Configuration Manual

MSc Research Project
Data Analytics

Ronan Flannery
Student ID: x19113269

School of Computing
National College of Ireland

Supervisor: Manaz Kaleel

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ronan Flannery
Student ID:	x19113269
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Manaz Kaleel
Submission Due Date:	17/08/2020
Project Title:	Configuration Manual
Word Count:	955
Page Count:	8

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:	11th August 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

Ronan Flannery
x19113269

1 Introduction

This configuration manual describes the software, environments and settings used in the research project "A Machine Learning Approach to Predicting Gross Domestic Product". This document may be used to replicate the technical work carried out in the research project.

2 Hardware Used

This research project was conducted on a Lenovo ideapad 320-15AST with the following configuration:

- **Operating System**: Windows 10
- **Processor**: AMD-A9-9420 RADEON R5, 5 Cores
- **RAM**: 8GB

Google Colab was used for creating and running the models used in this project.

3 Environment

Python was used to create the models in this project. The following environment was used in this research project:

- **Google Colaboratory**

Google Colab allows the user to write and run Python code in an online browser notebook. Colab provides a hosted online Jupyter Notebook service. It has advantages for creating and running machine learning models, particularly Artificial Neural Networks. There are less resource constraints than running models on a local machine and there is no requirement to install Python packages locally. Code written in Google Colab is saved to the users Google Drive and can be easily shared if working on a collaborative project. The following is required to use Google Colab:

- Internet Browser, for example Chrome, Firefox or Safari
- Google Account - A google account is required for using Google Colab

Google Colab is accessible here: `https://colab.research.google.com/notebooks/intro.ipynb`

A new notebook for writing code can be accessed through the File - New notebook option. An existing notebook can also be opened or uploaded via the File menu.

4 Implementation

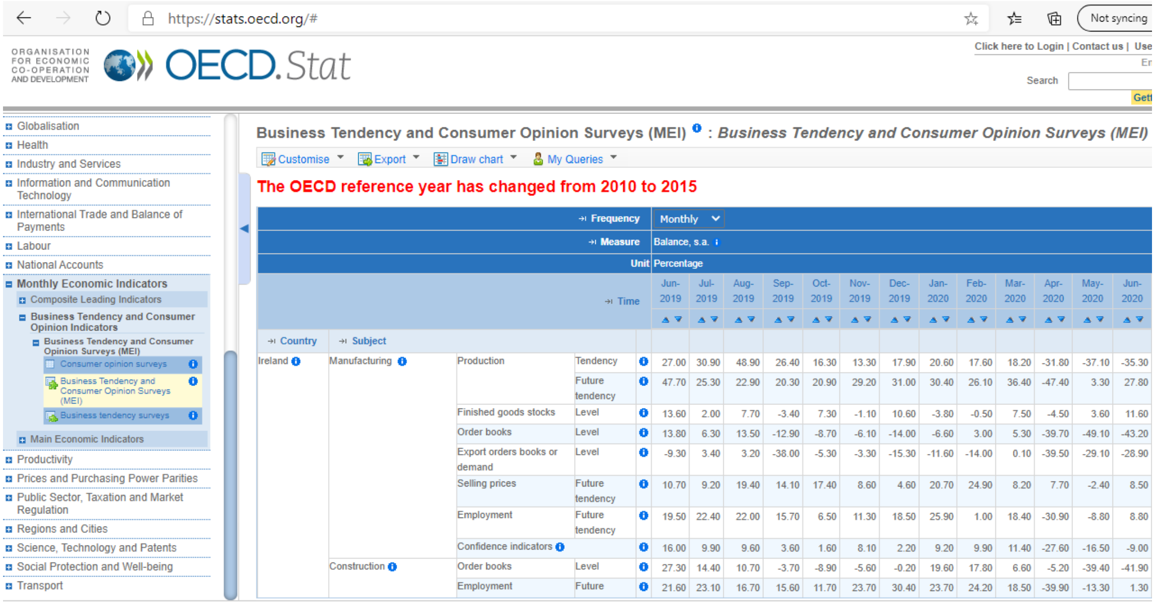
The following section outlines the implementation of the project from a technical perspective.

