

Data Imputation with Machine Learning

Dissertation Report

# 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.

**Keywords:** Machine Learning, Data Imputation, Missing Values

# Acknowledgements

My sincere gratitude to my dissertation supervisor <Insert Supervisor Name Here>. Their advice, constructive criticism and unwavering support have been invaluable throughout the process of this work. I will forever be grateful for their encouragement to pursue this avenue of research. I would also like to thank my family without whom this work would not have been possible.

Table of Contents

[Abstract 2](#_Toc79959760)

[Acknowledgements 3](#_Toc79959761)

[Table of Figures 4](#_Toc79959762)

[1. Introduction 6](#_Toc79959763)

[a. Background 6](#_Toc79959764)

[b. Aim/Objective 8](#_Toc79959765)

[c. Research Questions 8](#_Toc79959766)

[d. Ethical Considerations 9](#_Toc79959767)

[2. Literature Review 9](#_Toc79959768)

[3. Project Timeline 14](#_Toc79959769)

[4. Project Evaluation 14](#_Toc79959770)

[5. Methodology 15](#_Toc79959771)

[a. Dataset 15](#_Toc79959772)

[b. Machine Learning 16](#_Toc79959773)

[i. Neural Networks 16](#_Toc79959774)

[c. Results 19](#_Toc79959775)

[6. Conclusion 19](#_Toc79959776)

[7. Future Work 21](#_Toc79959777)

[8. References 21](#_Toc79959778)

# Table of Figures

[Figure 1 - Performance Metrics 8](#_Toc79959728)

[Figure 2 - Literature Review Workflow 9](#_Toc79959729)

[Figure 3 - Results for Missing Data Completely at Random 12](#_Toc79959730)

[Figure 4 - Results for the Wine Dataset 13](#_Toc79959731)

[Figure 5 - Results for the Glass Dataset 13](#_Toc79959732)

[Figure 6 - Results for the concrete compressive strength dataset. 13](#_Toc79959733)

[Figure 7 - Results for the liver patient dataset. 13](#_Toc79959734)

[Figure 8 - Results for the seeds dataset. 14](#_Toc79959735)

[Figure 9 - Project Timeline 14](#_Toc79959736)

[Figure 10 - Dataset Metadata. 15](#_Toc79959737)

[Figure 11 - Replacing -99 with "Nan”. 16](#_Toc79959738)

[*Figure 12 - Basic Architecture of a Neural Network* 17](#_Toc79959739)

[*Figure 13 - Popular Activation Functions* 18](#_Toc79959740)

# Introduction

Missing data can be defined as the class of problems that create chaos and difficulty by creating an absence of a part or more of a dataset (Efron, 1994). Decision making based on information from data is highly dependent on the truthfulness of the data (Bengtsson and Lindblad, 2020). For the analysis of data to be 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)

* Careful assessment of all examples of information to distinguish examples of missingness in the information and
* 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.

The research made on problems in the real world are victims of missing data in datasets. This problem Is highly common across the globe with all the researchers and this common problem causes high-level discrepancies in research results and forces researchers to take forceful steps like removing data instances with missing values to resolve and terminate this issue (Pantanowitz and Marwala, 2009). Furthermore, the ability to deduce information from the dataset using statistics and data visualization also decreases due to missing feature values in data instances (Pantanowitz and Marwala, 2009).

Missing data is classified by (Pantanowitz and Marwala, 2009) based on the pattern and mechanism of the absence of the data. The methods that can be implemented for data imputation on a specific dataset depends on this very classification (Pantanowitz and Marwala, 2009). Three broad classes are defined; monotone missingness, file matching and general missingness (Pantanowitz and Marwala, 2009). The mechanisms, in order from least to most dependent on other information, are: missing completely at random (MCAR); missing at random (MAR); and the non-ignorable case (Pantanowitz and Marwala, 2009). In the MCAR case, data cannot be predicted using any information in the set, known or unknown. For the MAR mechanism, there is a correlation between the missing data and the observed data, but not necessarily on the values of the missing data (Pantanowitz and Marwala, 2009).

## 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.

Missing information is a genuine issue regularly experienced in logical settings. Missing information is tricky as numerous factual investigations require total information (Tang and Ishwaran, 2017). This powers analyst who needs to utilize a factual investigation that requires total information to pick between ascribing information or disposing of missing values. In any case, to just dispose of missing information is certifiably not a sensible practice, as important data might be lost and inferential power bargained (Tang and Ishwaran, 2017). Consequently, ascribing missing information in such settings is a more sensible and commonsense approach. While numerous factual strategies have been created for crediting missing information, large numbers of these perform ineffectively in high dimensional and enormous scope information settings; for the model, in genomic, proteomic, neuroimaging, and other high-throughput issues (Tang and Ishwaran, 2017). Specifically, it is by and large suggested that all factors be remembered for numerous ascriptions to make it legitimate when all is said in done and altogether not to make inclination in the gauge of the connections (Tang and Ishwaran, 2017). Be that as it may, this can prompt overparameterization when there is an enormous number of factors also, the example size is moderate (Tang and Ishwaran, 2017).

