# ABSTRACT

In supervised learning, missing values usually appear in the training set. The missing values in a dataset may generate bias, affecting the quality of the supervised learning process or the performance of classification algorithms. These imply that a reliable method for dealing with missing values is necessary. In this paper, we analyze the difference between iterative imputation of missing values and single imputation in real-world applications. We propose an EM-style iterative imputation method, in which each missing attribute-value is iteratively filled using a predictor constructed from the known values and predicted values of the missing attribute-values from the previous iterations. Meanwhile, we demonstrate that it is reasonable to consider the imputation ordering for patching up multiple missing attribute values, and therefore introduce a method for imputation ordering. We experimentally show that our approach significantly outperforms some standard machine learning methods for handling missing values in classification tasks.

# TABLES OF CONTENTS

## LIST OF FIGURES



## LIST OF TABLES

# LIST OF ABBREVIATIONS

| ABBREVIATION | EXPANSION |
| --- | --- |
| MCAR | Missing Completely at Random |
| MAR | Missing at Random |
| NMAR | Not Missing at Random |
| K-NN | K-Nearest Neighbor |

# CHAPTER 1
# INTRODUCTION

The missing data problem is arguably the most common issue encountered by machine learning practitioners when analyzing real-world data. In many applications ranging from gene expression in computational biology to survey responses in social sciences, missing data is present to various degrees. As many statistical models and machine learning algorithms rely on complete data sets, it is key to handle the missing data appropriately. Missing data problem is a common issue in most real-world studies. Since most statistical models and data-dependent machine learning (ML) algorithms could only handle complete data sets, the issue of how to approach missing values plays an important role in statistical inferences. Let Y be an (N × K) data matrix with i-th row yi = (yi1, yi2, ..., yiK) where yij is the value of j-th feature for the i-th sample. Define the subset of observed values as Y obs and missing values as Y mis. Also, let M = [mij ] be the missing indicator matrix, where mij indicates whether yij is missing or not. Rubin (19776) defines three different missing mechanisms according to the conditional probability of the missingness, {mij = 1}, given the data. The mechanism of missing data is completely at random (MCAR) if the probability of missingness is independent of all data values, missing or observed, P(mij = 0|Y ) = g(ϕ), i = 1, ..., N, j = 1, ..., K, where g(.) is a known link function and ϕ is the vector of unknown mechanism parameters. The missing mechanism is called missing at random (MAR) if the probability of missingness depends only on the observed data values, P(mij = 0|Y ) = g(Y obs; ϕ), i = 1, ..., N, j = 1, ..., K. Finally, the mechanism is called missing not at random (MNAR) when the probability of missingness may also depend on the unobserved data even after conditioning on the observed ones. The missing mechanism for the likelihood inferences is ignorable when the MCAR or MAR assumptions hold with the additional condition of disjoint parameter spaces of the missing mechanism and the data model (see Little and Rubin, 2014; Tsiatis, 2007, for more details). One simple approach to analyze incomplete data is complete case (CC) analysis which discards all incomplete cases. This approach is logical only if the missing rate is considerably small or the missing data mechanism is MCAR (Little and Rubin, 2014). However, if the missing mechanism is MAR or MNAR or the missing rate is considerably high, the CC approach could highly influence statistical results. This is due to the fact that CC analysis makes no use of observed features of an incomplete case.

# CHAPTER 2

# LITERATURE SURVEY

1) **To generalize to multivariate settings, a chained equation process initializing using random sampling (IEEE 2011):**

Buuren and Groothuis-Oudshoorn, 2011 Missing completely at random (MCAR). This is the highest level of randomness. It occurs when the probability of an instance (case) having a missing value for an attribute does not depend on either the known values or the missing data. In this level of randomness, any missing data treatment method can be applied without risk of introducing bias on the data; 2. Missing at random (MAR). When the probability of an instance having a missing value for an attribute may depend on the known values, but not on the value of the missing data itself; 3. Not missing at random (NMAR). When the probability of an instance having a missing value for an attribute could depend on the value of that attribute. Several methods have been proposed in the literature to treat missing data. Many of these methods, such as case substitution, were developed for dealing with missing data in sample surveys, and have some drawbacks when applied to the Data Mining context. Missing data imputation can be harmful because even the most advanced imputation method is only able to approximate the actual (missing) value. The predicted values are usually more well-behaved, since they conform with other attributes values.

2) **These decision tree based imputation methods are non-parametric approaches that do not rely upon distributional assumptions on the data(IEEE 2012) :**

s(Stekhoven and B¨uhlmann, 2012) This method imputes each missing entry x_id as the mean of the dth dimension of the K-nearest neighbors that have observed values in dimension. A best attribute is selected to place at the root node of the tree and create one child node for each possible value of this selected attribute. For each child node, if

it isn't a leaf node, the entire process is then repeated recursively only using those training instances that actually reach this node. . Bayesian networks are often used for the classification problems, in which a learner attempts to construct Bayesian network classifiers from a given set of training instances with class labels. Assume that all attributes are fully independent given the class, then the resulting Bayesian network classifiers are called naive Bayesian classifiers (simply NB). We run our experiments on 36 UCI datasets published on the main web site of Weka platform [21], which represent a wide range of domains and data characteristics. We downloaded these data sets in the format of arff from the main web site of Weka. Due to the simplicity, effectiveness, and efficiency, C4.5 and NB are two very important algorithms for addressing the classification problems. In this paper, we propose a very simple, effective, and efficient algorithm based on C4.5 and NB. We simply denote it C4.5-NB. In C4.5-NB, C4.5 and NB are built and evaluated independently at the training time, and the class-membership probabilities are weightily averaged according to their classification accuracies.