4.1 Dataset

The data used in this research project was gathered from the OECD data repository. The link for this data source is: `https://stats.oecd.org/`

Figure 1 below shows the OECD data repository. The various indicators used in this projects data set can be selected from the menus on the left of the OECD screen. The data returned can be filtered by country, time period etc. The data can be downloaded in various formats, however in this project the data was downloaded in csv format. Data was taken for 10 different countries (Australia, Austria, Belgium, Denmark, Finland, Germany, Ireland, Korea, Sweden and the UK). Individual files for the economic indicators were downloaded, including Current and Constant GDP, Exchange Rates, Consumer Confidence Indicators and Business Confidence Indicators, Long Term and Short Term Interest Rates, Unemployment Rate, Exports, Government Expenditure, Private Consumption, Imports.

Figure 1: OECD Data Repository



https://stats.oecd.org/#
Not syncing
ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT
OECD.Stat
Click here to Login | Contact us | Use
Search
Gett
Globalisation
Health
Industry and Services
Information and Communication Technology
International Trade and Balance of Payments
Labour
National Accounts
Monthly Economic Indicators
Composite Leading Indicators
Business Tendency and Consumer Opinion Indicators
Business Tendency and Consumer Opinion Surveys (MEI)
Consumer opinion surveys
Business Tendency and Consumer Opinion Surveys (MEI)
Business tendency surveys
Main Economic Indicators
Productivity
Prices and Purchasing Power Parities
Public Sector, Taxation and Market Regulation
Regions and Cities
Science, Technology and Patents
Social Protection and Well-being
Transport
Customise
Export
Draw chart
My Queries

Business Tendency and Consumer Opinion Surveys (MEI) : *Business Tendency and Consumer Opinion Surveys (MEI)*

The OECD reference year has changed from 2010 to 2015

Frequency: Monthly

Measure: Balance, s.a.

Unit: Percentage

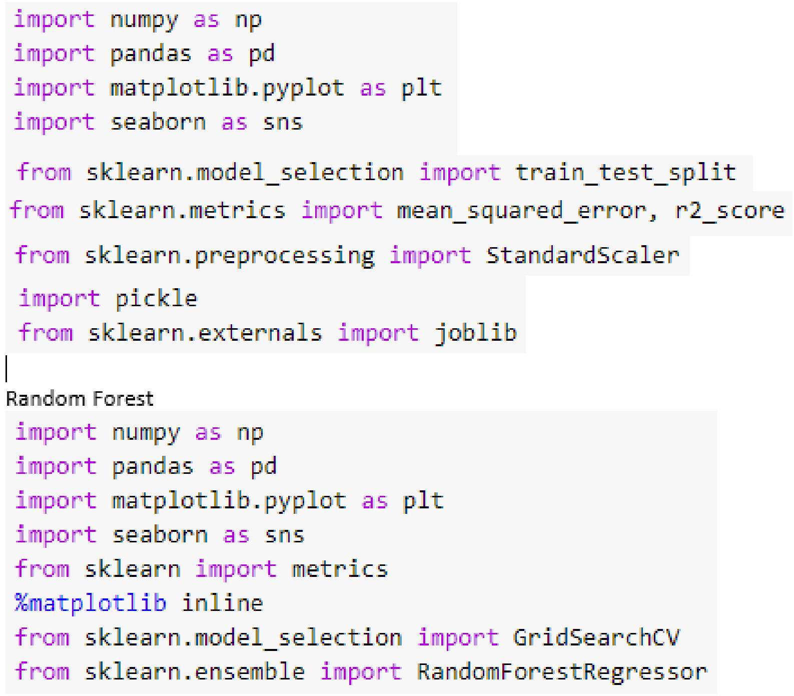
Country	Subject			Jun-2019	Jul-2019	Aug-2019	Sep-2019	Oct-2019	Nov-2019	Dec-2019	Jan-2020	Feb-2020	Mar-2020	Apr-2020	May-2020	Jun-2020
Ireland	Manufacturing	Production	Tendency	27.00	30.90	48.90	26.40	16.30	13.30	17.90	20.60	17.60	18.20	-31.80	-37.10	-35.30
			Future tendency	47.70	25.30	22.90	20.30	20.90	29.20	31.00	30.40	26.10	36.40	-47.40	3.30	27.80
		Finished goods stocks	Level	13.60	2.00	7.70	-3.40	7.30	-1.10	10.60	-3.80	-0.50	7.50	-4.50	3.60	11.60
		Order books	Level	13.80	6.30	13.50	-12.90	-8.70	-6.10	-14.00	-6.60	3.00	5.30	-39.70	-49.10	-43.20
		Export orders books or demand	Level	-9.30	3.40	3.20	-38.00	-5.30	-3.30	-15.30	-11.60	-14.00	0.10	-39.50	-29.10	-28.90
		Selling prices	Future tendency	10.70	9.20	19.40	14.10	17.40	8.60	4.60	20.70	24.90	8.20	7.70	-2.40	8.50
		Employment	Future tendency	19.50	22.40	22.00	15.70	6.50	11.30	18.50	25.90	1.00	18.40	-30.90	-8.80	8.80
		Confidence indicators		16.00	9.90	9.60	3.60	1.60	8.10	2.20	9.20	9.90	11.40	-27.60	-16.50	-9.00
	Construction	Order books	Level	27.30	14.40	10.70	-3.70	-8.90	-5.60	-0.20	19.60	17.80	6.60	-5.20	-39.40	-41.90
		Employment	Future	21.60	23.10	16.70	15.60	11.70	23.70	30.40	23.70	24.20	18.50	-39.90	-13.30	1.30