A promising methodology can be founded on Breiman’s (Breiman, 2001) irregular timberlands (Tang and Ishwaran, 2017) (contracted in the future as RF). RF have the wanted trademark that they: (1) handle blended kinds of missing information; (2) address associations and nonlinearity; (3) scale to high measurements while staying away from overfitting; and (4) yield proportions of variable significance valuable for variable determination. Presently there are a few diverse RF missing information calculations. This incorporates the first RF vicinity calculation proposed by Breiman (Breiman, 2001) (Tang and Ishwaran, 2017) carried out in the random Forest R-bundle (Tang and Ishwaran, 2017). An alternate class of calculations are the "on-the-fly-attribution" (OTFI) calculations executed in the random survival forest R-bundle (Tang and Ishwaran, 2017), which permit information to be attributed while at the same time growing an endurance tree. These calculations have been bound together inside the randomForestSRC R-bundle (abridged as RF-SRC) to incorporate endurance as well as order and relapse among different settings (Tang and Ishwaran, 2017). A third approach is missForest, a strategy as of late presented in (Tang and Ishwaran, 2017). Missforest adopts an alternate strategy by reworking the missing information issue as a forecast issue. Information is attributed by relapsing every factor thusly against any remaining factors and afterwards anticipating missing information for the reliant variable utilizing the fitted woods. MissForest has been shown (Tang and Ishwaran, 2017) to beat notable techniques, for example, k-closest neighbours (KNN) (Tang and Ishwaran, 2017) and parametric MICE (Tang and Ishwaran, 2017) (multivariate attribution utilizing affixed condition). Given that RF meets every one of the attributes for dealing with missing information, it appears to be attractive to utilize RF for attributing information (Tang and Ishwaran, 2017).

Missing data are a real issue that affects many data scientists and researchers across the globe. This problem is inherent in nature and very common in data collection, especially when dealing with large datasets collected from the real world. This impacts the decision-making ability of the Machine Learning algorithm or anyone in fact that uses data with missing feature values (Pantanowitz and Marwala, 2009). To quickly resolve this issue, some researchers tend to replace the missing values with random values, but this has more of a negative impact on the results of the research (Pantanowitz and Marwala, 2009). 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:

* Literature Review: Writing a review of the work done by specialists in a similar area to learn of the mainstream calculations utilized for information attribution.
* Data Collection and Preprocessing:
  + 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.
* Machine Learning:
  + 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.
* Performance Metrics
  + 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:



Figure 1 - Performance Metrics

## Research Questions

* 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 2.

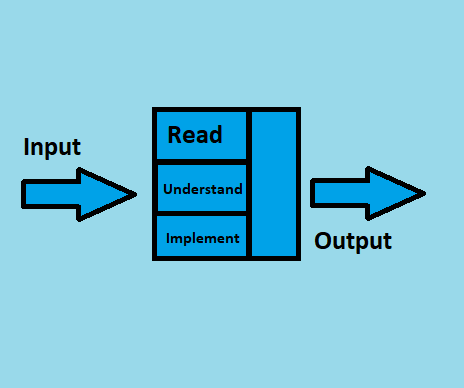


Figure 2 - Literature Review Workflow

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 be 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 issue of absent (or inadequate) information is moderately normal in numerous fields of exploration, and it might have various causes, for example, gear breakdowns, inaccessibility of hardware, refusal of respondents to address certain inquiries, and so forth These kinds of missing information are accidental and uncontrolled by the specialists, however, the general outcome is that the noticed information can't be examined because of the inadequacy of the informational indexes. Various specialists over the last quite a few years have researched strategies for managing missing information (Li, Deogun, Spaulding and Shuart, 2004). Techniques for dealing with missing information can be partitioned into three classes. The first is overlooking and disposing of information, and listwise erasure and pairwise cancellation are two generally utilized strategies in this classification (Li, Deogun, Spaulding and Shuart, 2004). The subsequent gathering is boundary assessment, which utilizes variations of the Expectation-Maximization calculations to gauge boundaries within the sight of missing information (Li, Deogun, Spaulding and Shuart, 2004). The third classification is attribution, which signifies the way toward filling in the missing qualities in a piece of information sent by some conceivable qualities’ dependent on data accessible in the informational index (Li, Deogun, Spaulding and Shuart, 2004). Among all attribution draws near, numerous alternatives are differing from a straightforward strategy like mean ascription, to some more hearty and muddled strategies dependent on the examination of the connections among ascribes. Head ascription strategies by and by incorporate (a) Mean attribution; (b) Regression ascription; (c) Hot deck ascription; and (d) Multiple ascriptions (Li, Deogun, Spaulding and Shuart, 2004). Bunching calculations have been broadly utilized in hot deck ascription. Quite possibly the most well-known bunching calculation is the K-implies technique, which takes the number of alluring groups, K, as info boundary, and yields a dividing of K groups on a bunch of items (Li, Deogun, Spaulding and Shuart, 2004).

