National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Zainab Makrani
Student ID: 22190082

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Zainab Makrani
Student ID:	22190082
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	15/12/2023
Project Title:	Configuration Manual
Word Count:	XXX
Page Count:	9

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:	13th December 2023

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

Zainab Makrani
22190082

Contents

1 Introduction

This document discusses how the machine learning models are implemented for prediction of readmissions. The total code consists of 8 ML models which are differentiated between simple ML models and Neural networks. Each model is trined and used for prediction and then evaluated. The code is developed on kaggle IDE and is best to run it there, however, an alternative is to run it on local environment. Each of these ways is discussed in details in this manual.

2 Packages and Libraries

This is contained in the utilities section of the code whic first checks the version of python implemented and neccessary pip installments

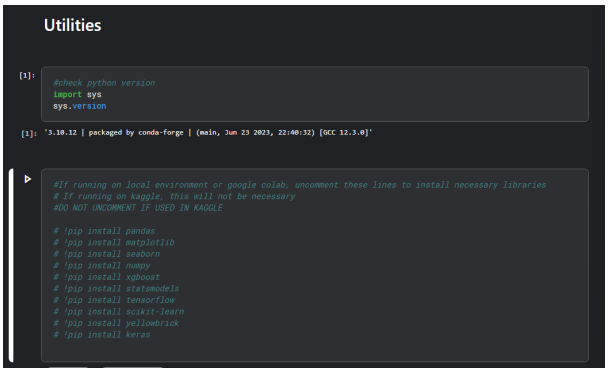

Figure 1: Version and Pip install commands

The utilities also contain the libraries that are later implemented to create necessary machine learning methods.

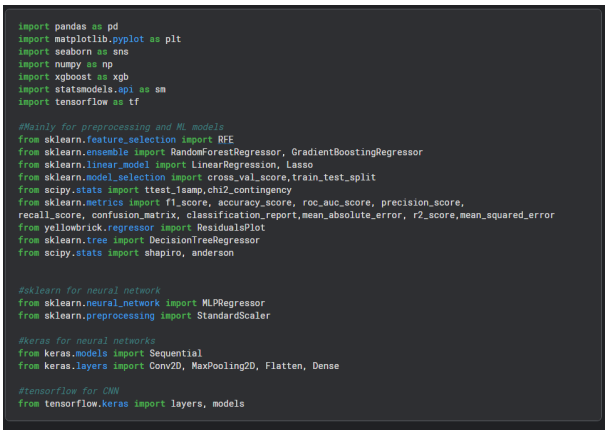

Figure 2: Python Libraries implemented

3 System and Software specification

To build this project on kaggle, u will need to create a kaggle account to run the kaggle notebook [1]. Based on their technical specifications, [2], each notebook is allocated;

1. 12 hours of execution time for CPU and GPU ran sessions while providing 9hrs for TPU sessions.
2. 20GB automatically saving disk space
3. Scratchpad disk space which is not saved outside the current session.

To build this on a local environment, the prerequisites is that python 3 is installed and set up to the path [3].

3.1 Hardware specifications for local environment

These are the device specifications for the local environment in which the research was tested after compilation on kaggle

- Processor Intel(R) Core(TM) i5-8265U
- CPU @ 1.60GHz 1.80 GHz
- Installed RAM 8.00 GB (7.89 GB usable)
- Device ID 6DC9A036-AA3A-4B5D-ACCC-492EA2F9BF3E
- Product ID 00327-30666-87111-AAOEM
- System type 64-bit operating system, x64-based processor
- Pen and touch Touch support with 2 touch points

Windows specification include:

- Edition Windows 11 Home Single Language
- Version 22H2
- OS build 22621.2715
- Experience Windows Feature Experience Pack 1000.22677.1000.0

4 Data Analysis

In this section data is explored and prepared to be used for machine learning. In the EDA, data is loaded and thereafter graphs are made to understand the data. Not all codes are displayed here due to complexity of code but rather just snippets

[1] How to run kaggle notebook. [https://www.kaggle.com/docs/notebooks]
[2] Technical specifications[https://www.kaggle.com/docs/notebookstechnical-specifications]
[3] How to install python 3 [https://www.python.org/downloads/]

