Project Draft

Data Imputation with Machine Learning



|  |  |
| --- | --- |
| Name: | Supervisor: |
| Student ID: |  |
| Word Count: 2791 (Without References) |  |
| Date: |  |

Table of Contents

[Table of Figures 2](#_Toc69897951)

[Abstract 3](#_Toc69897952)

[Introduction 3](#_Toc69897953)

[Background 4](#_Toc69897954)

[Aim/Objective 4](#_Toc69897955)

[Research Questions 4](#_Toc69897956)

[Ethical Considerations 5](#_Toc69897957)

[Literature Review 5](#_Toc69897958)

[Project Timeline 9](#_Toc69897959)

[Methodology 9](#_Toc69897960)

[Project Evaluation 10](#_Toc69897961)

[Conclusion 10](#_Toc69897962)

[References 11](#_Toc69897963)

# Table of Figures

[Figure 1 - Literature Review Workflow 5](#_Toc69897902)

[Figure 2 - Results for data missing completely at Random 7](#_Toc69897903)

[Figure 3 - Results for the wine dataset. 7](#_Toc69897904)

[Figure 4 - Results for the glass dataset. 8](#_Toc69897905)

[Figure 5 - Results for the concrete compressive strength dataset. 8](#_Toc69897906)

[Figure 6 - Results for the liver patient dataset. 8](#_Toc69897907)

[Figure 7 - Results for the seeds dataset. 8](#_Toc69897908)

[Figure 8 - Project Timeline 9](#_Toc69897909)

[Figure 9 - Interactive Trend Plot for the Dataset Features 10](#_Toc69897910)

# Abstract

Machine learning (ML) as a methodology for implementing AI-infused ideas has gained significant traction over the past 2 decades. This has led to an exponential increase in the development of Machine Learning and the number of researchers that focus on improving ML. The rapid development of Machine Learning wouldn’t have been possible if there wasn’t enough data to go on.

All ML projects are heavily dependent and there is no ML without data. The importance and significance of data are unreplaceable to ML. As such, the data available for ML projects must be perfect for precise results. But often this isn’t the case. The process of dataset collection is now getting automated through web scrappers and other similar tools. This results in datasets that aren’t as perfect and accurate. They might be missing some values or the values might not be of the proper format.

This research aims to use the prowess of Machine Learning, especially regression to counter this issue of missing values in datasets. The output of this research would be the experimental analysis of applying ML to datasets to counter their missing values.

# Introduction

Decision making based on information from data is highly dependent on the truthfulness of the data (Bengtsson and Lindblad, 2020). For analysis of data being as accurate as possible, it follows that the data needs to be as accurate as possible (Bengtsson and Lindblad, 2020). Accurate data imply that the data is complete since incomplete data increase the risk of weakening the validity (Bengtsson and Lindblad, 2020). However, in the real world, data tends to be incomplete (Bengtsson and Lindblad, 2020). In many cases, the incompleteness is due to the challenging problem of missing values (Bengtsson and Lindblad, 2020). A missing value occurs when an observation does not have a collected value for a variable (Bengtsson and Lindblad, 2020). With missing values, information about the population is missing which risks having data that does not reflect the population truthfully (Bengtsson and Lindblad, 2020). This can affect the conclusions drawn from the data (Bengtsson and Lindblad, 2020).

Generally measurable and AI calculations are not powerful enough to deal with missing qualities (Jadhav, Pramod and Ramanathan, 2019). They get influenced by missing information. Missing information presents a component of uncertainty while dissecting information and that can influence properties of measurable assessors and results in loss of force and deceiving ends (Jadhav, Pramod and Ramanathan, 2019). Fittingly managing missing qualities is a significant and testing task since it requires: (Jadhav, Pramod and Ramanathan, 2019)

1. Careful assessment of all examples of information to distinguish example of missingness in the information and
2. Clear comprehension of various ascription procedures. This research project will result in an algorithmic approach that will be able to tackle this issue regarding missing data instances.

## Background

Data is what drives the 21st century. The advent of technology and the exponential rise in the availability of technology to the masses has caused a rapid surge in data consumption and production. Handling such a massive data flow at every moment becomes taxing and sometimes mistakes are made while data recording. These mistakes cause gaps that cause flaws in the statistics that are performed on the data and affect the complete following process and results.

This issue persists since the methods for dealing with this aren’t developed but the data consumption and production are skyrocketing every second. This is what initiated the research for a methodology that can tackle this.