An essential issue in missing information ascription is to fill in missing data about an article dependent on the information on other data about the object (Li, Deogun, Spaulding and Shuart, 2004). As perhaps the most well-known procedure in information mining, the grouping technique works with the way toward taking care of this issue. (Li, Deogun, Spaulding and Shuart, 2004) in his research employs the use of clustering algorithms to achieve missing data imputation. The calculation for missing information ascription with the K-implies bunching technique can be separated into three cycles. In the first place, arbitrarily select K complete information objects as K centroids (Li, Deogun, Spaulding and Shuart, 2004). Second, iteratively adjust the parcel to lessen the amount of the distances for each item from the centroid of the group to which the item has a place. The interaction ends once the summation of distances is not exactly a client determined edge ε (Li, Deogun, Spaulding and Shuart, 2004). The last cycle is to fill in all the non-reference ascribes for each deficient item dependent on the group data. Information protests that have a place with a similar bunch are taken as closest neighbours of one another, and we apply the closest neighbour calculation to supplant missing information (Li, Deogun, Spaulding and Shuart, 2004).

(Li, Deogun, Spaulding and Shuart, 2004) also expands the first K-implies grouping strategy to a fuzzy variant to credit missing information. The justification for applying fuzzy methodology is that fuzzy bunching gives a superior depiction instrument when the groups are not all-around isolated, similar to the case in missing information ascription (Li, Deogun, Spaulding and Shuart, 2004). Besides, the first K-implies grouping might be caught in a nearby least status if the underlying focuses are not chosen as expected. Notwithstanding, consistent enrollment esteems in fuzzy grouping make the subsequent calculations less helpless to stall out in neighbourhood least circumstance (Li, Deogun, Spaulding and Shuart, 2004).

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.

Many methods and approaches have been formulated concerning dealing with data imputation. The simplest procedure for dealing with this issue is the removal of data instances from the data that contain missing feature values (Pantanowitz and Marwala, 2009). This method is simple and in some of the cases where the complexity of the problem statement is not alarming, this can be implemented. The issue with this method is that the barbaric removal of data instances from the dataset can create inherent biases and partial results from the dataset. Other techniques include the available case procedures, weighting procedures and imputation-based procedures. The imputation techniques that involve the prediction of missing values using Machine Learning are applied to the previously discussed MAR and MCAR cases.

Two classes of techniques exist based on the fact that prediction can be used to impute data i.e., non-model-based approach and model-based approach. Non-model based approaches consist of statistical approaches like the mean imputation and hot-deck imputation (Pantanowitz and Marwala, 2009). The working procedure of these non-model techniques is to reduce the variance in the datasets using various statistical tools. Model-based approaches include regression-based techniques, multiple imputations (Pantanowitz and Marwala, 2009), expectation-maximization (Pantanowitz and Marwala, 2009) and neural network (NN) based approaches.

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 cases (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 factor. 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:

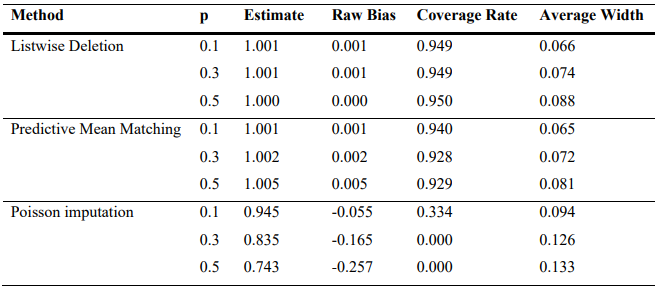


Figure 3 - Results for Missing Data Completely at Random

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:

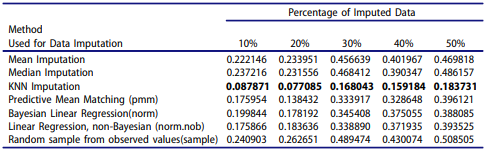


Figure 4 - Results for the Wine Dataset

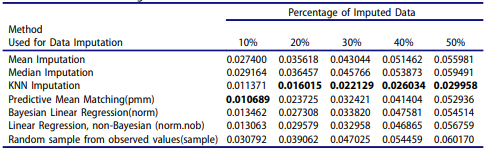


Figure 5 - Results for the Glass Dataset

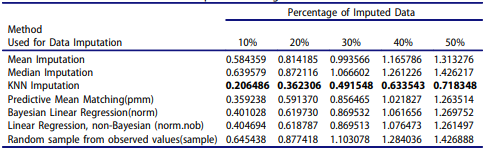


Figure 6 - Results for the concrete compressive strength dataset.

