# Feb 2023 – SB+ - MSc in Data Analytics

Author: Sajjan Bhattarai

e-mail: [sbs23010@student.cct.ie](mailto:sbs23010@student.cct.ie)

Student ID: sbs23010

# Abstract

# Introduction

# Materials and Methods

## Selection of Programming Language and Libraries

For this study, two programming languages Python and R were considered for their simplicity and availability of wide range of libraries for Data Analytics and Machine Learning. Ultimately, Python was chosen because it provides more human readable syntax, and versatility in terms of object-oriented and modular programming. Additionally, it has a larger ecosystem of libraries and frameworks, that offer more choices and easy integration with other technologies. Therefore, for the purposes of performing statistical analysis, machine learning and visualization on dashboard, Python is the most effective language within the scope of this study.

Similarly for data manipulation, three Python libraries Pandas, PySpark and NumPy were considered. The study will mainly utilize Pandas, along with NumPy in some situations. PySpark was also explored and tested, however it was deemed infeasible for the scale of data being analyzed in this study. Additionally, the differences and limitations in PySpark’s syntax made it more challenging to perform small or basic tasks. For example, lack of *apply()* method in PySpark DataFrame requires the use of User Defined Functions (UDF). Such additional complexities would have been justified if the concerned dataset was huge and required distributed processing (SparkByExamples, 2023). Nevertheless, it was kept as an option during this study in case a much larger dataset needs to be processed.

## Project Management Framework

Two commonly used frameworks CRISP-DM and SEMMA were considered for this study. CRISP-DM stands for Cross Industry Standard Process, and was developed by a consortium of numerous data-mining companies (Kotu and Deshpande, 2014). It comprises of iterative phases that include business/research understanding, data understanding, modelling, evolution and deployment. On the other hand, SEMMA stands for Sample, Explore, Modify and Assess, was developed by Statistical Analysis System (SAS).

Both frameworks follow similar model of iterative progress and feedback loop. CRISP-DM was chosen for this study because it offers a more comprehensive framework from understanding of research requirements to final deployment of the model. Whereas business/research understanding isn’t a defined phase in SEMMA framework. With this, the project deliverables were planned as below:

# Results

# Discussion/Conclusions

# References

* SparkByExamples (2023). *Pandas vs PySpark DataFrame With Examples*. Available at: <https://sparkbyexamples.com/pyspark/pandas-vs-pyspark-dataframe-with-examples/> (Accessed: 14 April 2023).
* Kotu v. and Deshpande B. (2014) Predictive Analytics and Data Mining : Concepts and Practice with RapidMiner. Amsterdam: Morgan Kaufmann. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=e250xww&AN=919334&site=eds-live (Accessed: 12 April 2023).