Data Management Plan for Research Students

1. **Overview**

|  |
| --- |
| **Researcher:** Karthik Srini Muthusamy; Amaboh Ngu; Tanvi Mathur; Miguel Ángel Palacios; Nehal Gajraj |
| **Project title:** Banking Customer Credit-Score Bracket Classification |
| **Project duration:** 15 Weeks |
| **Project context:**   * The project is focused on the field of Finance. * The main objective of this project is to identify people’s bank details and credit related information to classify them into credit score brackets. * The classification is done using various Machine Learning Algorithms by the team. |

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

|  |
| --- |
| **2.1 Where will your data/research sources come from?**   * The source for this dataset is Kaggle which is a public domain dataset with no copyright. * The link to dataset is: [Credit Score Dataset](https://www.kaggle.com/datasets/parisrohan/credit-score-classification) |
| **2.2 How often will you get new data?**   * The intended update frequency is set to "Never," as mentioned in the dataset description referenced in the box above. * Given the size of the dataset (50.000 rows), we may want to use sampling approaches to test the Machine Learning Algorithms on multiple sets of stances. |

|  |
| --- |
|  |
| **2.3 How much data/information will you generate?**   * Initial Dataset Size: 29.6MB * The dataset size reduced to 16.2MB after all the data preparation is performed on the initial dataset. |
| **2.4 What file formats will you use?**   * The software required to access data is Excel. * The file format of the dataset is Comma Separated Values File (.csv) |

1. **Organising your data**

|  |
| --- |
| **3.1How will you structure and name your folders and files?**   * The names of the folders and files are according to the research project. * Main Folder Name: DDA project * File names will be according to the function and operations performed inside the file. |
| **3.2 What additional information is required to understand each data file?**   * The appendix contains additional data about the data, which is sufficient for the reader to understand the project. |
| **3.3 What different versions of each data file or source will your create?**   * The dataset has different versions. * The initial dataset, the cleaned dataset, the dataset after implementing EDA. |

1. **Looking after your data**

|  |
| --- |
| **4.1 Where will you store your data?**  The data is store on following locations:   * Laptop of each team member * Github Repository * OneDrive |
| **4.2 How will your data be backed up?**   * There are 5 copies of the data, one with each member. * Its stored on everyone’s local system, Github Repository and One Drive * The Github Repository is updated after any major change is made in the local files by any one team member. * Local files are updated every day as we work. |
| **4.3 How will you test whether you can restore from your backups?**   * The data is stored on Github Repository and so it can easily be traced back to previous version. |

1. **Sharing your data**

|  |  |  |
| --- | --- | --- |
| **5.1 Who owns the data you generate?**   * The data is owned by Kaggle which is a public domain with no copyrights. | | |
| **5.2 Who else has a right to see or use this data?**   * *The dataset is obtained from* an open repository so anyone can use it. | | |
| **5.3 Who else should reasonably have access to this data when you share it?**   * General public will have access to data since it originally is available on an open repository (Kaggle) | | |
|  | This template is licensed under a Creative Commons Attribution 3.0 **4**  Unported License. |  |

|  |
| --- |
|  |
| **5.4 What should/shouldn’t be shared and why?**   * Any document with a non-disclosure or confidentiality clause, as well as any ethical document, cannot be disseminated freely. |

1. **Archiving your data**

|  |
| --- |
| **6.1What should be archived beyond the end of your project?**   * Data, code, and report should be archived as it can be used further in any new research. |
| **6.2 For how long should it be stored?**   * **Following the EPSRC guidelines we will store the data for** 10 years from the date of last access. |
| **6.3 When will files be moved into the data archive/repository?**   * The files will be moved to the repository after completing and submitting the project. |
| **6.4 Where will the data be stored?**   * One Drive * Github Repository |
| **6.5 Who is responsible for moving data to the data archive and maintaining it?**   * The data is moved to repository by me and my team members. |

|  |
| --- |
| **6.6 Who should have access and under what conditions?**   * The repository is under my account so I will have access to it. * The other team members are also added to the repository as contributors, so they do have equal access to the repository. |

1. **Executing your plan**

|  |
| --- |
| **7.1 Who is responsible for making sure this plan is followed?**   * It is a group project so everyone in the team namely Karthik Srini, Amaboh Ngu, Tanvi Mathur, Miguel Ángel and Nehal Gajraj are responsible to follow the plan. |
| **7.2 How often will this plan be reviewed and updated?**   * The dataset & research question were first discussed with supervisor. * Further plan was discussed and finalized among group members. * Its reviewed ones each section is completed and is updated on repository. |
| **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)