 | 6048 | 34 | 0.47 | 1.32 | 84.81 | 1 | | |
| 28 | 47 | 48 | 10.457 | 3147449 | 306 | 36781 | 34 | 1236 | 1237 | 123 | 24 | 13003 | 19 | 3.04 | 2.65 | 58.15 | 1 | | |
| 29 | 51 | 31 | 3.75 | 17147382 | 179 | 25061 | 94 | 1428 | 1428 | 50 | 99 | 14394 | 40 | 0.53 | 1.46 | 81.74 | 1 | | |

data1

Figure 3: Microsoft Excel

## 3.3 Overleaf

The reporting of the project is carried out using overleaf[1] in latex as shown in Figure 4.

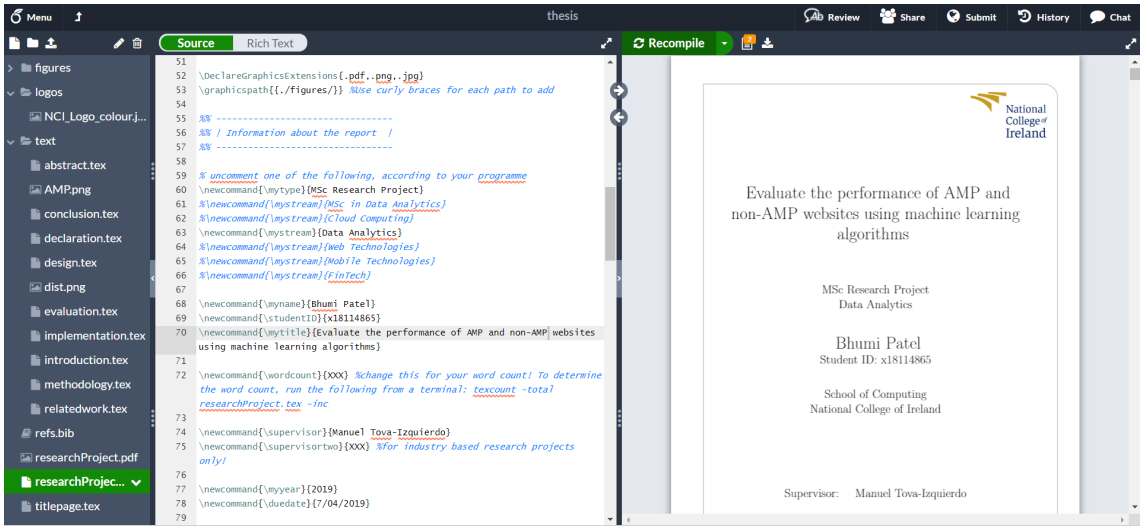



Figure 4: Overleaf

## 3.4 Tableau

Tableau was used for analyzing models outcomes as shown in Figure 5.