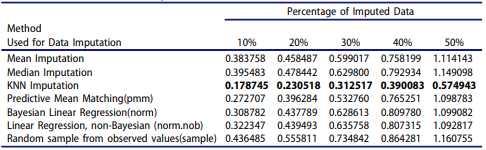


Figure 7 - Results for the liver patient dataset.

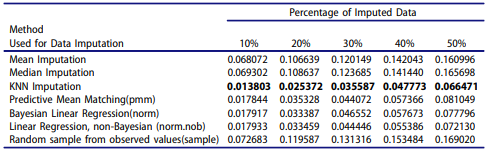


Figure 8 - Results for the seeds dataset.

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 9. 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.

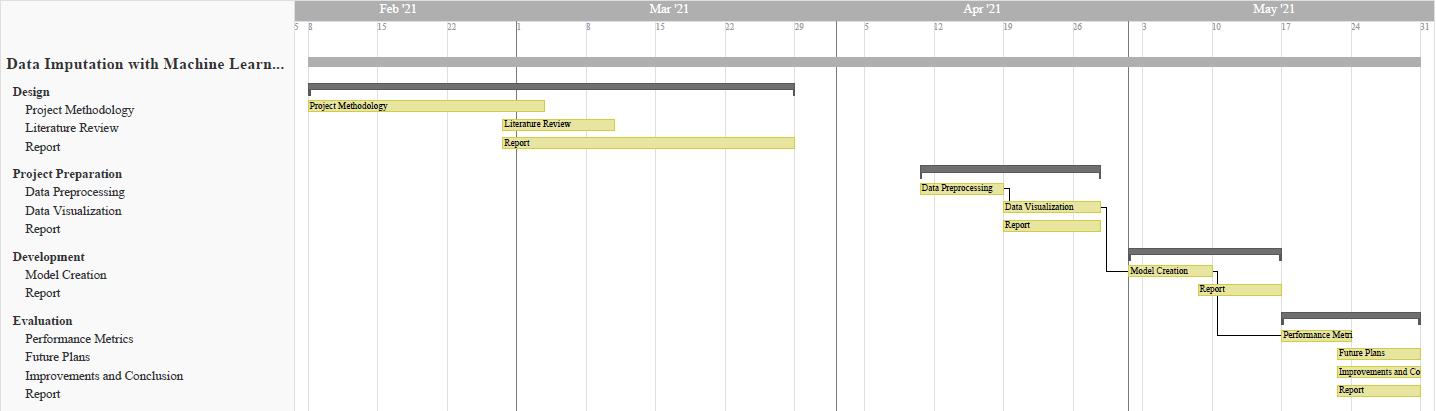


Figure 9 - Project Timeline

# Project Evaluation

The project will be evaluated by:

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

# Methodology

## Dataset

The dataset employed for this research is the city temperature dataset. This dataset contains the temperature readings for various states from various countries belonging to different regions from across the world. The features of this dataset are as follows:

* State
* Region
* City
* Country
* Month
* Day
* Year
* AvgTemperature

From the features in the dataset, it is easy to understand that the dataset contains extensive data for the temperatures from various countries. Average temperatures daily are recorded from the year 1995 to 2019. As a result, the dataset is enormous. The application of the dataset for this research requires data preprocessing since not all features are of use in this research. Upon inspecting the research, it can be seen that the main aim of this research is to show the average temperature of a specific area over the years. The area under consideration can be selected in 2 different ways, by Region i.e., the continent or by country. The state and city features are just extensions of the country feature and can be either grouped or ignored. For this research, the area considered is by country. This makes the features “Region”, “State” and “City” obsolete since the main focus is to observe the changing of the average temperature of each country daily from the year 1995 to 2019. The final dataset now contains “City”, “Month”, “Day”, “Year” and “AvgTemperature” as features.

The “City” feature is of the categorical type i.e., it consists of textual data. To be useful in this research, the feature must be encoded into a numerical data type so that it is easier to decide upon a Machine Learning technique for supervised learning. To this extent, Label Encoding will be employed to encode the dataset. The metadata of the dataset is studied after encoding.

