National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Vivek Kumar
Student ID: x19201885

School of Computing
National College of Ireland

Supervisor: Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Vivek Kumar
Student ID:	x19201885
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Christian Horn
Submission Due Date:	16/08/2021
Project Title:	Configuration Manual
Word Count:	749
Page Count:	6

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:	Vivek Kumar
Date:	15th August 2021

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):	

Student Grade Prediction

Vivek Kumar
x19201885

1 Overview

In this research, I have tried to predict the grades of the student by using using machine learning techniques. The main aim of this research was to investigate the main reasons of student failures and does alcohol impact the grades of the student. this research, we have used Multiple Regression, Stepwise Regression, Logistic Regression, k-Nearest Neighbor and Principal Component Analysis. All of the programming was done using R programming Language. We have four different sections in this file which are Overview, System Requirements, Installation, Implementation and Results.

2 System Requirements

2.1 Hardware Requirements

Minimum Hardware Requirements:
CPU : Intel Pentium 4 1500MHz
Graphics Card : NVIDIA GeForce 6100
Memory : 4 GB
OS : Windows 7

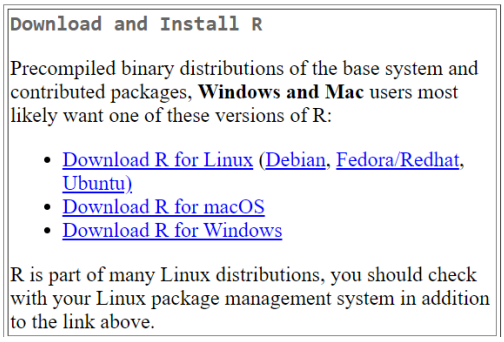
2.2 Software Requirements

This research requires applications like MS Office, R Language (4.1.1) and RStudio Desktop (1.4.1717).

3 Installation

3.1 Softwares

R Language for the specific Operating System needs to be downloaded and installed. In Figure 1, we can see the OS version of R Language to be downloaded. The link for downloading it is *R Programming* (2021).
In Figure 2, we can see the OS version of RStudio to be downloaded. The link for downloading it is *RStudio* (2021)
In Figure 3, we can see the version of MS Office to be downloaded. The link for downloading it is *MS Office* (2021)



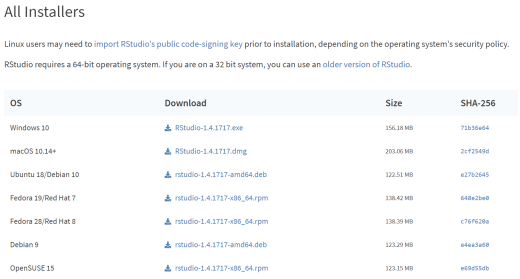
Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Figure 1: Download R Language



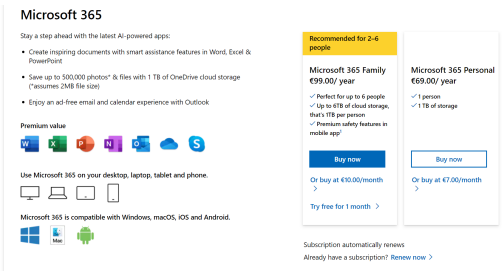
All Installers

Linux users may need to import RStudio's public code-signing key prior to installation, depending on the operating system's security policy.

RStudio requires a 64-bit operating system. If you are on a 32 bit system, you can use an older version of RStudio.

OS	Download	Size	SHA-256
Windows 10	RStudio-1.4.1717.exe	156.18 MB	71b36e64
macOS 10.14+	RStudio-1.4.1717.dmg	203.06 MB	2cf2549d
Ubuntu 18/Debian 10	rstudio-1.4.1717-amd64.deb	122.51 MB	e27b2645
Fedora 19/Red Hat 7	rstudio-1.4.1717-x86_64.rpm	138.42 MB	648e2be0
Fedora 28/Red Hat 8	rstudio-1.4.1717-x86_64.rpm	138.39 MB	c76f628a
Debian 9	rstudio-1.4.1717-amd64.deb	123.29 MB	e4ea3a60
OpenSUSE 15	rstudio-1.4.1717-x86_64.rpm	123.15 MB	e69d51db

Figure 2: RStudio version to be Installed



Microsoft 365

Stay a step ahead with the latest AI-powered apps:

- Create inspiring documents with smart assistance features in Word, Excel & PowerPoint
- Save up to 500,000 photos* & files with 1 TB of OneDrive cloud storage (*assumes 2MB file size)
- Enjoy an ad-free email and calendar experience with Outlook

Premium value

Use Microsoft 365 on your desktop, laptop, tablet and phone.

Microsoft 365 is compatible with Windows, macOS, iOS and Android.

