Configuration Manual

MSc Research Project
MSc in Data Analytics

Zaid Siddiqui
Student ID: x23135433

School of Computing
National College of Ireland

Supervisor: Harshani Nagahamulla

Student Name:	Zaid Siddiqui		
Student ID:	x23135433		
Programme:	MSc in Data Analytics	**Year:**	2024
Module:	MSc Research Project		
Lecturer:	Harshani Nagahamulla		
Submission Due Date:	12/12/2024		
Project Title:	Configuration Manual		
Word Count:	XXX **Page Count:** 4		

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.

<u>ALL</u> 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:	Zaid Siddiqui
Date:	12/12/2024

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

Zaid Siddiqui
x23135433

1   Setting up Google Colab

Access Google Colab:

Launch your browser and go to Google Colab.
Use your Google account to log in.
Make a fresh notebook:
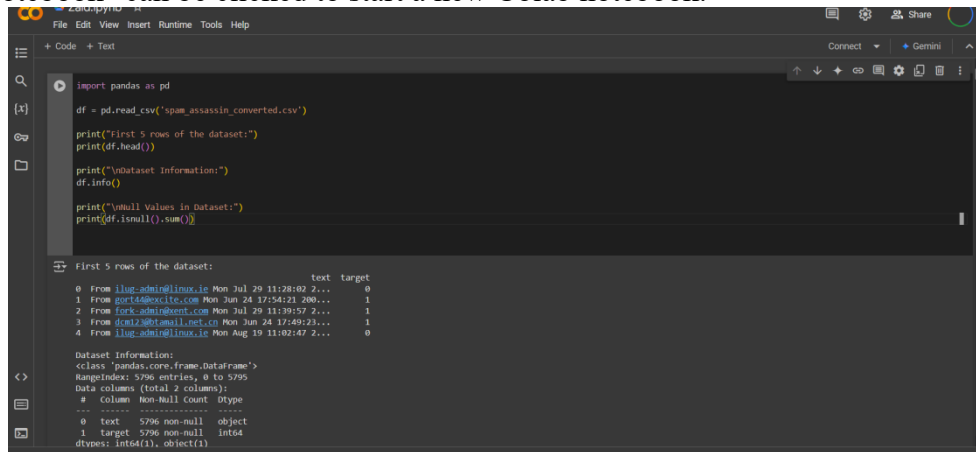"New Notebook" can be clicked to start a new Colab notebook.

Fig 1. Colab Notebook outline

Enable GPU:
You can turn on GPU acceleration for quicker training, particularly for deep learning models.
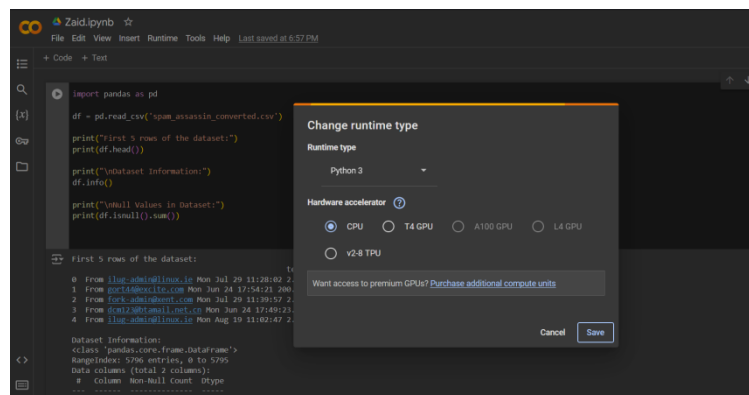Click "Save" after selecting "GPU" under "Hardware accelerator" under "Runtime" -> "Change runtime type."

Fig 2. Run time Environment

2   Setting up the Environment

Install necessary libraries:
To install the necessary libraries, execute the following instructions in a code cell:

Tensorflow torch torch_geometric scikit-learn imblearn!pip install pandas matplotlib seaborn

Libraries:
At the start of your notebook, include the following import statements:

```
import matplotlib
import pandas
Import seaborn
```