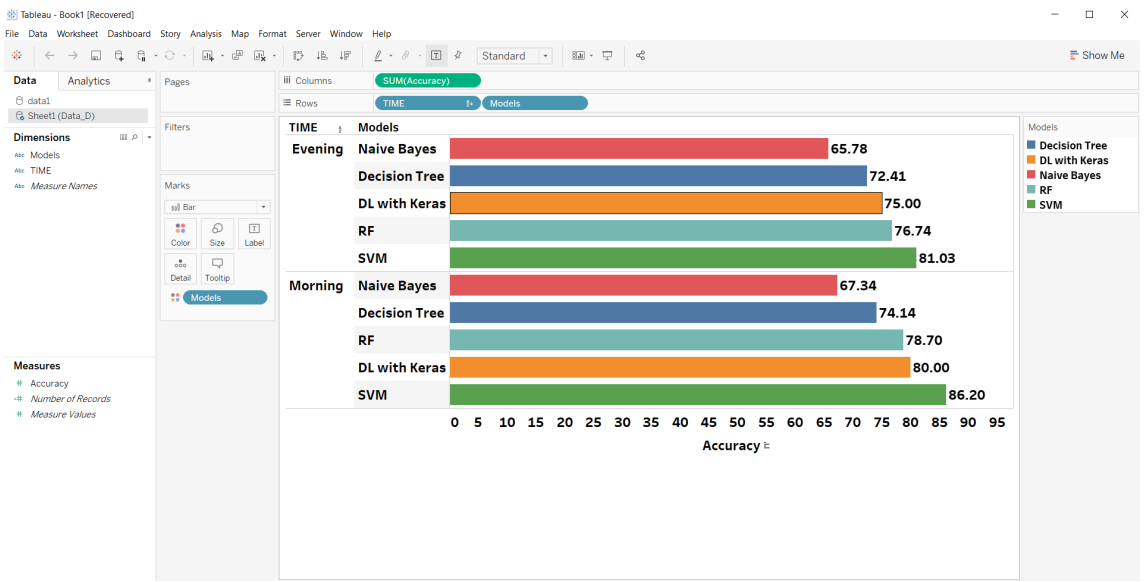



Figure 5: Tableau

[1] https://www.overleaf.com/project/5d307a2c02272418ffd3e2dd

## 3.5 GTmetrix

GTmetrix tool are used to fetched the data as shown in Figure 6.

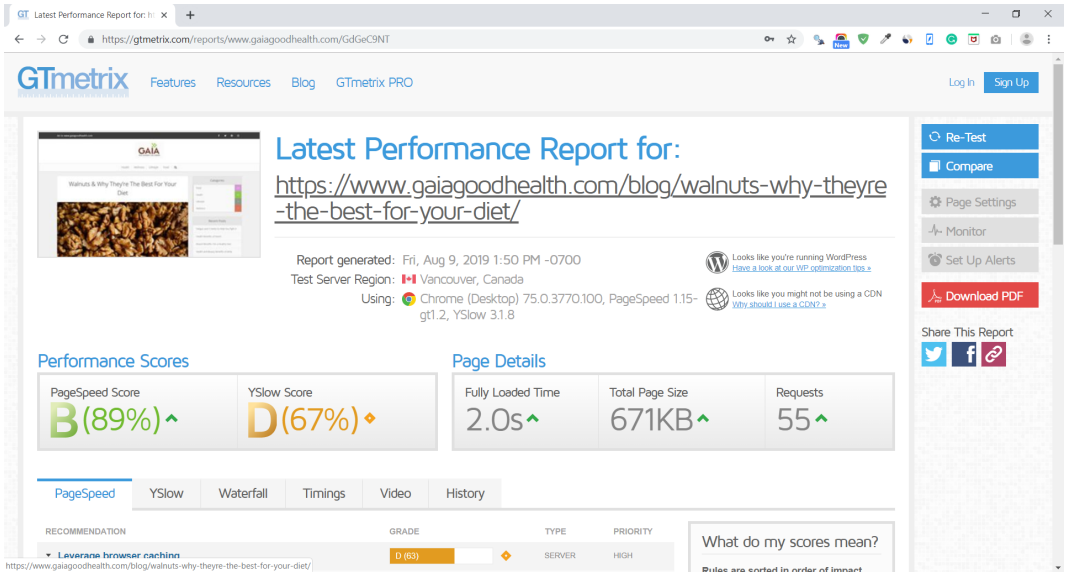



Figure 6: GTmetrix

## 3.6 Accessing project from ICT solutions

Figure 7 shows 3 folders : Data-preprocessing, Models, and Datasets. Data-preprocessing consist of R code file of Exploratory Data Analysis(EDA) and cleaning, Models consist of machine learning algorithms, and Data sets consist of AMP and non-AMP data.

