Data Management Plan for Research Students

1. **Overview**

|  |
| --- |
| **Researcher:** Ama N, Karthik S, Tanvi M, Miguel P. |
| **Project title:** Banking Customer Credit-Score Bracket Classification |
| **Project duration:** 15 Weeks |
| **Project context:**  *This project lies in the field of finance, and the main objective is to classify people into credit score brackets through the use of machine learning algorithms based on their bank details and credit-related information.* |

1. **Defining your data/research sources**

|  |
| --- |
| **2.1 Where will your data/research sources come from?**  *The dataset utilized for this project has been obtained from kaggle (linked down below), and it is a public domain dataset with no copyright.*  [*https://www.kaggle.com/datasets/parisrohan/credit-score-classification*](http://www.kaggle.com/datasets/parisrohan/credit-score-classification) |
| **2.2 How often will you get new data?**  *As stated in the dataset description linked in the box above, the expected update frequency is set at "Never". However, given the size of the dataset (50.000 rows), we might consider samplig techniques to test the Machine Learning Algorithms on different sets of stances.* |

|  |
| --- |
|  |
| **2.3 How much data/information will you generate?**  *The dataframe has a size of 45.000 KB. However, this size will be reduced after*  *the process of data cleaning and preparation.* |
| **2.4 What file formats will you use?**  *The data is provided as CSV files, containing numerical and categorical data,*  *and can be easily accessed and opened with programming environments like*  *Python, Rstudio or Google Colab. Another open alternative to view the data is Google sheets.* |

1. **Organising your data**

|  |
| --- |
| **3.1How will you structure and name your folders and files?**  *The folders and files will be stored in a Github repository, separated iin folders for each process of the project (Cleaning and preparation, EDA, ML). The names of the folders and files will be self-descriptive for an easier understanding.* |
| **3.2 What additional information is required to understand each data file?**  *To reproduce the results uploaded, previous knowledge about data processing and Machine Learning would be beneficial. Nonetheless, the files will contain comments explaining most cells of code.* |
| **3.3 What different versions of each data file or source will your create?**  *To keep things simple, the final version of each file will be the file uploaded, as we will be working and improving a single file rather than creating duplicates.* |

1. **Looking after your data**

|  |
| --- |
| **4.1 Where will you store your data?**  *As previously mentioned, the data will be stored in a Github repository, so it is easily accessible for contributors and posteriorly, to openly share the results with the community.* |
| **4.2 How will your data be backed up?**  *The original version of the data will be stored in the Github repository, and a copy*  *of the dataframe will be also stored in the computer of each of the contributors.*  *There will be no need of updating the original dataframe obtained, as it is not*  *real-time data.* |
| **4.3 How will you test whether you can restore from your backups?**  *The data generated will be of public domain, as it is originally obtained from an open source (Kaggle).* |

1. **Sharing your data**

|  |  |  |
| --- | --- | --- |
| **5.1 Who owns the data you generate?**  *The data generated will be of public domain, as it is originally obtained from an open source (Kaggle).* | | |
| **5.2 Who else has a right to see or use this data?**  *During the development of the project, the data will only be used by colaborators. Posteriorly, it will be openly available.* | | |
| **5.3 Who else should reasonably have access to this data when you share it?**  *The general public will be able to access the different models created for this project.* | | |
|  | This template is licensed under a Creative Commons Attribution 3.0 **4**  Unported License. |  |

|  |
| --- |
|  |
| **5.4 What should/shouldn’t be shared and why?**  *Since the data used was obtained from a secondary data source and is openly available to everyone, there is no need to worry about ethical issues regarding data privacy or consent.* |

1. **Archiving your data**

|  |
| --- |
| **6.1 What should be archived beyond the end of your project?**  *As mentioned, what will be archived will be the final document for each section of the project. Any other files that will be needed to run such documents, like the data, will also be included and available.* |
| **6.2 For how long should it be stored?**  *As it will be stored in the cloud, it will remain there indefinitely, so it follows the EPSRC guidelines for “10 years from the date of last access”.* |
| **6.3 When will files be moved into the data archive/repository?**  *The data included in the repository will be moved to each folder during the*  *development of the project, so other members can interact with those files in real time.* |
| **6.4 Where will the data be stored?**  *The data will be stored in a Github repository available to first the colaborators, and then to the open public.* |
| **6.5 Who is responsible for moving data to the data archive and maintaining it?** |

|  |
| --- |
| *All colaborators will be equaly responsible for updating the*  *repository. Karthik Srini will be responsible for mantaining it.* |
| **6.6 Who should have access and under what conditions?**  *After the project is finalized, the open public will have open access to the files. In the meantime, only colaborators will be able to access such files, to not affect the outcome of the research.* |

1. **Executing your plan**

|  |
| --- |
| **7.1 Who is responsible for making sure this plan is followed?**  *Every colaborator will be equally responsible to make sure this plan is followed and achieved.* |
| **7.2 How often will this plan be reviewed and updated?**  *This plan will be reviewed (and changed if necessary) after the completion of each section (Data cleaning and preparation, EDA, Machine Learning implementation and validation, Discussion of findings) to tackle any possible issue faced.* |
| **7.3 What actions have you identified from the rest of this plan?**  *List them here with timescales* |
| **7.4 What further information do you need to carry out these actions?**  *Where can you find this information? Who might you be able to ask?* |

|  |
| --- |
|  |

**Notes on completing this form**

* Type as much (or as little) as you feel you need to into each box: it will expand to accommodate what you write;
* You can leave or remove the prompts in grey once you’re done;
* For help with completing this DMP, please contact [researchdata@brunel.ac.uk](mailto:researchdata@brunel.ac.uk)