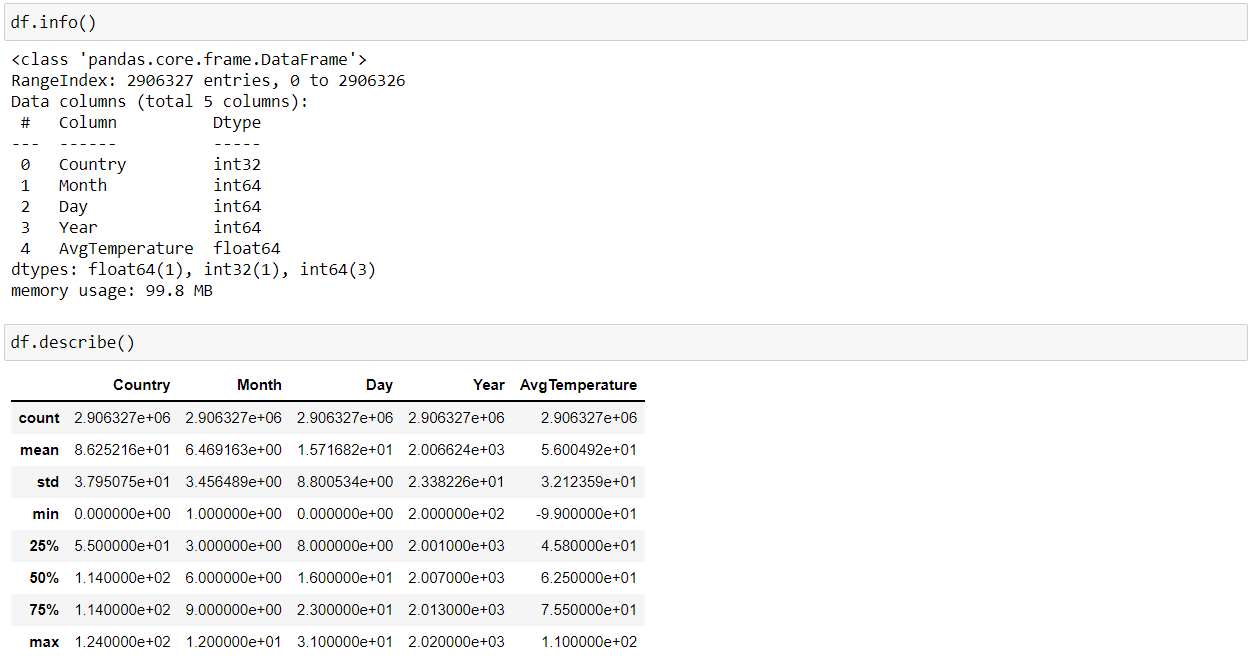


Figure 10 - Dataset Metadata.

The main objective of this research is to explore the possibility of applying Machine Learning for data imputation and to measure its effectiveness in the respective field and possibly create an algorithm for data imputation that can be practically implemented in real-world scenarios. Therefore, the dataset must contain missing data. However, upon observation, it is seen that the dataset does not have any missing data but instead assigns the value “-99” in the “AvgTemperature” feature to signify missing data. For this research, the formerly mentioned value will be replaced with “NaN” which represents the null value in Python.

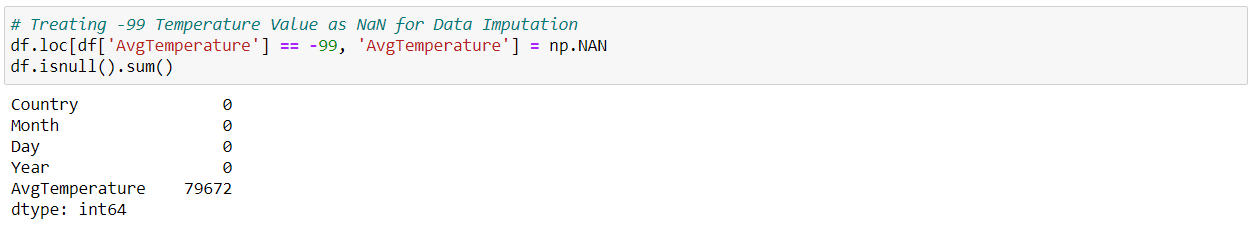


Figure 11 - Replacing -99 with "Nan”.

From Figure 11, it can be seen that there are approximately 80,000 null values that need to be imputed in the “AvgTemperature” feature and no other feature in the dataset has any null values. Therefore, the “AvgTemperature” feature is the primary concern for this research.