In this project, data showing the dates of national elections in various countries was also sourced. This was used to create a variable during the feature engineering portion of the project. The variable indicated the months when an election was held in the country. This data was sourced from the following links: `http://www.parties-and-elections.eu/countries.html`, `http://elections.uwa.edu.au/index.lasso` and `https://en.wikipedia.org/wiki/Elections_in_South_Korea`

4.2 Google Colab environment setup

In this project Multilayer Perceptron and Random Forest Models were created. Figure 2 below shows the libraries imported for creating these models in Google Colab.

Figure 2: Libraries used for creating the models



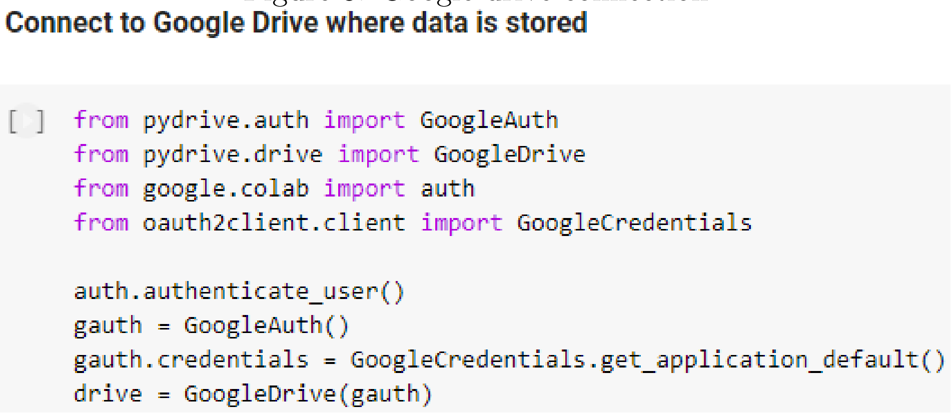
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import StandardScaler
import pickle
from sklearn.externals import joblib
```

Random Forest

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import metrics
%matplotlib inline
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestRegressor
```

The data files used in this project were stored on a google drive. In order to connect to the google drive to load/store files an authentication needs to be made with the google drive. Figure 3 below shows the code for authenticating with the Google Drive. A file can be shared from a google drive by marking it as sharable and creating a file id that can be incorporated in the code.

Figure 3: Google drive connection



Connect to Google Drive where data is stored

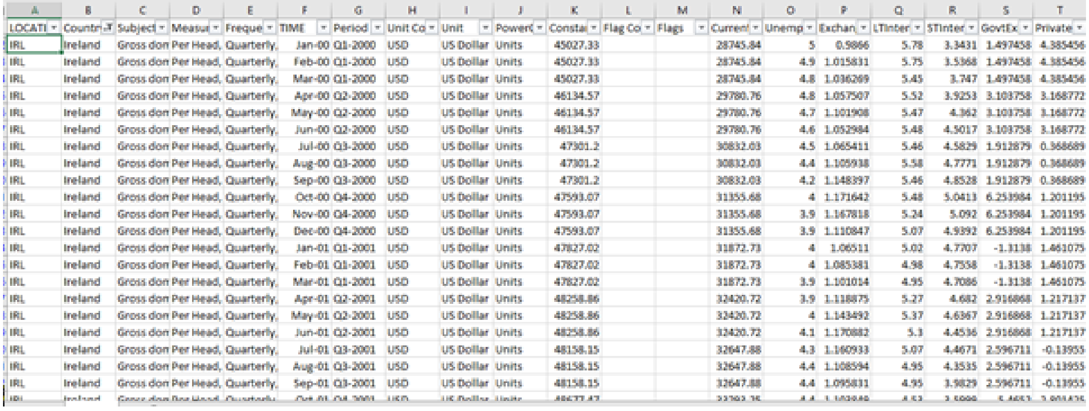
```
[ ] from pydrive.auth import GoogleAuth
    from pydrive.drive import GoogleDrive
    from google.colab import auth
    from oauth2client.client import GoogleCredentials

    auth.authenticate_user()
    gauth = GoogleAuth()
    gauth.credentials = GoogleCredentials.get_application_default()
    drive = GoogleDrive(gauth)
```