4.1 EDA

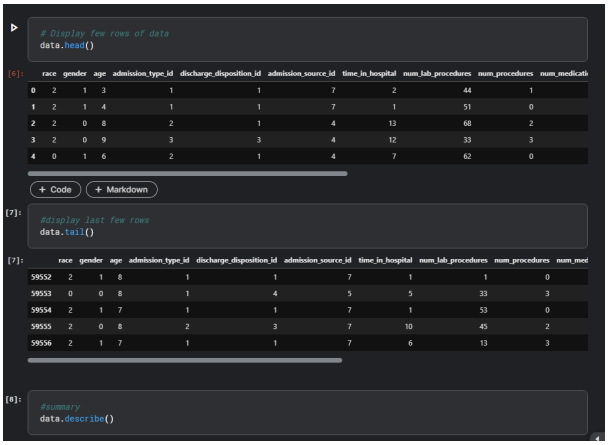

Figure 3: Checking Loaded data in Kaggle

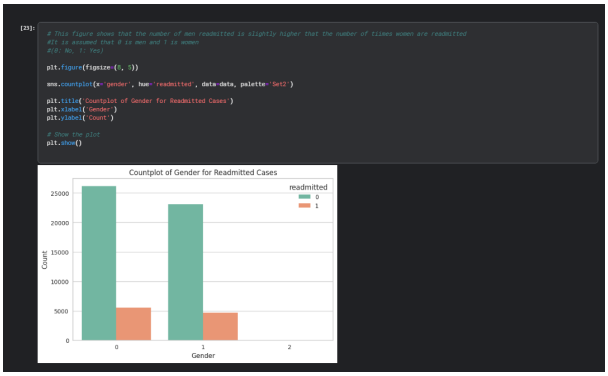

Figure 4: Counterplot to show gender relation to readmitted cases

Figure 5: Barplot to show time spent in hospital vs readmission

4.2 Pre-processing

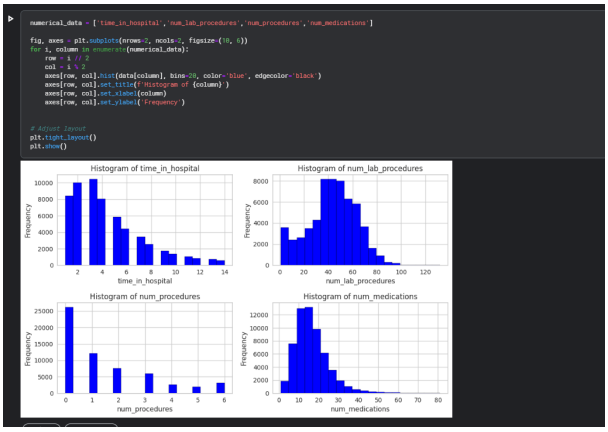

Figure 6: Histogram to check distribution

Figure 7: Check normalization of data

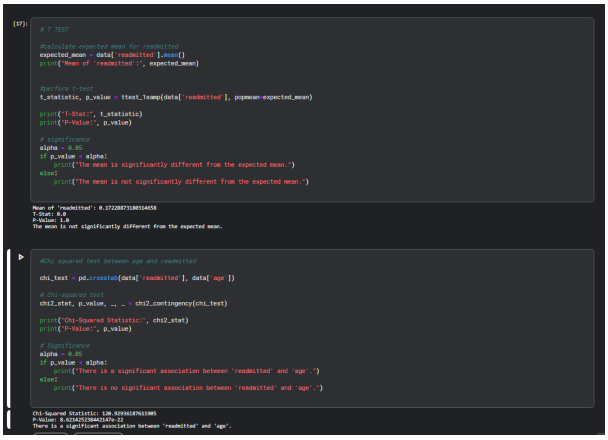

Figure 8: Enter Caption

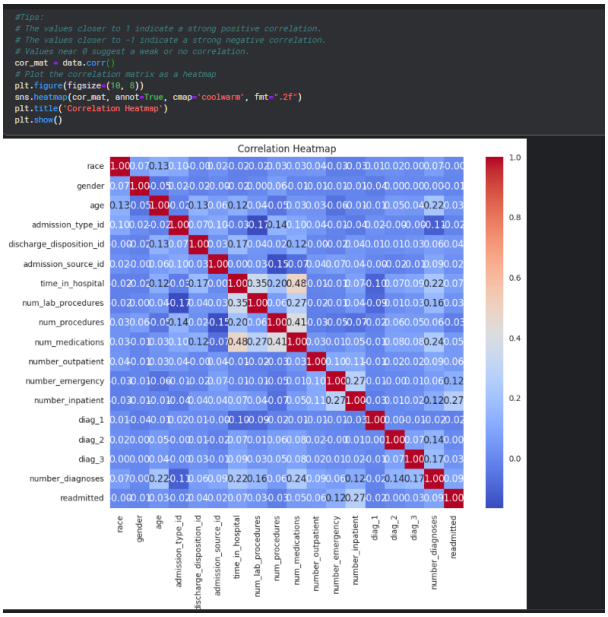