## Machine Learning

The basic data visualization and data preprocessing are done and the datasets are now ready for the application of various imputation techniques. The observation of the dataset shows us that the dataset contains features with multiple output classes. Since the dataset defines the type of algorithm to be used in a Machine Learning project, the algorithm used for this project will be from the classification realm of Machine Learning, i.e., the Random Forest Classifier. The algorithm will be used to impute the dataset using six different statistical techniques applied to the dataset before the process of imputation.

The machine learning algorithm and the statistical techniques will be explained in detail before diving into the implementation and the results of the Machine Learning and Statistics combination.

## Experimental Design

### Using Regression to Impute Missing Data

When we have multiple variables with missing values, we can't just directly use Regression Imputation to impute one of them as the predictors contain missing data themselves. This Catch-22 situation can be avoided by initially imputing all the variables with missing values using some trivial methods like Simple Random Imputation (we impute the missing data with random observed values of the variable) which is later followed by Regression Imputation of each of the variables iteratively.

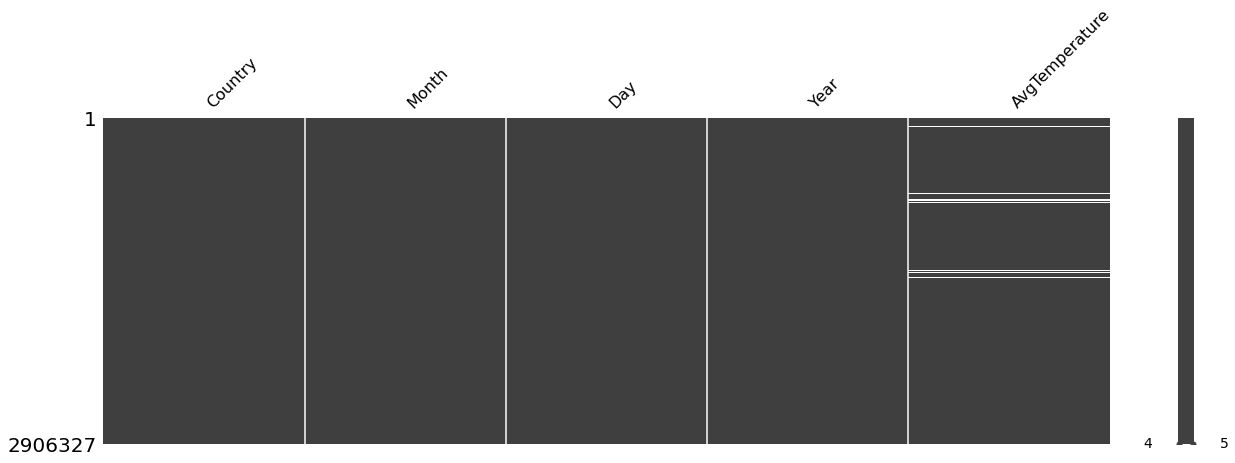


Figure 12 - Specialized Missing Number Graph to Visualize the Extent of Missing Values in the Target Feature.

Since Regression is employed in this research for Data Imputation, there are two popular methods in Regression used for Data Imputation, Deterministic Regression and Stochastic Regression. This will be explored in the experiment for this research and the effectiveness of each will be decided by using accuracy-based performance metrics.

### Deterministic Regression

In Deterministic Regression Imputation, we replace the missing data with the values predicted in our regression model and repeat this process for each variable. A major disadvantage in this method is that we reduce the inherent variability in the imputed variable. In other words, since we substitute the missing data with regression outputs, the predicted values lie along the regression hyperplane where the variable would have contained some noise/bias. We can visualize the above fact in several ways. The first one is plotting histograms for both the incomplete data and the complete data in which we can observe that the plot of the completed data is taller and narrower when compared to that of the incomplete data. In other words, the complete data has a lesser standard deviation (thus lesser variability) than the incomplete data. Another method would be plotting a boxplot in which we can observe that the IQ Range is pretty compressed for the complete data when compared to that in the incomplete data.



Figure 13 - Results of Deterministic Regression using the Specialized Missing Values Graph shows that all the Missing Values have been Accounted for.

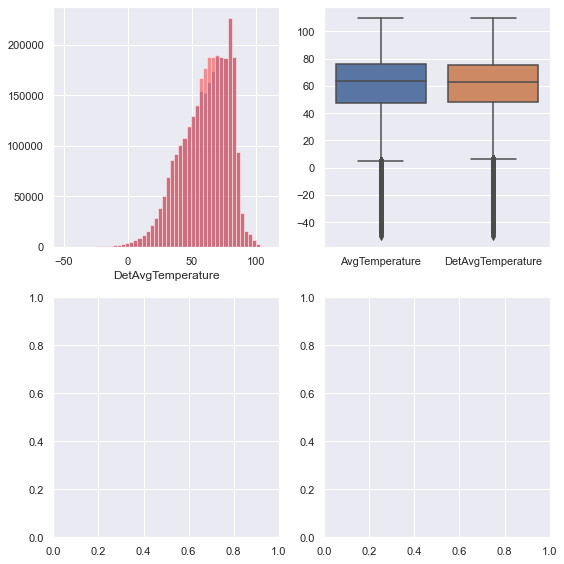


Figure 14 - A Boxplot in which we can observe that the IQ Range is pretty compressed for the complete data when compared to that in the incomplete data.