## Aim/Objective

The aim here would be to create a machine learning algorithm/ensemble that provides recovery or replacement for any missing or inconsistent data in a respective dataset. The dataset used will be of the medical domain. The following steps when traced will help achieve this aim:

* Writing review of the work done by specialists in similar area to learn of the mainstream calculations utilized for information attribution.
* Dataset determination and information pre-preparing. In any case, during the time spent pre-handling, the missing qualities won't be prepared since that will be finished by the calculation.
* Planning the Machine Learning Algorithm. The dataset measurements will be noticed and the appropriate calculation will be utilized to decide the missing qualities in the information. The worldview for this Machine Learning Research Project will be Regression. Along these lines, Regression calculations like SVM Regressor, MLP Regressor, ARIMA and Multivariate Linear Regression calculations will be utilized for this reason. These calculations will be accessible from the scikit-learn library for Python.
* Execution Metrics for the Algorithms. Since exactness can't be determined for Regression calculations, their presentation must be estimated as a blunder in its forecasts. To quantify these blunders, the accompanying ways are utilized:

|  |  |
| --- | --- |
| **Performance Metric** | **Formula** |
| Mean Squared Error (MSE) | If **n** predictions are produced from a dataset with **n** data instances; where **Y** is the vector of labels in the dataset and is the vector of predicted labels, then |
| Root Mean Squared Error (RMSE) |  |
| Mean Absolute Error (MAE) |  |

## Research Questions

1. Can machine learning provide the assistance needed in combating the issue of missing data instances?

## Ethical Considerations

The UK Data Service division additionally gives rules to moral exploration with an explicit connection to Big Data. These rules will frame the reason for this report’s moral methodology. The points that need to be focused on are:

* Keeping data confidential that disregards bunch protection,
* Referring to hotspots for all data utilized inside the examination project,
* Guaranteeing all information is put away in the right area.

## Literature Review

The amount of missingness provides a clue to what extent the missing values affect the results, as it is related to its impact on research conclusions (Bengtsson and Lindblad, 2020). Generally, larger proportions of missing values tend to have a greater impact on statistical inference and generalizability since it indicates that more information about the population is missing (Bengtsson and Lindblad, 2020). The sample data might reflect a bias as a lot of observed data gets deleted due to a lot of observations obtaining missing values, leading to biased parameter estimates and misleading statistical inference (Bengtsson and Lindblad, 2020).

This makes the Literature Review, the process of referring to the research done by peers in the same domain mandatory. Since there is a lot of work done in this field, a method to proceed with this is made as shown in *Figure 1*.

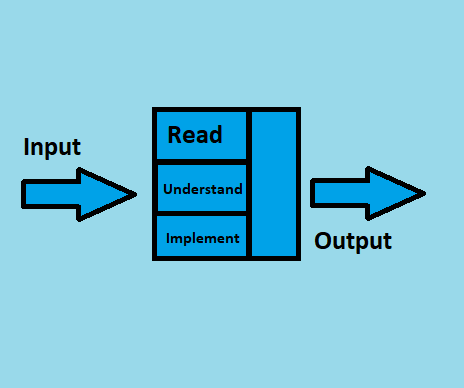


Figure 1 - Literature Review Workflow

As alluded to above, work is being done on this point since quite a while back. In this time, a ton of audit material, explicitly examination papers have amassed for anybody that necessities to do a Literature Review. Notwithstanding, investigating the entirety of the papers at this point isn't feasible. Along these lines, a game plan for picking quality papers needs to in place. The strategy is as per the going with:

* English should be the solitary language in the papers.
* Should be spread in journals with a high impact factor.
* Complete and free access ought to be available for the Paper and the Journal.
* All the assessment datasets and code for the papers ought to be open free.
* 10 years old examination isn't permitted.

The papers selected for study in this research are discussed in this section. The methodology/experiments conducted in these research papers and the results that those experiments bore will be the main focus of this section.

The missing information issue is ostensibly the most well-known issue experienced by machine learning specialists while examining true information. In numerous applications going from quality articulation in computational science to study reactions in sociologies, missing information is available to different degrees (Bertsimas, Pawlowski and Daisy Zhuo, 2018). As numerous measurable models and AI calculations depend on complete informational indexes, it is vital to deal with the missing information properly. Sometimes, basic methodologies may do the trick to deal with missing information. For instance, a complete-case examination utilizes just the information that is completely known and precludes all perceptions with missing qualities to lead to factual examination (Bertsimas, Pawlowski and Daisy Zhuo, 2018).