Figure 9: Cor relation matrix

4.3 Feature Engineering

Figure 10: Experimenting with models for RFE

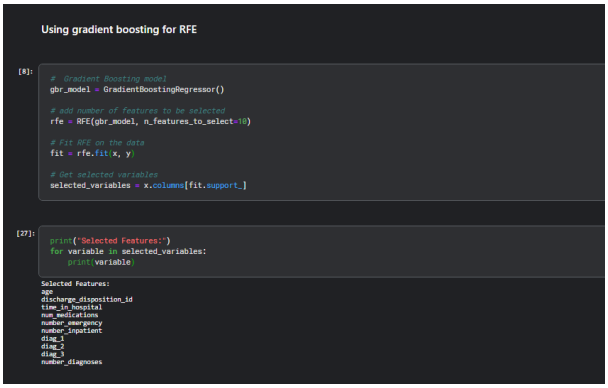

Figure 11: Using gradient boosting for RFE

4.4 Splitting data

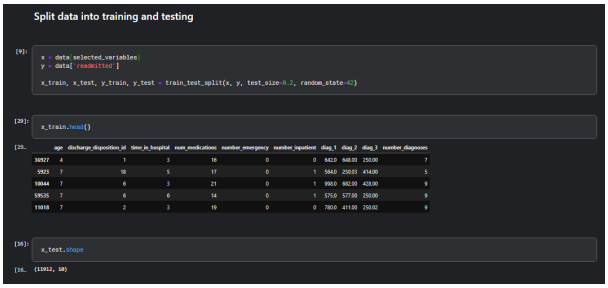

Figure 12: Splitting data into training and testing

5 Simple Machine learning models

There are 5 simple models applied, however as a snippet to show the steps taken in the model creation the best performing from the 5 is displayed here which is random forest.