4.3 Data Pre-processing

After downloading the data an exploratory data analysis was carried out and the individual data files were combined into a main data file in Microsoft Excel. Figure 4 below shows the combined data file used in this research project. The main data file used in the project contains data for 10 different countries over a period of 20 years (2000 - 2019).

Data pre-processing was carried out in Python. A number of unnecessary columns were removed from the data set and some other data pre-processing steps were performed.

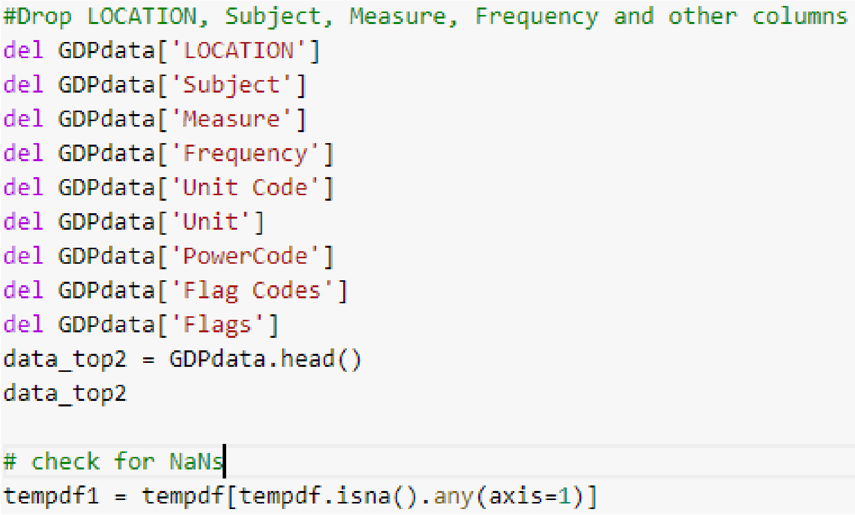
Figure 4: Main data file



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
LOCATI	Countr	Subject	Measu	Freque	TIME	Period	Unit Co	Unit	Power(	Consta	Flag Co	Flags	Current	Unemp	Exchan	LTInter	STInter	GovtEx	Private
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Jan-00	Q1-2000	USD	US Dollar	Units	45027.33			28745.84	5	0.9866	5.78	3.3431	1.497458	4.385456
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Feb-00	Q1-2000	USD	US Dollar	Units	45027.33			28745.84	4.9	1.015831	5.75	3.5368	1.497458	4.385456
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Mar-00	Q1-2000	USD	US Dollar	Units	45027.33			28745.84	4.8	1.036269	5.45	3.747	1.497458	4.385456
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Apr-00	Q2-2000	USD	US Dollar	Units	46134.57			29780.76	4.8	1.057507	5.52	3.9253	3.103758	3.168772
IRL	Ireland	Gross dom	Per Head,	Quarterly,	May-00	Q2-2000	USD	US Dollar	Units	46134.57			29780.76	4.7	1.101908	5.47	4.362	3.103758	3.168772
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Jun-00	Q2-2000	USD	US Dollar	Units	46134.57			29780.76	4.6	1.052984	5.48	4.5017	3.103758	3.168772
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Jul-00	Q3-2000	USD	US Dollar	Units	47301.2			30832.03	4.5	1.065411	5.46	4.5829	1.912879	0.368689
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Aug-00	Q3-2000	USD	US Dollar	Units	47301.2			30832.03	4.4	1.105938	5.58	4.7771	1.912879	0.368689
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Sep-00	Q3-2000	USD	US Dollar	Units	47301.2			30832.03	4.2	1.148397	5.46	4.8528	1.912879	0.368689
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Oct-00	Q4-2000	USD	US Dollar	Units	47593.07			31355.68	4	1.171642	5.48	5.0413	6.253984	1.201195
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Nov-00	Q4-2000	USD	US Dollar	Units	47593.07			31355.68	3.9	1.167818	5.24	5.092	6.253984	1.201195