Recommended for 2-6 people

Microsoft 365 Family
€99.00/ year

✓ Perfect for up to 6 people
✓ Up to 6TB of cloud storage, that's 1TB per person
✓ Premium safety features in mobile app

Buy now

Or buy at €10.00/month

Try free for 1 month

Microsoft 365 Personal
€69.00/ year

✓ 1 person
✓ 1 TB of storage

Buy now

Or buy at €7.00/month

Subscription automatically renews

Already have a subscription? Renew now

Figure 3: MS Office Installation

3.2 Install R Libraries

In addition to RStudio, the libraries shown in Figure 4 needs to be installed for this project.

```
install.packages('caTools')
install.packages("class")
install.packages("dummies")
install.packages("lattice")
install.packages("ggplot2")
```

Figure 4: Additional Packages to be installed

4 Implementation

In this section, we can see the steps we followed in creating student grade prediction models and results we get from these models implementation.
We have used following files for implementation:
MulR.R,KNN.R,LogR.R and PCA.R.

4.1 Data Selection

We have downloaded the dataset from Kaggle dataset repository. The 'Student Alcohol Consumption' dataset can be seen in Figure 5 which we have used in our project.

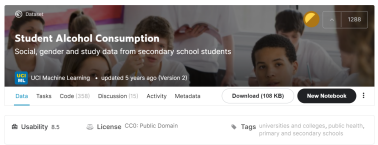


Figure 5: Student Alcohol Consumption Dataset

4.2 Load Dataset

We have imported both Maths and Portuguese students comma-separated value (CSV) into RStudio as shown in Figure 6. First we have imported both CSV files into our RStudio environment and merged them into one file.

Figure 6: Import dataset into RStudio

4.3 Data Pre-processing

After importing the dataset we have checked the dataset for null values using "is.null" function but no null values were found in the dataset as shown in Figure 7.

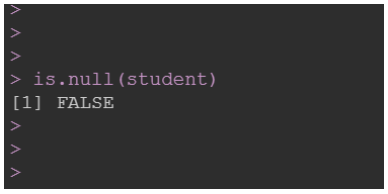


Figure 7: Checking dataset for Null values

4.4 Data Transformation

We transformed each variable and recorded them as factor as it was useful for some classification and regression models. The transformation can be seen in Figure 8.

Figure 8: Variables Encoding

4.5 Dataset Splitting

After all the transformation and pre-processing, we have splitted our dataset into 80:20 ratio where 80% of the dataset had been used for the training purpose and 20% of the dataset had been used for the testing purpose. We have used "caTools" library and "set.seed" function for the dataset splitting. As we can see that in the Figure 9.

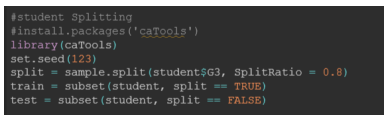


Figure 9: Dataset Splitting

4.6 Modelling

In this step we'll implement our Multiple Regression, Stepwise Regression, KNN, Logistic Regression and Logistic Regression with PCA models.
First we have implemented Multiple Regression and Stepwise Regression. We used Multiple and Stepwise Regression to choose the most significant variables from our dataset. We can shoe that in the Figure 10.

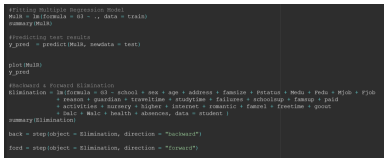


Figure 10: Multiple and Stepwise Regression

After choosing the significant variables with the help of Multiple and Stepwise regression, we implemented k-nearest Neighbor model with 5 as the value of K which we can see in the Figure 11.

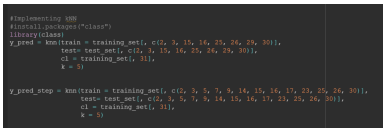


Figure 11: K-nearest Neighbor

After K-nearest Neighbor, we implemented Logistic Regression but instead of applying it only onto the significant variables, In logistic Regression modelling we implemented it onto the full dataset which is shown in the Figure 12.

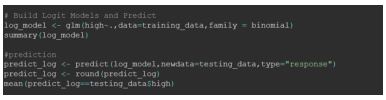


Figure 12: Logistic Regression

And after Logistic Regression, we implemented Logistic Regression onto the Principal Components. First we added 50 new dummy variables into our dataset using "Dummies" libraries and then applied Principal Components onto them and then applied Logistic Regression onto first 15 Principal Components. We performed Principal Component Analysis to check if that made any difference to the normal modelling results. The modelling can be seen in the Figure 13.

Figure 13: Logistic Regression with Principal Components

5 Other Software Used

We have used the overleaf for the documentation of the website. We can see in Figure 14, how Overleaf can be used for documentation.

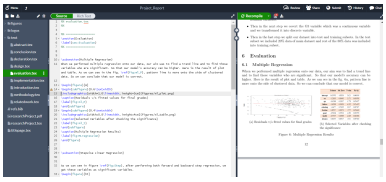


Figure 14: Overleaf

References

MS Office (2021). https://www.microsoft.com/en-ie/microsoft-365/microsoft-office. Windows Version.

R Programming (2021). https://mirror.niser.ac.in/cran. Windows Version.

RStudio (2021). https://www.rstudio.com/products/rstudio/download/#download. Windows Version.