**Project Proposal Form**

Please refer to the **Project Handbook Section 4** when completing this form

|  |  |
| --- | --- |
| **Degree Title:**  Computing | **Student’s Name: Joshua Morgan** |
| **Supervisor’s Name: Mohammad Naiseh** |
| **Project Title/Area: Using machine intelligence for the classification of the imputation types of data sets and the selection of an imputation algorithm / methodology** |

# Section 1: Project Overview

|  |
| --- |
| **1.1 Problem definition - use one sentence to summarise the problem:**  Using machine learning to determine the optimal data imputation method to use for a data set with missing values  **1.2 Project description - briefly explain your project:**  Using a machine learning and both sourced and generated data to train a multiclass classification algorithms to classify a target data set (With missing values in some variable slots). The output of this (the given class) will help determine the imputation method selected in the next section of the system.  Furthermore we will also train another multiclass classification algorithm on the output of the previous model along with meta-features and heuristics from the data set to select the “BEST” imputation method to generate the missing values  **1.3 Background - please provide brief background information, e.g., client, problem domain:**  Missing values are a common problem for both industry and research projects ***“Missing data in medical research is a common problem that has long been recognised by statisticians and medical researchers alike” (Derrick A. Bennett, 2001)*** There are three main problems with missing values; ***“loss of efficiency, complications in handling and analysing the data, bias resulting from differences between missing and complete data” (Jiří Kaiser, 2014)*** Loss of efficiency is produced by time-consuming processes used to deal with the missing values and Complications in handling and analysing the data occur due to most methods not being able to handle missing data. There are many different types of missing data that can occur in a data set ***“The missing data mechanism is usually classified as missing completely at random, missing at random or not missing at random” (Lakshminarayan, K., Harp S. A. & Samad, T. 1999)***  One of the most common and most effective ways of dealing with missing values in a dataset is data imputation. The goal of imputation is to replace missing data with plausible values, based on the other available information in the dataset for analysis. The problem with data imputation is that not all methods work as well on some data sets as others; there are many factors that can affect how effective an imputation method is on a data set, such as the missing data mechanism ***“Many imputation methods handle this degree of missing structure in the data well,***[***22***](https://journals.sagepub.com/doi/full/10.1177/2050312118822912#bibr22-2050312118822912)***and the assumption is thus often assumed to hold. However, when data are MNAR, it is difficult to identify, and consequently respond to, the missing mechanisms as this is unverifiable.” (Marianne Riksheim Stavseth, Thomas Clausen, and Jo Røislien, January 8, 2019)***    This means determining the missing data type is important, but determining the actual type can be challenging as it often relies on assumptions that could be incorrect. Missing data mechanism aside choosing the right data imputation method for your dataset is also important as data imputation done incorrectly can introduce bias (to the centre when using mean imputation) and distort the data and imputed values that are not accurate can lead to poor data analysis  **1.4 Aims and objectives – what are the aims and objectives of your project?**  Aims:   * This project aims to use machine learning and the given information about a dataset to “Recommend” an imputation method that can be used to fill in the missing data with the highest possible degree of accuracy * This project also aims to evaluate how the missing data mechanism affects the effectiveness of each data imputation method on a dataset   Objectives   * Revise literature about multiclass classification models, learning how to implement one as effectively as possible with regards to the purposes of my project * Revise literature about imputation techniques and identify on which datasets they are most effective * Identify effective methodologies (optimisation technique, regularisation technique, etc) for multiclass classification models to recommend the “BEST” solution to a given problem * Identify optimal Meta-features / Meta-heuristics to use as inputs in both machine learning models * Design interface for the system that adheres to the best practises of UI/UX design * Implement a machine learning pipeline with dataset input and algorithm recommendation output as an implementation of the solution to the proposed problem * Evaluate the final given solution next to permutations of the solution and other methodologies |