This functions admirably if a couple of perceptions contain missing qualities, and when the information is missing totally at irregular, complete case examination doesn't prompt one-sided results (Bertsimas, Pawlowski and Daisy Zhuo, 2018). Then again, a few AI calculations normally represent missing information, and there is no requirement for pre-handling. For example, CART and K-implies have been adjusted for issues with missing information. In numerous different circumstances, missing qualities should be attributed preceding running measurable examinations on the total informational index (Bertsimas, Pawlowski and Daisy Zhuo, 2018).

Researchers have been working on this problem since long before and have come up with innovative, new and novel approaches to handling Data Imputation. The Literature Review for this research focuses on approaches to Data Imputation after 2014. The approaches for Data Imputation are mostly statistical, but recently more innovative and novel approaches are appearing in the Literature i.e., deep learning, generative adversarial learning, fuzzy logic, Autoencoders and many more.

Missing information is omnipresent in huge information clinical preliminary. Albeit numerous examinations don't unequivocally report how they handle missing information, some verifiable techniques are utilized in measurable programming. Subsequently, various bundles may deal with missing information in an unexpected way (or the default strategies are unique) and results may not be reproduced precisely by utilizing diverse measurable programming bundles (Zhang, 2016). Once in a while, this may not lead to fundamentally various outcomes, yet the logical adequacy of the investigation is undermined. The best practice is to expressly state how missing qualities are taken care of. For effortlessness, numerous specialists essentially erase deficient case (Listwise cancellation), which is likewise the default technique in numerous relapse bundles (Zhang, 2016).

This strategy gets dependable outcomes just when the quantity of missing qualities isn't enormous and the missing example is missing indiscriminately (MCAR) or missing MAR. Another disservice of complete case examination is data misfortune. This can be a major issue when there is an enormous number of factors (segments) (Zhang, 2016). A generous number of cases can be erased because cancellation depends on missingness on at least one factors. Moreover, a complete case investigation can prompt erratic inclination (Zhang, 2016). The answer to this issue is attribution. Missing qualities are supplanted by ascribed values. Since ascription is a space of dynamic examination, there are various techniques and bundles created for attribution (Zhang, 2016).

The missing values are roughly estimated using central tendency measures like mean, median and mode in many types of research (Zhang, 2016). The mean and standard deviation are one-sided. Attributions with mode and middle work in a similar way and they are left to users for training (Zhang, 2016). Albeit harsh attribution gives quick and basic techniques to missing qualities, it belittles change, bargains connection among factors, and inclinations rundown insights. Hence harsh ascriptions must be utilized when a modest bunch of qualities are missing, they are not for general use (Zhang, 2016).

Some researchers use Listwise Deletion, Predictive Mean Matching and Poisson Imputation for tackling the data imputation problem (Bengtsson and Lindblad, 2021). The results for these techniques from the paper by (Bengtsson and Lindblad, 2021) are as follows:

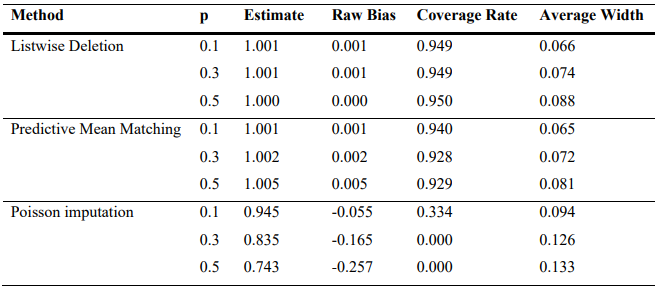


Figure 2 - Results for data missing completely at Random

Imputations using Central Tendency measures are very popular amongst researchers in this domain. Other approaches like the Predictive Mean matching Discussed above is also popular amongst researchers. Moderately used approaches for Data Imputation also include Imputing using clustering techniques like k-NN, imputing using probabilistic methods like the Bayes theorem and also generic regression algorithms like the Linear regression. Results of these algorithms from the research done by (Jadhav, Pramod and Ramanathan, 2019) are as follows:

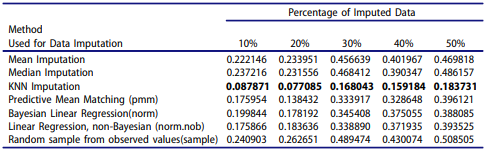


Figure 3 - Results for the wine dataset.

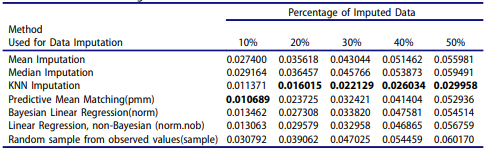


Figure 4 - Results for the glass dataset.