**3) Evidence from recent literature suggests that recent advances in optimization have driven significant progress in machine learning. (IEEE 2017):**

s (Bertsimas and Van Parys, 2017; Bertsimas and Mazumder, 2014), Integer and convex optimization have been applied successfully but it does apply for median and sparse regression problems. Despite the frequent occurrence and the relevance of missing data problem, many Machine Learning algorithms handle missing data in a rather naive way. However, missing data treatment should be carefully thought, otherwise bias might be introduced into the knowledge induced. In this work we analyses the use of the k-nearest neighbor as an imputation method. Imputation is a term that denotes a procedure that replaces the missing values in a data set by some plausible values. One advantage of this approach is that the missing data treatment is independent of the learning algorithm used.Missing data imputation can be harmful because even the most advanced imputation method is only able to approximate the actual (missing) value. The predicted values are usually more well-behaved, since they conform with other attributes values. In the experiments carried out, as more attributes

with missing values were inserted and as the amount of missing data increased, more simple were the induced models. Our analysis indicates that missing data imputation based on the k-nearest neighbor algorithm can outperform the internal methods used by C4.5 and CN2 to treat missing data, and can also outperform the mean or mode imputation method, which is a method broadly used to treat missing values.

**4) We reconsider the missing data problem from this perspective(IEEE 2017)**

n (Bertsimas and Dunn, 2017; Bertsimas et al., 2017) . order to develop optimization-based methods for imputation with improved out-of-sample performance. Imputation techniques replace missing values with approximated ones depending on the data set. Options range from simple approaches like mean imputation to more complex methods based on attribute correlations. Here are some commonly used imputation methods: 1. Case swapping Most sample surveys employ this strategy. A nonsampled instance replaces a missing data instance (for example, a person who cannot be reached); 2. This is a popular strategy. It involves replacing missing data for a property with the mean (quantitative) or mode (qualitative) of all known values; 3. Hot/cold deck. The hot deck approach replaces a missing attribute value with a value from an approximated distribution for the current data. Hot deck usually comes in two phases. The data are first clustered. In the second step, each missing data instance is assigned to a cluster. A cluster's full cases are utilised to fill in missing data. Calculate the attribute's mean or mode within a cluster. 4. Prediction model. Prediction models handle missing data in complex ways. These strategies include building a predictive model to estimate values to replace missing data. It is utilised as a class attribute, and the rest as input for the prediction model. This method is supported by the fact that qualities usually have links (correlations). A predictive model for classification or regression for qualitative and quantitative features with missing data might be built using correlations. Some of these links may be preserved if the prediction model captures them. Because the missing values are anticipated using a set of qualities, the predicted values are likely to be more consistent with this set of attributes than the genuine (unknown) values. The second flaw is the necessity for attribute correlation

## 2.1 EXISTING SYSTEM:

- In data mining process the biggest task of data preprocessing is missing value imputation. Imputation is a statistical process of replacing missing data with substituted values.
- Many clinical diagnostic dataset are usually incomplete. Excluding incomplete dataset from the original dataset can bring more problem than simplification. In this paper the machine learning techniques for missing value imputation have been explored using Ionosphere data from UCI repository.
- The data imputation problem has been approached using well-know machine learning techniques.
- The experiments have shown that the final classifier performance when the algorithm is used. Experiments show that popular machine learning classifier techniques were found to outperform than standard mean/mode imputation techniques.

## 2.2 PROPOSED SYSTEM:

- In this proposed work, a Python-based data mining system capable of diagnosing the HD using a Decision Tree has been developed. In the methodology, the UCI data repository was taken into consideration with 14 Attributes.
- In the dataset, there are few missing values (yet found to be hyperparameter), and pre-processing with such missing values is a common yet challenging problem.
- A mere substitution will give biased results from the data to be observed for HD diagnosis and will certainly affect the value of the learning process in Machine Learning.
- Therefore, in the proposed work, a missing value imputation is done, which gave better accuracy, and it is trustable.
- In this approach, we impute each missing data attribute value by predicting its data value from non-missing data attributes.

# CHAPTER 3

# METHODOLOGY

## 3.1 MACHINE LEARNING

### Steps of Machine Learning

- Step 1: Gathering Data.
- Step 2: Preparing that Data.
- Step 3: Choosing a Model.
- Step 4: Training.
- Step 5: Evaluation.
- Step 6: Hyper parameter Tuning.
- Step 7: Prediction.

**Introduction:**

In this blog, we will discuss the workflow of a Machine learning project this includes all the steps required to build the proper machine learning project from scratch.

We will also go over data pre-processing, data cleaning, feature exploration and feature engineering and show the impact that it has on Machine Learning Model Performance. We will also cover a couple of the pre-modelling steps that can help to improve the model performance. Python Libraries that would be need to achieve the task:

1) Numpy
2) Pandas
3) Sci-kit Learn
4) Matplotlib

We can define the machine learning workflow in 3 stages.

1) Gathering data
2) Data pre-processing
3) Researching the model that will be best for the type of data
4) Training and testing the model
5) Evaluation

The machine learning model is nothing but a piece of code; an engineer or data scientist makes it smart through training with data. So, if you give garbage to the model, you will get garbage in return, i.e., the trained model will provide false or wrong prediction

**1. Gathering Data**

The process of gathering data depends on the type of project we desire to make, if we want to make an ML project that uses real-time data, then we can build an IoT system that using different sensors data. The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done.We can also use some free data sets which are present on the internet. **Kaggle** and **UCI Machine learning Repository** used the most for making Machine learning models. Kaggle is one of the most visited websites that is used for practicing machine learning algorithms, they also host competitions in which people can participate and get to test their knowledge of machine learning.