# Section 2: Artefact

|  |
| --- |
| **2.1 What is the artefact that you intend to produce?**  A machine learning pipeline that takes a target dataset as an input. This pipeline will contain;   * A preprocessing component that extracts the necessary information about the missing dataset; general dataset heuristics, information about how the data is missing (which variables and of what quantity compared to the rest of that dataset) * Two multiclass classification machine learning models;   + The first of which will classify what type of imputation problem the data set possess using the information about how the data is missing, which was extracted in the preprocessing stage,   + The second of which will output an imputation algorithm recommendation based on the dataset heuristics from the preprocessing stage and the classification from the first model. * The pipeline will also be operatable via both a small user interface and as a command line tool   **2.2 How is your artefact actionable (i.e., routes to exploitation in the technology domain)?**  The artifact can be programmed by a variety of languages, using both self-develop code and a variety of data analysis/processing, machine intelligence and mathematical libraries and frameworks. |

# Section 3: Evaluation

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **3.1 How are you going to evaluate your work?**  The missing type classification model can be tested by inputting a given dataset where the classification type of missing values are known and testing that the model outputs the same classification type for a certain percentage of instances.  The imputation algorithm recommendation model can be tested in the same way by knowing the best imputation algorithm for a given dataset and testing that the model gives the same result.  Further evaluation can be performed be testing the recommended algorithm on the data set and assuring the results obtained from it and equal to or better than the result from all the other possible recommendations.  My work can also be measured against the current industry and research based solutions that have been proposed to this problem to assess the effectiveness and diversity of the solution.  **3.2 Why is this project honours worthy?**  This project will require research into data imputation algorithms, multiclass classification machine learning models and extraction and curation of data heuristic and metafeatures, etc.  It will also involve the development and implementation of the high-functioning and experimental software solution that has been described throughout this proposal.  **3.3 How does this project relate to your degree title outcomes?**  This project involves the research of a database and data analysis problem, and the development and implementation of a machine learning and software based solution to the aforementioned problem.  **3.4 How does your project meet the BCS Undergraduate Project Requirements?**  Solving the problem of missing data in datasets and how to impute this missing data (what methods to choose) most accurately is a substantial problem to solve:   * Missing or poorly imputed data can cause issues when analysing data (bias and overall reduced accuracy) Solving this problem is important in both and research and industrial context as data analysis can be relevant to both areas * Missing data can mean that access to some functionality in software or web apps is not available, as long as this is not user caused this is an important problem to solve in industry * Missing data could potentially rest in a machine learning neural network not having an input, this means that the model won’t be able to either run most effectively or run at all. Furthermore badly imputed data could cause poor learning in the training stage or bad results in the testing stage. Solving this problem would be important in both an industrial and research context where this technology is used   This project would allow me to employ skills and knowledge that I have accumulated throughout the studies of my course such as:   * Developing and optimising neural networks as well as developing machine learning solutions to address given problems * Software development and application software development * Implementation of software development planning and the execution of the active software development process * Implementation of software principles to develop a functional peace of software with a consideration for User interface and experience * Database management and relational and non-relational databases * Web programming / API’s   **3.5 What are the risks in this project and how are you going to manage them?