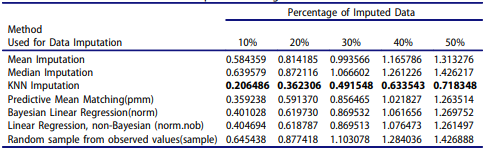


Figure 5 - Results for the concrete compressive strength dataset.

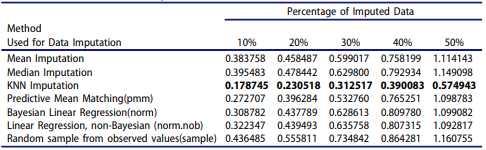


Figure 6 - Results for the liver patient dataset.

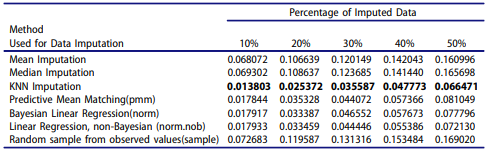


Figure 7 - Results for the seeds dataset.

The results from these researches prove the validity and the strength of these approaches. It proves why these approaches are so popular and repeatedly used in this domain. Although the approach is not the only factor affecting the results. The implementation of the approach on selected datasets also matters a lot. That is why (Jadhav, Pramod and Ramanathan, 2019) used five datasets to prove the strength of the performance of various popular Data Imputation techniques.

## Project Timeline

Research projects have time limitations and the ability to consent to a period requirement is imperative to advance. The course of occasions in a task undeniably fans out key endeavour assumptions and the degree of their perfection. This investigation project recognizes time as its unmistakable benefit. By gainfully distributing time to various endeavours resource over-trouble is restricted. Thwarting resource over-trouble restricts the threat of significant worth lessening.

The Gantt chart is perceived as a strong instrument for the time the load up. The Gantt layout expected for this endeavour is fanned out underneath in *Figure 1*. Endeavours are fanned out in consecutive solicitation on the left-hand side. The time frame for their completion is found along the X-Axis. By sticking to this plan, the endeavour will be passed on advantageously to a raised assumption.

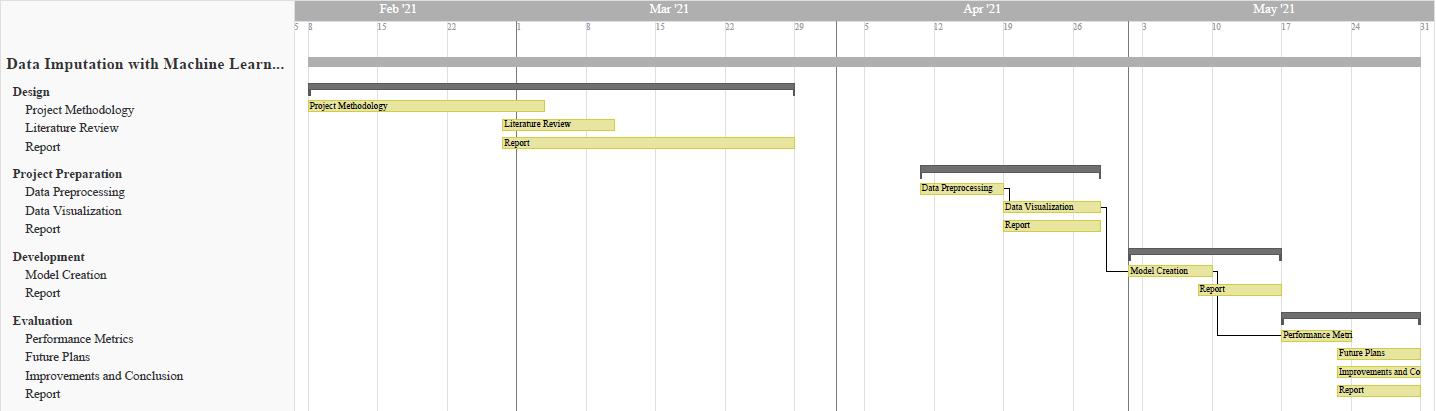


Figure 8 - Project Timeline

# Methodology

The dataset used will be the RAPID (Real-time Adaptive and Predictive Indicator of Deterioration) project that collects and analyses real-time patient data and alerts if the patient health is declining. The paradigm from the analysis of the dataset and the review of the Research Questions has been deduced to be Regression.