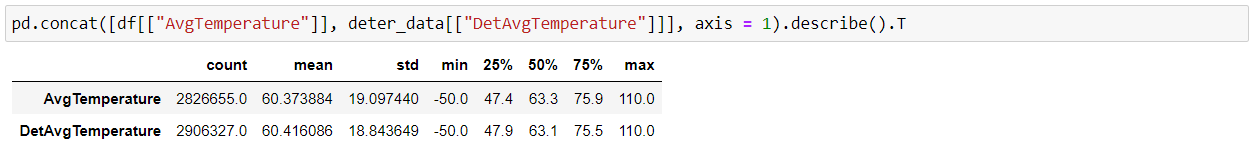


Figure 15 - Comparison of the Original "AvgTemperature" feature and the Updated "AvgTemperature" feature using Deterministic Regression.

### Stochastic Regression

To add uncertainty back to the imputed variable values, we can add some normally distributed noise with a mean of zero and the variance equal to the standard error of regression estimates. This method is called Random Imputation or Stochastic Regression Imputation. When we introduce this Gaussian noise, we may end up imputing some negative values for the missing data due to the spread of the distribution for a particular pair of mean and standard deviation. But, as per our discussion earlier, there might be some variables whose values can never be zero. For example, a negative value for Insulin concentrations would be meaningless. We can avoid this situation by retaining the values introduced by simple random imputation which is discussed above. This reduces the variability that we introduce, but it's something we have to deal with, especially in the case of these variables whose values are restricted to certain parts of the real number line.

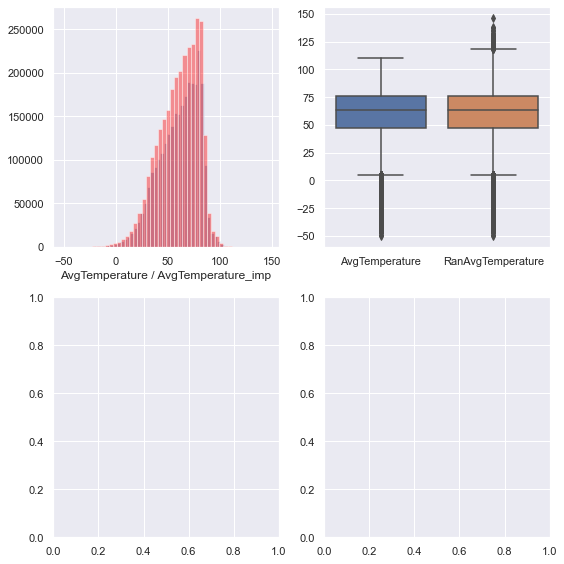


Figure 16 - A Boxplot in which we can observe that the IQ Range is pretty compressed for the complete data when compared to that in the incomplete data for stochastic regression.

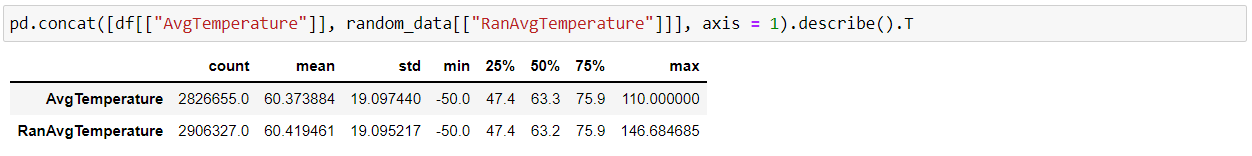


Figure 17 - Comparison of the Original "AvgTemperature" feature and the Updated "AvgTemperature" feature using Stochastic Regression.

# Conclusion

The project aimed to compare and contrast existing Machine Learning methods for data imputation to shed light on how Machine Learning can be greatly helpful in providing an excellent solution for handling missing data. The aim of the project was achieved by following defined objectives that lead to the successful completion of the research with results to boot. The objectives that were required to follow were the following:

Dataset Selection: the most critical part of this research since the complete research revolves around replacing and handling missing values in data, the requirement of a dataset for experimenting with was of monumental consequence. Since there was no requirement of predicting anything and using the prediction to solve a real-world problem, the possible route of selecting a good toy dataset to display the prowess of Machine Learning in data imputation was considered. The Titanic dataset, which recorded demographic information about the passengers of the titanic including whether the passengers survived or not was discovered to be an excellent choice as a toy dataset for this research.