**   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Risk ID | Risk | Description | Risk Rating | Mitigation | | R1 | Time management | Risk of underestimating the amount of time that each section of the project will take up or outside factors affecting availability for working on the project. This might results in later stages of the project being rushed or not having enough time to be completed | 6 | Schedule the entire timeline of the project in the Gannt chart and allow for redundancy in case anything outside of control goes wrong. If it is within the bounds consult the supervisor about an extension if possible | | R2 | Difficulty in gather data | There is much data that will need to be gathered in order to train both of the multiclass classification networks  For the imputation type classification network you will need multiple datasets for each imputation problem type  For the algorithm recommendation network you will also need multiple datasets for each imputation type and multiple datasets where each imputation algorithm is the best choice  This is a lot of data to source and doing so while trying to maintain quality could be marginally difficult | 3 | The best mitigation strategy here would be to gather as much natural and high quality data as possible from multiple different sources. Then if this falls short we can generate our own data.  For the imputation typing data we can take complete datasets (from sourced datasets) and cut data out to make the required datasets  For the imputation algorithm recommendation datasets we can run the algorithm on different imputation typing datasets or others if need be | | R3 | Poor performance | The tools and methodologies I will be using have all been implemented to solve similar problems / perform likewise functions e.g. using machine learning for classification.  However the usage of the classification results as an imputation algorithm recommendation could cause performance issues as even if the network recommends the best imputation algorithm it knows that algorithm when implemented might fall short.  This could be from the network of not being aware of the best algorithm (e.g. for being too obscure) or because the imputation problem is complex and requires a more specialised solution (machine learning imputation) | 4 | Make sure that the network is aware of as many different imputation algorithms as possible (even more obscure ones)  Also allow for the fact that some datasets may not have an imputation algorithm that can perform significantly enough to generate missing data (Maybe allow for an N/A result or suggest specialisation of the best algorithm or a neural network  Finally make sure to thoroughly research optimisation and regularisation techniques and properly assess that the integrate with my solution / implementation | | R4 | Scope creep | Risk of going outside of the requirements and objectives of the initially proposed project or focussing too much on the smaller aspects of the project that are not as important and do not need as much attention / innovation  Another risk could be that requirements get tweaked | 2 | Have a clearly defined project that has clearly defined requirements that are in line with the aims of the project but also allowing enough flexibility for more middle-level changes in the project where necessary  Also having clearly defined objective / outcomes that specify what is important about the project (the main things we want to get out of it) will stop more trivial details / features getting too much attention | | R5 | Loss of work | Risk of work being lost due to accidentally and unforeseen circumstances such as; storage / hardware damage, software (OS) fault, etc | 2 | Store multiple copies of the whole project workspace on at least two different local devices and on cloud storage solutions:   * GitHub for programming work / artifact * GoogleDrive or OneDrive for documentation work | | R6 | Logical errors / deep foundational issues occur late into development | Risk of their being a logical error in the initial plan for the project, most likely with the artifact. This might be because the proposed solution actually does not solve the problem or is not able to be implemented at least in the way described in the project proposal.  If it can’t be implemented this might not be immediately obvious until the later stages of development / implementation | 4 | Thoroughly research the domain of the problem that you propose to solve to make sure the solution you put forth satisfies its needs. Furthermore do some medium to light research into the techniques and technologies you will be using or could use to implement this solution. Use all this information in the initial stages of the project to create a plan and design that is logically sound and implementable  In the case that the logical problem occurs later in the project attempt to either find a work around or adjust the project as minimally as possible to mitigate or remove the issue | |