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Dec-00	Q4-2000	USD	US Dollar	Units	47593.07			31355.68	3.9	1.110847	5.07	4.9392	6.253984	1.201195
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Jan-01	Q1-2001	USD	US Dollar	Units	47827.02			31872.73	4	1.06511	5.02	4.7707	-1.3138	1.461075
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Feb-01	Q1-2001	USD	US Dollar	Units	47827.02			31872.73	4	1.085381	4.98	4.7558	-1.3138	1.461075
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Mar-01	Q1-2001	USD	US Dollar	Units	47827.02			31872.73	3.9	1.101014	4.95	4.7086	-1.3138	1.461075
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Apr-01	Q2-2001	USD	US Dollar	Units	48258.86			32420.72	3.9	1.118875	5.27	4.682	2.916868	1.217137
IRL	Ireland	Gross dom	Per Head,	Quarterly,	May-01	Q2-2001	USD	US Dollar	Units	48258.86			32420.72	4	1.143492	5.37	4.6367	2.916868	1.217137
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Jun-01	Q2-2001	USD	US Dollar	Units	48258.86			32420.72	4.1	1.170882	5.3	4.4536	2.916868	1.217137
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Jul-01	Q3-2001	USD	US Dollar	Units	48158.15			32647.88	4.3	1.160933	5.07	4.4671	2.596711	-0.13955
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Aug-01	Q3-2001	USD	US Dollar	Units	48158.15			32647.88	4.4	1.108594	4.95	4.3535	2.596711	-0.13955
IRL	Ireland	Gross dom	Per Head,	Quarterly,	Sep-01	Q3-2001	USD	US Dollar	Units	48158.15			32647.88	4.4	1.095831	4.95	3.9829	2.596711	-0.13955

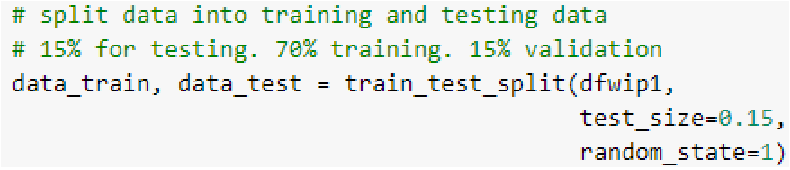
Figures 5 and 6 below show an example of the pre-processing code from Google Colab.

Figure 5: Data Pre-processing



```
#Drop LOCATION, Subject, Measure, Frequency and other columns
del GDPdata['LOCATION']
del GDPdata['Subject']
del GDPdata['Measure']
del GDPdata['Frequency']
del GDPdata['Unit Code']
del GDPdata['Unit']
del GDPdata['PowerCode']
del GDPdata['Flag Codes']
del GDPdata['Flags']
data_top2 = GDPdata.head()
data_top2

# check for NaNs
tempdf1 = tempdf[tempdf.isna().any(axis=1)]
```

Figure 6: Training, test, validation split of the data



```
# split data into training and testing data
# 15% for testing. 70% training. 15% validation
data_train, data_test = train_test_split(dfwip1,
                                         test_size=0.15,
                                         random_state=1)
```

4.4 Models

After the data pre-processing is completed, the the models are created.