The features in the dataset are as follows:

* <Timestamp>: The time at which the data instance was recorded.
* <Lifetouch Heart Rate>: The heart rate of the Patient record.
* <Lifetouch Respiration Rate>: The respiration rate of the Patient record.
* <Oximeter SpO2>: Blood oxygen levels on the Patient record.
* <Oximeter Pulse>: Patient pulse recorded.

This data has been recorded in real-time from the patients. On observation, the dataset has issues like missing values, values that don’t follow the pattern i.e. -1. These issues have to be checked to increase the quality of the dataset and the results of the Machine Learning algorithm. Therefore, data pre-processing is mandatory for this project.

The number of features isn’t that many. Also, from observing the dataset, the paradigm of this machine learning project can be selected as regression. The dataset is time-series data providing diagnosis on the patient. This data will be used to predict the upcoming values for these features in advance. This will help predict if the patient’s health is going to decline or not. Regression algorithms will be implemented for this research project. Machine Learning libraries for Python like scikit-learn provide many Regression algorithms like ANN Regressor, SVM Regressor, ARIMA and many more. These algorithms will be used to train and predict this dataset.

Figure 9 - Interactive Trend Plot for the Dataset Features

The dataset is not a large dataset with just four highlights to work with. The pattern line diagram above is of the dataset. It shows how the four highlights in the dataset circle back to time. Each element is special and essential to the dataset. The justification this is that there are as of now less highlights to work with. The lone conceivable information cleaning for this dataset will be to fill the missing qualities. That will be finished by the finalized ML ensemble made in this examination.

# Project Evaluation

The project will be evaluated by:

* Performance Metrics for Regression Algorithms
* Quality of the Statistical Data Analysis

# Conclusion

The importance of Data imputation increases with the increase in data. The increase in data has been exponential. It is no longer a human feat to collect and manage data. Data collection is being automated. As such, the resulting datasets might contain faults or may have a case of missing values. This research focuses on handling this issue using the regression paradigm in Machine Learning RAPID (Real-Time Adaptive and Predictive Indicator of Deterioration).

# References

1. Bengtsson, F. and Lindblad, K., 2021. Methods for handling missing values: A simulation study comparing imputation methods for missing values on a Poisson distributed explanatory variable. Uppsala University, [online] Available at: <https://www.diva-portal.org/smash/get/diva2:1520218/FULLTEXT01.pdf> [Accessed 5 February 2021].
2. Jadhav, A., Pramod, D. and Ramanathan, K., 2019. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence, 33(10), pp.913-933.
3. Jason Anastasopoulos, L. and G.Hunter, S., 2020. Fast and Robust Missing Data Imputation. researchgate.net, [online] Available at: <https://www.researchgate.net/publication/347741781\_Fast\_and\_Robust\_Missing\_Data\_Imputation> [Accessed 5 February 2021].
4. Jin, L., Bi, Y., Hu, C., Qu, J., Shen, S., Wang, X. and Tian, Y., 2021. A comparative study of evaluating missing value imputation methods in label-free proteomics. Scientific Reports, 11(1).
5. Zhang, Z., 2016. Missing data imputation: focusing on single imputation. Annals of Translational Medicine, [online] 4(1). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4716933/> [Accessed 5 February 2021].
6. Amiri, M. and Jensen, R., 2016. Missing data imputation using fuzzy-rough methods. Neurocomputing, 205, pp.152-164.
7. BEAULIEU-JONES, B. and MOORE, J., 2016. MISSING DATA IMPUTATION IN THE ELECTRONIC HEALTH RECORD USING DEEPLY LEARNED AUTOENCODERS. Biocomputing 2017.
8. Bertsimas, D., Pawlowski, C. and Daisy Zhuo, Y., 2018. From Predictive Methods to Missing Data Imputation: An Optimization Approach. Journal of Machine Learning Research, pp.1-39.
9. Duan, Y., Lv, Y., Liu, Y. and Wang, F., 2016. An efficient realization of deep learning for traffic data imputation. Transportation Research Part C: Emerging Technologies, 72, pp.168-181.
10. García-Laencina, P., Abreu, P., Abreu, M. and Afonoso, N., 2015. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Computers in Biology and Medicine, 59, pp.125-133.
11. Li, Z., Sharaf, M., Sitbon, L., Sadiq, S., Indulska, M. and Zhou, X., 2013. A web-based approach to data imputation. World Wide Web, 17(5), pp.873-897.
12. Yoon, J., Jordon, J. and van der Schaar, M., 2018. GAIN: Missing Data Imputation using Generative Adversarial Nets. Proceedings of the 35th International Conference on Machine Learning.