Configuration Manual

MSc Research Project
Master of Science in Data Analytics

Emma O’Hanlon

Student ID: 19210451

School of Computing
National College of Ireland

Supervisor: Zahid Iqbal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Emma O’Hanlon

Student ID: X19210451

Programme: Master of Science in Data Analytics **Programme:** Master of Science in Data Analytics

Module: MSc Research Project

Supervisor: Zahid Iqbal

Submission Due Date: 12/12/2022

Project Title: Using Supervised Machine Learning to Predict the Final Rankings of the 2021 Formula One

Word Count: 552 **Page Count:** 3

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: 12/12/2022

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

Emma O’Hanlon
Student ID: 19210451

1 Introduction

This configuration manual is to accompany the research paper entitled “Using Supervised Machine Learning to Predict the Final Rankings of the 2021 Formula One Championship”. All relevant information about the software and hardware used in the research project can be found below. It will also outline any libraries and technologies that were imported in Python and R Studio and show to configure a working environment. The aim of this manual is to allow reproducibility to the end users.

2 Hardware Specifications

The entirety of this project is carried out using the following hardware as seen in Table 1.

Table 1 - Hardware Specification

Hardware	**Configuration**
System Model	HP Pavilion Laptop
Operation System	Windows 11 Home
Processor	11th Gen Intel(R) Core (TM) i5-1155G7
RAM	8.00 GB

3 Creating a Working Environment

There are two programming languages used in this project, Python and R Programming Language. For the latest version of Python, the 64-Bit Graphical Anaconda Installer (621 MB) was installed for Windows. [1] For R Programming Language, the latest version of RStudio was downloaded for Windows. [2]
Python is used to gather, clean, explore, transform, and visualise the data. R is used to run the two supervised machine learning models, Multiple Linear Regression (MLR) and Artificial Neural Networks (ANN) for regression. Figure 1 is an illustration of the workflow for this project.

[1] https://www.anaconda.com/products/distribution#download-section

[2] https://posit.co/download/rstudio-desktop/

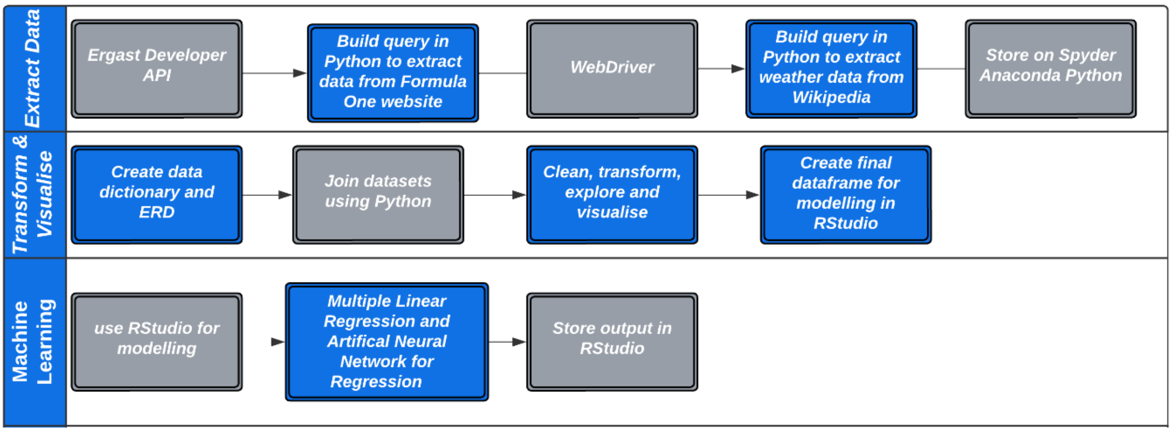
Extract Data
Ergast Developer API
Build query in Python to extract data from Formula One website
WebDriver
Build query in Python to extract weather data from Wikipedia
Store on Spyder Anaconda Python
Transform & Visualise
Create data dictionary and ERD
Join datasets using Python
Clean, transform, explore and visualise
Create final dataframe for modelling in RStudio
Machine Learning
use RStudio for modelling
Multiple Linear Regression and Artifical Neural Network for Regression
Store output in RStudio

Figure 1 – Workflow

4 Datasets

There are two sources used to gather a total of six datasets, the open source Ergast Developer Application Programming Interface (API) and WebDriver from Selenium. Both are run using Python. Five of the six datasets gathered with the API contains Formula One race data from the Formula One official website, while the WebDriver gathers weather information of each race from Wikipedia. The link to the weather data in Wikipedia can be found in the *circuits* dataset.

5 Libraries

Table 2 is a list of all the libraries used for both Python and RStudio

Table 2 - Libraries

Python	**R**
os	dplyr
pandas	corrplot
numpy	car
requests	olsrr
webdriver	keras
re	mlbench
string	magrittr
unicodedata	neuralnet
nltk	
WordNetLemmatizer	
dateutil.relativedelta	
matplotlib.pyplot	
seaborn	
plotly.express	

6 Machine Learning

There were many iterations of the both the MLR and ANN models. The code below shows the code of the final models that were used for the results and comparison piece in the research report.

MLR:

```
model_13_mlr_test <- lm(driver_standings_pos_after_race~ podium:points + round          +
circuit_name:country + circuit_name:driver_name + constructor_name +
                 grid + status + qualifying_best +
                 driver_wins_after_race + constructor_wins_after_race +constructor_standings_pos_after_race +
dry  + cloudy + driver_age + new_time, data = test_df)
```

ANN:

```
model_1_ANN <- keras_model_sequential()
model_1_ANN  %>%
  layer_dense(units = 5, activation = 'relu', input_shape = c(27)) %>%
  layer_dense(units = 1)

#Model Compilation
model_1_ANN  %>% compile(loss = 'mse',
               optimizer = 'rmsprop',
               metrics='mae')

#Model Fitting
set.seed(0)
model_1_ANN_fit <- model_1_ANN %>%
  fit(X_train, y_train,
      epochs = 100,
      batch_size = 32,
      validation_split = 0.2)

#Prediction on train
model_1_ANN %>% evaluate(X_train, y_train)
pred_train <- model_1_ANN %>% predict(X_train)

#Prediction on test
model_1_ANN %>% evaluate(X_test, y_test)
pred <- model_1_ANN %>% predict(X_test)
pred
```

7 Conclusion

This manual explained the main technologies employed throughout the project. It also clarified how those technologies were configured and implemented. This manual will allow the end user to set up their working environment to be able to run the code linked to this project. The results for each user will be identical to those stated in the research paper.