**Section 4: References**

|  |
| --- |
| **4.1 Please provide references.**  Lakshminarayan, K., Harp S. A. & Samad, T., 1999: Imputation of Missing Data in Industrial Databases, Applied Intelligence 11, pp. 259–275. Available at: <http://www.ime.unicamp.br/~wanderson/Artigos/imputation_industrial_databases.pdf>  Jiří Kaiser, 2014: Dealing with Missing Values in Data, pp. 42. Available at: [178-697-1-PBHandlingMissingVaues.pdf](file:///C:\Users\jjmor\Downloads\178-697-1-PBHandlingMissingVaues.pdf)  Derrick A. Bennett, 2001: How can I deal with missing data in my study. Available at: [How can I deal with missing data in my study?](https://pdf.sciencedirectassets.com/783243/1-s2.0-S1326020023X61603/1-s2.0-S1326020023036488/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEDYaCXVzLWVhc3QtMSJHMEUCIQDwzTwsv83Jyfm1g2rKF%2FnVAJXpcnirnblxva5UKFI7UAIgFicIOZz%2BOboq4jDDTGeoaGZK6AF3gJPhzVXUTEFkj1gqvAUI3%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FARAFGgwwNTkwMDM1NDY4NjUiDHDK0oWkiuGpQleBwCqQBT43TqgpsHPbvqs19ipPIMVvVd7Decvh%2FxPjRn5ULPbUljXtZIkn5Ff1ShKBKGDx29aV9LRfVabSZgY%2FGBtuhRgJRktc3QgJzLa9MZlU4UCS7TDZdO14MFd7w7x%2FH3J8WfareBci2SZruXvi95lVYVUhN5fIdSuxom0HnlRUtMkw3S7%2F1QOY2%2BlR8JTPlBvQPSIext3rFMRf3cbUg7XkiNHgVLvNKEVgpUGyM0xviqeLw%2FyGjCkG9si0DsF%2Bo%2FxRxgtpIruKRry8kCC1uQcEj94pc7pawUTZTBj7Y%2FuAa%2BNhHJvhN2uSwZ9V23%2BKgWz7FW%2F2QQTBHd8s%2ByMBWzhjqrm4kNDlYXh6OrAGq94TZXyzfBX6zes%2BzFh1wWQqOe3w7ieFb1h%2BE47Im%2BSuKGee8IGReZtyf8hVUGrC7iYvsuYE7LxkB07P8NXK10y7JwedMre6vQgdPW0qK1Mt12O4J%2B3Lmw2AlmckUO76V%2BybaFcXNv3V7hivLdv%2BVauz7XGkfsDT90uo4YJQW49aQyLujS0uDiiXaVFaN1nta1xOANG%2F4xQeaR3u%2FkH2zgFQ4OK3aDLBINsFj4I8KYspBuU7gYc9TPmKQLd%2FvpSIgEdnt3kPb%2FtiLdiGGsuz7lTcRyqk2%2FjGqf%2B8zNu%2BbntsHiZR4h6M5HOSnX3UutaJ%2Bt0zDS33Ts%2F1XixKzWCk72qEnjqzBN6jDHMqkglMW8L%2BR5uvHlVB8Bp4Fm4ZgjVwAjennTORNOoSUgpcuuUyST6l7hEOpUJLsgfUMpBz08po4K0XmAmKtxBkZgS06iZ3Jvbl2uarCyz9c5Xpl4irISFkWky8CKpaCJ3Qs2gFwjLp7gQLdfNUjZRsdM4EeljjDLbQwfjqMIHwvboGOrEBzWefEasf5Hrfm%2FIk3mw78b52ODSdSwG1GMznIJJknSaDIfbwJDxsRc%2FqK%2FA5Qck8DAdcYKEZ5OBoSHdOKnHzl8XW4%2FJrFL35kj%2FvYKRsx0Us9IUx0gRPIG6F60VFzBTs%2F6IIXYi6Uc80tYgCnSSBhGqYvw5vGgfVm4hQjwcCGKoj4FVuw538zLlec7YQn4S7s9m9A6Crifer5NiGKGZPW53ylgJiUWgVMSryx1IyJUlG&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20241203T221431Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTYVREE24Y3%2F20241203%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=31912a0c6c2f92cd508429f6cdf649fcbdce0cabc22b00f261fe1aef236d2aee&hash=37b8d7aa1fed9eafb17d5db533dba56aa8e643654f059963f2ec66bcb144b636&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S1326020023036488&tid=spdf-d368552e-1314-4906-9d9d-54dbffa40299&sid=e87a6ba94b367846bf7ad7023acfabf350fbgxrqb&type=client&tsoh=d3d3LnNjaWV)  Marianne Riksheim Stavseth, Thomas Clausen, and Jo Røislien, January 8, 2019. How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. Available at: [How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data](https://journals.sagepub.com/doi/epub/10.1177/2050312118822912) |

# Section 5: Proposed Plan (please attach your Gantt chart below)

# Note: Please complete the research ethics checklist once the proposal has been approved by your supervisor.