Data preprocessing: The data has to be analyzed to discover the features with missing values so that the concentration of the experiment can revolve around those features. Two features from the dataset were discovered with missing values namely “Age” and “Cabin”. Since the scope of the project was not to completely rid the data of missing values but to show that Machine Learning can be used to impute missing values in the dataset, the data instances that contained missing values for the “Cabin” feature were dropped from the dataset, completely discarded. The focus of the data imputation was on the remaining feature with missing data values i.e., “age”.

Data visualization: The dataset when accessed was discovered to be already split into training and test dataset for the convenience of anyone that wishes to use it for Machine Learning. Because the dataset was pre-split, the comparison of the data instances in both the datasets was to be done to make sure that both the datasets were totally in sync and completely related. This was done by plotting probability distribution graphs for each of the features in both datasets. The distribution graphs were then analyzed and the observation determined that both the datasets were indeed in full sync. The distribution graphs however did not cover the “Survival” feature of the dataset since this feature contains binary output classes i.e., 0 and 1 and probability distributions cannot be made for this type of feature.

Machine Learning: The processes involved in this were the finalization of an algorithm, the selection of the statistical and Machine Learning methods to combine and use for data imputation. The literature review done for this research made the author discover that one of the most popular techniques for data imputation is to employ the use of a Random Forest classifier. The literature review is the reason behind the selection of the random forest classifier as the main testing algorithm for this research. But the Random Forest classifier is not used directly in the process of data imputation. The algorithm used for data imputation is Linear Regression. Two forms of Linear Regression are used for data imputation i.e., Deterministic Regression and Random Regression. The other techniques that were used for Data imputation were mean imputation and the Iterative Imputer. The random forest classifier is used as the performance metric for the testing of the data imputation techniques applied in this research.

Results were generated graphically using the 10-fold cross-validation for accurate error estimation. The results were a little surprising, especially the behaviour of Random regression and the behaviour of baseline data. The observations and their interpretation are discussed in the Results subsection 5.b.3.

Having the option to successfully attribute missing information is of extraordinary significance to researchers working with genuine information today. An AI strategy like RF, known for its great forecast execution and capacity to deal with all structures of information, addresses a potentially appealing answer for this testing issue. In any case, because no orderly relative investigation of RF had endeavoured in missing information settings, we attempted an enormous scope trial investigation of various RF ascription strategies to figure out which techniques performed best, and under what sorts of settings.

Missing information causes critical data misfortune in examinations as data is squandered and understanding can't be acquired into the fundamental reasons for the on attendance. Using information coming about because of an HIV seroprevalence overview, this paper examines and looks at five AI ideal models to credit missing information: RFs, AANN-GA, AANF-GA, RF-AANN-GA and AANN-GA-RF. It is obvious from the introduced results that the RF calculation as a relapse framework to attribute the missing information beats different ideal models examined for the considered informational collection. This is valid for both calculation time and calculation precision, with RFs beating different ideal models by up to 32 % on normal for certain classes.

# Future Work

An issue of concern is that Regression Imputation might not serve as the best method when a variable is the missing majority of its data, as in the case of insulin. In these cases, we have to use more powerful approaches as Maximum Likelihood Imputation and Multiple Imputation:

* Regression Imputation assumes that the data are Missing At Random.
* For a better Regression model, we might have to follow different Data Transformation methods depending on our data.
* Do observe that we have included Outcome as one of our predictors even though it is caused by the other variables under scrutiny.
* This notebook does not describe the best method for many cases, rather it just demonstrates Regression Imputation as one of the methods.

# 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.
13. Pantanowitz, A. and Marwala, T., 2009. Missing Data Imputation Through the Use of the Random Forest Algorithm. Advances in Intelligent and Soft Computing, pp.53-62.
14. Shataee, S., Kalbi, S., Fallah, A. and Pelz, D., 2012. Forest attribute imputation using machine-learning methods and ASTER data: comparison of k-NN, SVR and random forest regression algorithms. International Journal of Remote Sensing, 33(19), pp.6254-6280.
15. Stekhoven, D. and Buhlmann, P., 2011. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), pp.112-118.
16. Tang, F. and Ishwaran, H., 2017. Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal, 10(6), pp.363-377.
17. Li, D., Deogun, J., Spaulding, W. and Shuart, B., 2004. Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method. Rough Sets and Current Trends in Computing, pp.573-579.
18. Tang, F. and Ishwaran, H., 2017. Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data Science Journal, 10(6), pp.363-377.
19. Breiman, L., 2001. Machine Learning, 45(1), pp.5-32.