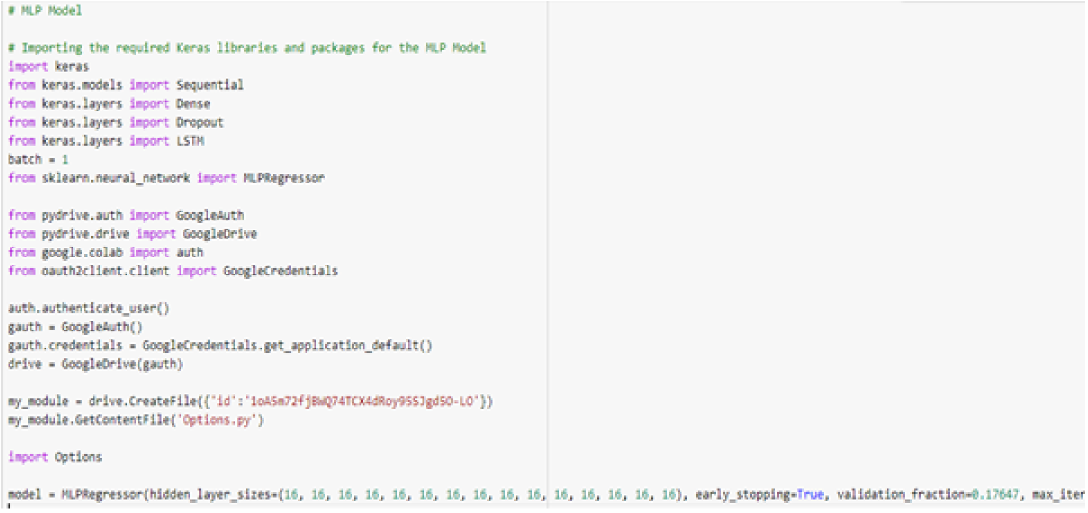
4.4.1 Multi-Layer Perceptron Model

The Multilayer Perceptron model calls individual modules within the code to perform tasks as follows:

- Options.py - contains parameters for the model
- Dataset.py - this contains the code for loading the dataset
- Model.py - this contains code for the MLP model
- plot.py - this shows the final plots created for analysing the output of the model

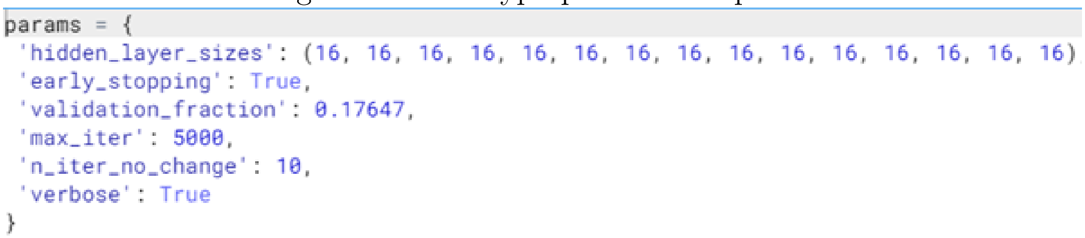
Figure 7 below shows a sample of the code from the implementation of the MLP model. Figure 8 shows some of the hyperparameters configured for the MLP model.

Figure 7: Multi-Layer Perceptron Implementation



```
# MLP Model

# Importing the required Keras libraries and packages for the MLP Model
import keras
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import Dropout
from keras.layers import LSTM
batch = 1
from sklearn.neural_network import MLPRegressor

from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials

auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)

my_module = drive.CreateFile({'id':'1oA5m72fjBWQ74TCX4dRoy95SJgd5O-LO'})
my_module.GetContentFile('Options.py')

import Options

model = MLPRegressor(hidden_layer_sizes=(16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16), early_stopping=True, validation_fraction=0.17647, max_iter
```