3   Loading the Data

Upload the dataset:
Upload the spam_assassin_converted.csv file to your Google Colab environment. You can do this by
Using the file explorer pane on the left side of the Colab interface.
Using the google.colab library:
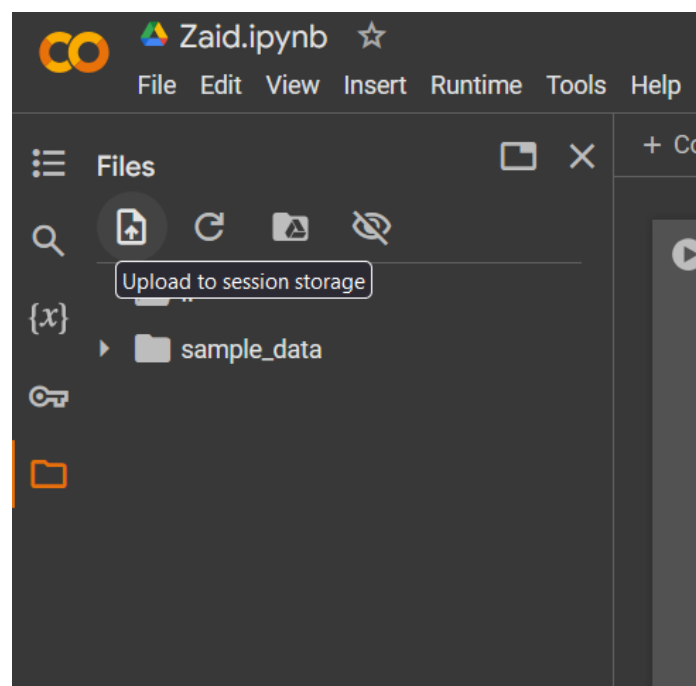from google.colab import files
upl= files.upload()

Fig 3. Colab Notebook File System

Read the data:

Use the following code to read the data into a pandas DataFrame:

dataf = pd.read_csv('spam_assassin_converted.csv')

4   Execution and Customization

Run the code cells:

Execute each code cell sequentially to perform the data preprocessing, analysis, model training, and evaluation.
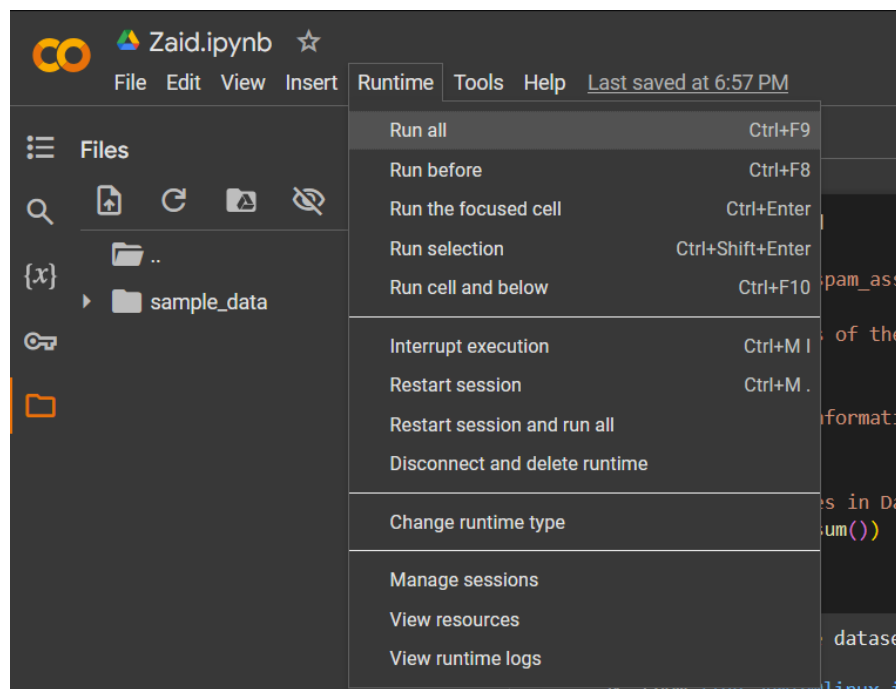

Fig 4. Colab Option to Run all Cells

5 System Software and Hardware Requirements

Software:
Operating System: Windows, macOS, or Linux
Python: Version 3.7 or higher
Libraries: pandas, matplotlib, seaborn, scikit-learn, imblearn, tensorflow, torch, torch_geometric

Hardware:
Processor: A modern multi-core processor (e.g., Intel Core i5 or AMD Ryzen 5)
Memory: 8GB RAM or more (16GB recommended for optimal performance)
Storage: Sufficient disk space to store the dataset and any generated files
GPU: Optional, but highly recommended for faster training of deep learning models (e.g., NVIDIA GeForce GTX 1060 or higher)

6 References

Google Colaboratory. (2024). Welcome to Colaboratory. [Online]. Available: https://colab.research.google.com/
McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56). [Online]. Available: https://pandas.pydata.org/

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95. [Online]. Available: https://matplotlib.org/

Waskom, M. L. (2021). Seaborn: Statistical data visualization. Journal of Open Source Software, 6(60), 3021. [Online]. Available: https://seaborn.pydata.org/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. [Online]. Available: https://scikit-learn.org/stable/

Python Software Foundation. (2024). Python Language Reference, version 3.12. [Online]. Available: https://docs.python.org/3/