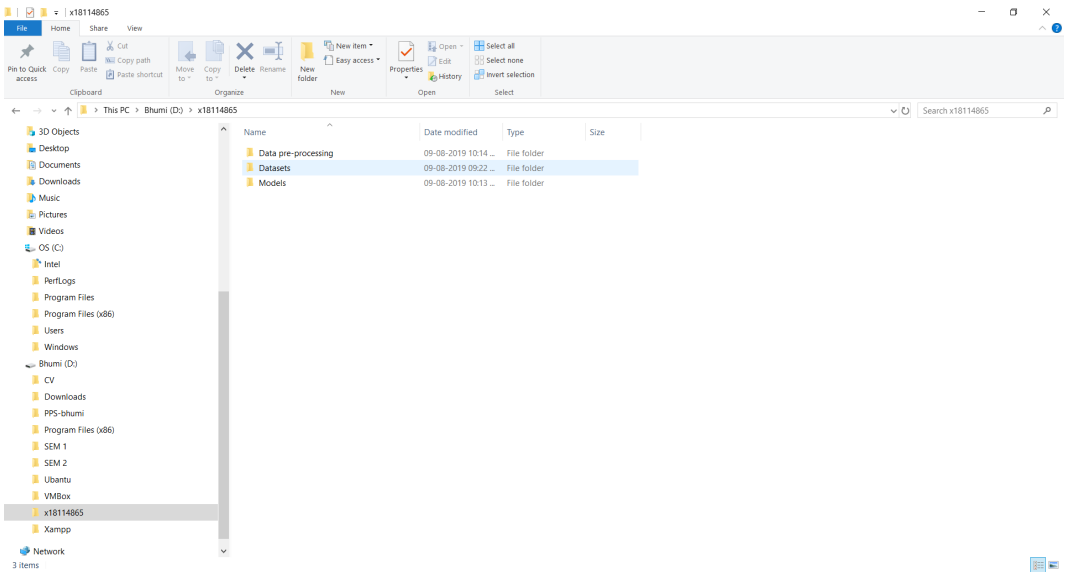



Figure 7: Guidelines to access project from ICT solutions

# 4 System Specifications

- **Processor :** Intel(R) Core(TM) i5-7300HQ CPU @ 2.50GHz
- **Installed Memory (RAM) :** 8 GB
- **GPU ;** NVIDIA GEFORCE GTX 1050
- **Operating System :** Windows 10 64-bit

# References


Allaire, J. and Chollet, F. (2019). *keras: R Interface to 'Keras'*. R package version 2.2.4.1.
**URL:** *https://CRAN.R-project.org/package=keras*

from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. and Hunt., T. (2019). *caret: Classification and Regression Training*. R package version 6.0-82.
**URL:** *https://CRAN.R-project.org/package=caret*

Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the Boruta package, *Journal of Statistical Software* **36**(11): 1–13.
**URL:** *http://www.jstatsoft.org/v36/i11/*

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest, *R News* **2**(3): 18–22.
**URL:** *https://CRAN.R-project.org/doc/Rnews/*

Majka, M. (2019). *naivebayes: High Performance Implementation of the Naive Bayes Algorithm*. R package version 0.9.4.
**URL:** *https://CRAN.R-project.org/package=naivebayes*

Neuwirth, E. (2014a). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.
**URL:** *https://CRAN.R-project.org/package=RColorBrewer*

Neuwirth, E. (2014b). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.
**URL:** *https://CRAN.R-project.org/package=RColorBrewer*

Ooms, J. (2019). *pdftools: Text Extraction, Rendering and Converting of PDF Documents*. R package version 2.2.
**URL:** *https://CRAN.R-project.org/package=pdftools*

Therneau, T. and Atkinson, B. (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13.
**URL:** *https://CRAN.R-project.org/package=rpart*

Ushey, K., Allaire, J. and Tang, Y. (2019). *reticulate: Interface to 'Python'*. R package version 1.13.
**URL:** *https://CRAN.R-project.org/package=reticulate*

Wei, T. and Simko, V. (2017). *R package "corrplot": Visualization of a Correlation Matrix*. (Version 0.84).
**URL:** *https://github.com/taiyun/corrplot*

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
**URL:** *http://ggplot2.org*

Wickham, H. (2019). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0.
**URL:** *https://CRAN.R-project.org/package=stringr*

Wickham, H., Franois, R., Henry, L. and Mller, K. (2019). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.0.1.
**URL:** *https://CRAN.R-project.org/package=dplyr*

Wickham, H. and Henry, L. (2019). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.8.3.
**URL:** *https://CRAN.R-project.org/package=tidyr*

Wickham, H., Hester, J. and Francois, R. (2018). *readr: Read Rectangular Text Data*. R package version 1.3.1.
**URL:** *https://CRAN.R-project.org/package=readr*