Figure 8: MLP hyperparameter Options

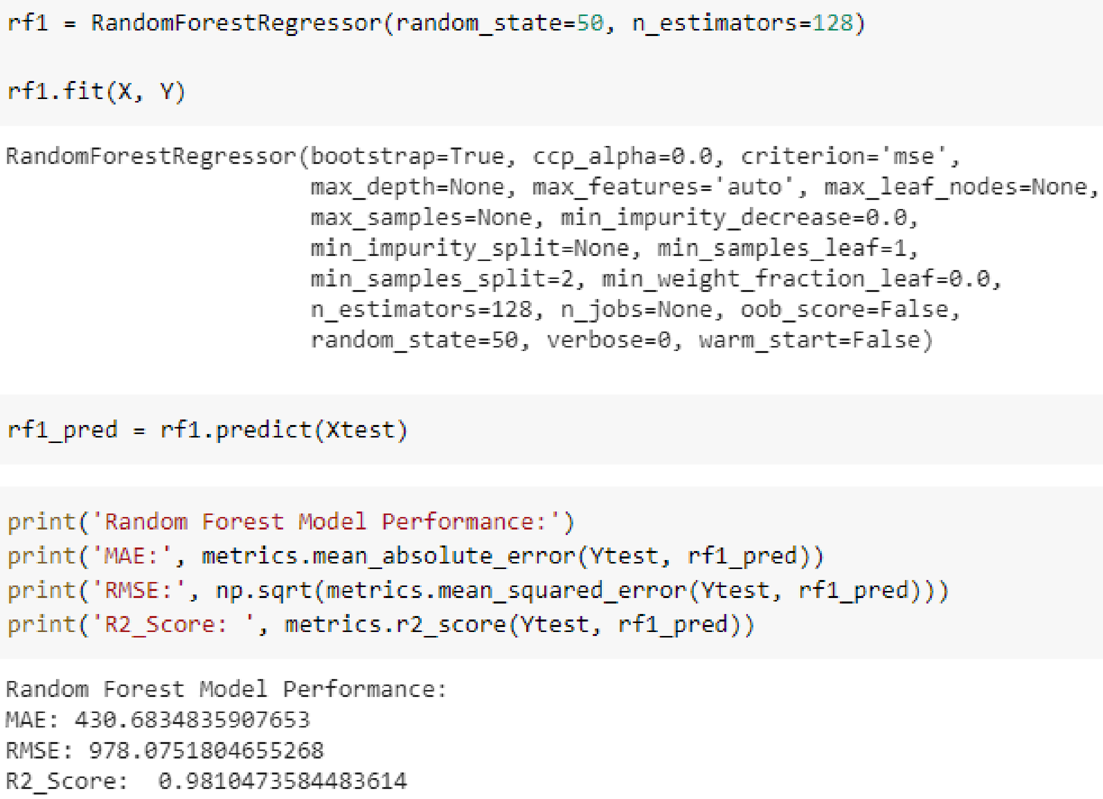


```
params = {
 'hidden_layer_sizes': (16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16)
 'early_stopping': True,
 'validation_fraction': 0.17647,
 'max_iter': 5000,
 'n_iter_no_change': 10,
 'verbose': True
}
```

4.4.2 Random Forest Model

Figure 9 below shows an example of the Random Forest model implementation, including the parameters of the Random Forest Regressor and some of the evaluation metrics.

Figure 9: Random Forest Implementation



```
rf1 = RandomForestRegressor(random_state=50, n_estimators=128)

rf1.fit(X, Y)
```

```
RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                      max_depth=None, max_features='auto', max_leaf_nodes=None,
                      max_samples=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      n_estimators=128, n_jobs=None, oob_score=False,
                      random_state=50, verbose=0, warm_start=False)
```

```
rf1_pred = rf1.predict(Xtest)
```

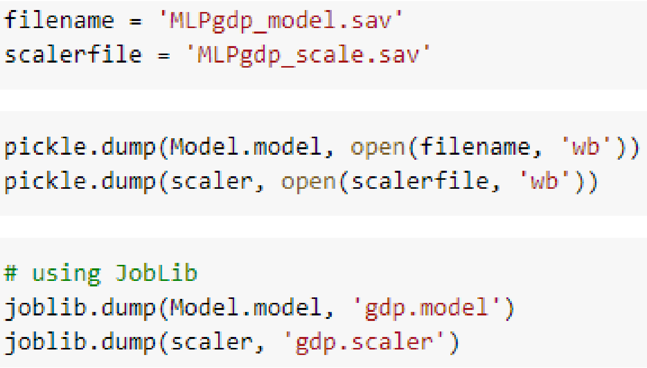
```
print('Random Forest Model Performance:')
print('MAE:', metrics.mean_absolute_error(Ytest, rf1_pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(Ytest, rf1_pred)))
print('R2_Score: ', metrics.r2_score(Ytest, rf1_pred))
```

```
Random Forest Model Performance:
MAE: 430.6834835907653
RMSE: 978.0751804655268
R2_Score:  0.9810473584483614
```

5 Output and Evaluation

This project uses the Python 'pickle' module for saving the MLP model in order to allow it to be called in a separate instance if required. The output of the MLP is also redirected to a log file, which can be used for further evaluation of the training iterations, if required.