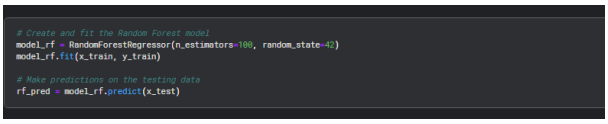

Figure 13: Creating model

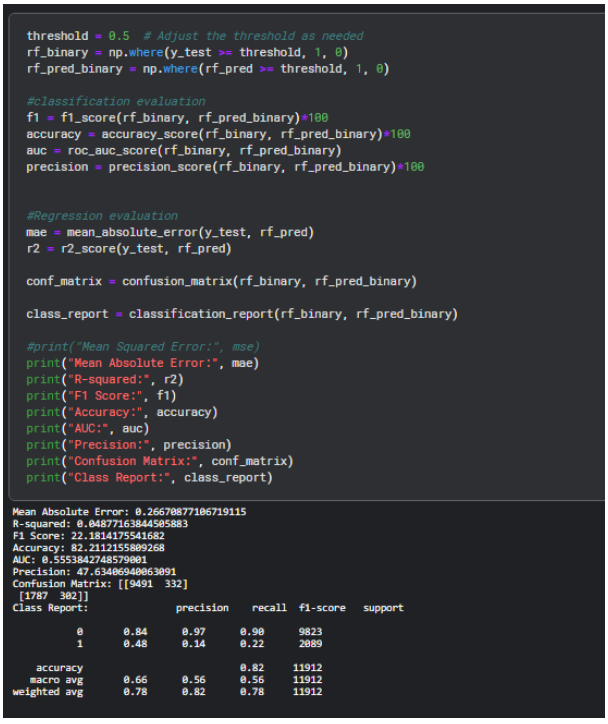

Figure 14: Evaluating Random Forest model

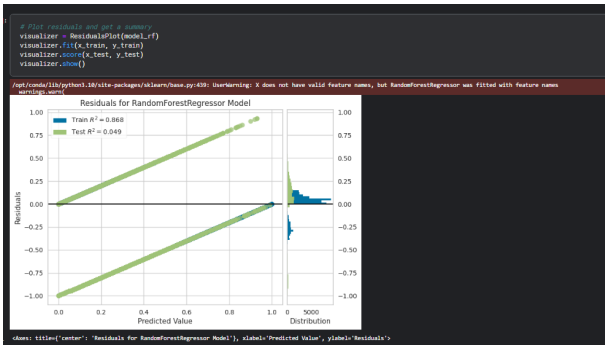

Figure 15: Visualizing the performance of random forest model predictions

6 Neural Networks

Just as done previously, here only the best performing model from the 3 neural networks is displayed which is dense neural network.

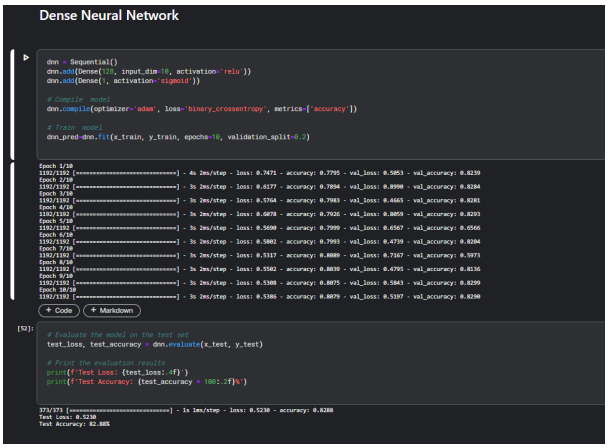

Figure 16: Creating and testing the dense neural network

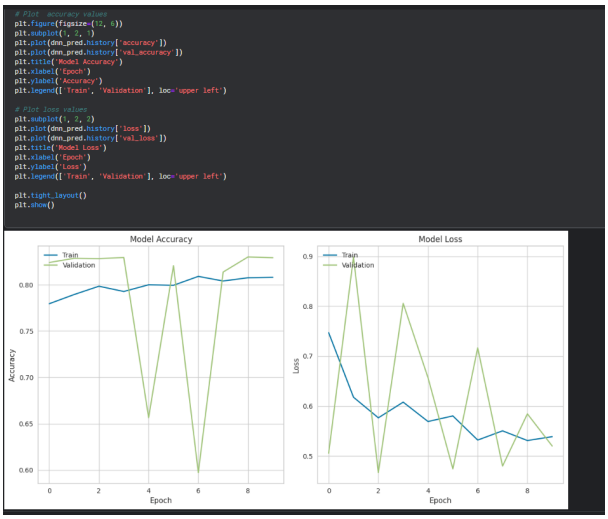

Figure 17: Visualizing the perfomance of the dense neural network

7 How to run the code

The code itself contains comments of how to run and what to uncomment when running in different environments.

7.1 Kaggle Notebook

The pre requisites for this is that, you must create a kaggle account. Thereafter an easy way is to just follow the comments, but since the project is developed in kaggle notebook, you can opt to just hit the run all button after uploading it. It may take a while to run the whole code.

7.2 Jupyter Notebook

As for this, jupyter and python must be installed and correctly set up on your path. After this uncomment the pip install commands as shown below 18 to make sure that the libraries are imported correctly.

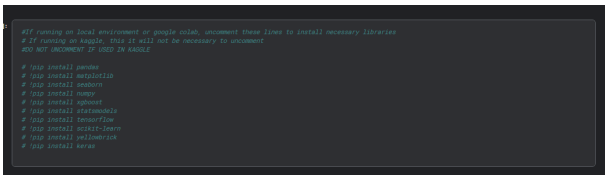

Figure 18: Pip install commands for local envirornment

After this follow the commented instructions and uncomment the first cell in the exploratory data 19. Before doing so make sure to download the dataset from kaggle [4] and that the jupyter notebook and dataset are in the same location.

Figure 19: Uncomment the file path

Once that is done, comment this cell as it is only applicable if used on kaggle notebook to run the codes 20

Figure 20: Comment for local Env

Once the relevant changes have been made you can opt to run all cells in the notebook.

References

[4] Dataset [https://www.kaggle.com/datasets/omnamahshivai/dataset-hospital-readmissions-binary/]