Figure 10: Saving the model with pickle

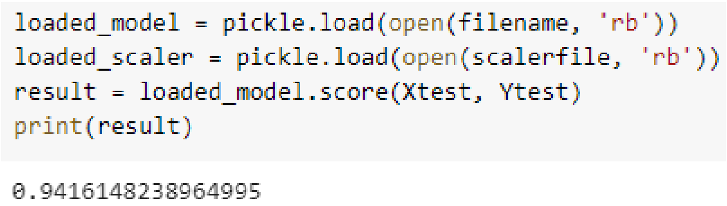


```
filename = 'MLPgdp_model.sav'
scalerfile = 'MLPgdp_scale.sav'
```

```
pickle.dump(Model.model, open(filename, 'wb'))
pickle.dump(scaler, open(scalerfile, 'wb'))
```

```
# using JobLib
joblib.dump(Model.model, 'gdp.model')
joblib.dump(scaler, 'gdp.scaler')
```

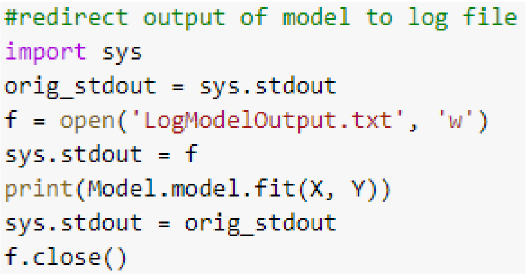
Figure 11: Calling the Model and Displaying result



```
loaded_model = pickle.load(open(filename, 'rb'))
loaded_scaler = pickle.load(open(scalerfile, 'rb'))
result = loaded_model.score(Xtest, Ytest)
print(result)
```

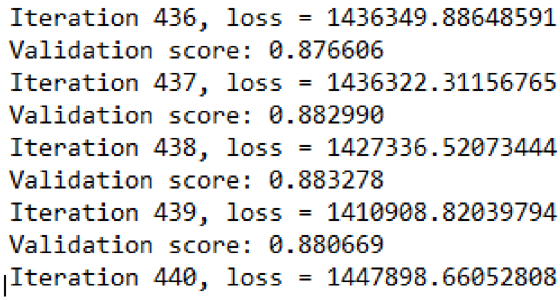
```
0.9416148238964995
```

Figure 12: Redirecting the model output to a log file



```
#redirect output of model to log file
import sys
orig_stdout = sys.stdout
f = open('LogModelOutput.txt', 'w')
sys.stdout = f
print(Model.model.fit(X, Y))
sys.stdout = orig_stdout
f.close()
```

Figure 13: Output log showing the iterations of training

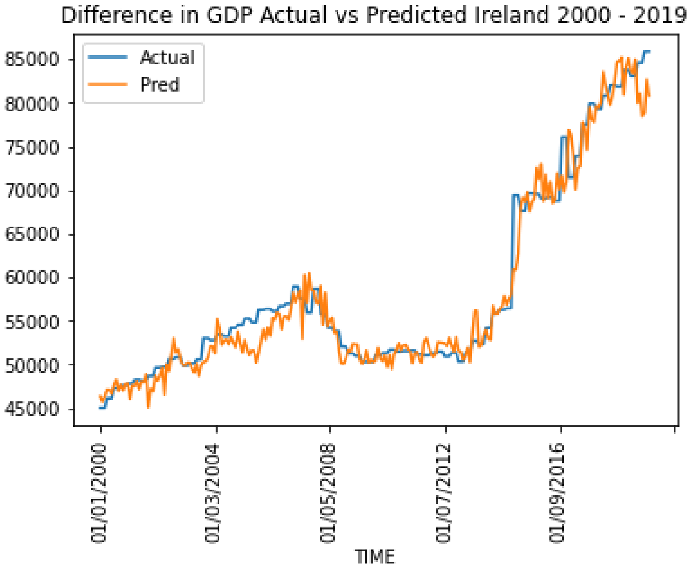


```
Iteration 436, loss = 1436349.88648591
Validation score: 0.876606
Iteration 437, loss = 1436322.31156765
Validation score: 0.882990
Iteration 438, loss = 1427336.52073444
Validation score: 0.883278
Iteration 439, loss = 1410908.82039794
Validation score: 0.880669
Iteration 440, loss = 1447898.66052808
```

5.1 Plots

This research project uses the Python 'matplotlib' library for creating graphs from the output of the models. Various graphs were created from the output, for example graphs showing the predictions of GDP vs the actual GDP figures. Figure 14 below shows an example of one of the output graphs.

Figure 14: Actual vs Predicted GDP MLP model


Difference in GDP Actual vs Predicted Ireland 2000 - 2019
Actual
Pred
01/01/2000
01/03/2004
01/05/2008
01/07/2012
01/09/2016
TIME