MISSING VALUES IMPUTATION AND FUTURISTIC PREDICTION USING DEEP LEARNING TECHNIQUE

A Thesis

Submitted in fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

By

ASHOK KUMAR TRIPATHI

Enrollment No. 186207

IN

COMPUTER SCIENCE & ENGINEERING

Under the Supervision of

Dr. PRDEEP KUMAR GUPTA

(Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING AND INFORMATION TECHNOLOGY
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGYWAKNAGHAT SOLAN-173234,
HIMACHAL PRADESH, INDIA

Under the External- Supervision of

Dr. HEMRAJ SAINI

(Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING, DEHRADUN INSTITUTE OF

TECHNOLOGY, INDIA

Under the External - Supervision of

Dr. GEETANJALI RATHEE

(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING, NETAJI SUBHAS UNIVERSITY

OF TECHNOLOGY, INDIA

AUGUST 2024

DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the Ph.D. thesis entitled "**Missing Values Imputation And Futuristic Prediction Using Deep Learning Technique**" submitted at the Jaypee University of Information Technology, Waknaghat, Himachal Pradesh, India is an authentic record of my work carried out under the supervision of Dr.Pradeep Kumar Gupta. I have not submitted this work elsewhere for any degree or diploma. I am fully responsible for the contents of my Ph.D. thesis.

(Signature of the Scholar)

Ashok Kumar Tripathi

Enrollment No.: 186207

Department of Computer Science& Engineering

Jaypee University of Information Technology, Waknaghat, Solan

(HP), India

Date: 17/08/2024

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled "**Missing Values Imputation And Futuristic Prediction Using Deep Learning Technique**" submitted by Ashok KuamrTripathi at Jaypee University of Information Technology, Waknaghat, Himachal Pradesh, India, is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

(Signature of Supervisor)

Dr.Prdeep Kumar Gupta

(Professor)

Department of Computer Science &Engineering

Jaypee University of InformationTechnology,

Waknaghat, Solan (HP), India

EXTERNAL SUPERVISOR-1 CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled"**Missing Values Imputation And Futuristic Predication Using Deep Learning Technique**" submitted by Ashok KuamrTripathi at Jaypee University of Information Technology, Waknaghat, Himachal Pradesh, India, is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

(Signature of External Supervisor)

Dr. Hemraj Saini
(Professor)

Department of Computer Science &Engineering

Dehradun Institute Of Technology, India

Dehradun Uttarakhand, India

EXTERNAL SUPERVISOR-2 CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled"**Missing Values Imputation And Futuristic Predication Using Deep Learning Technique**" submitted by Ashok KuamrTripathi at Jaypee University of Information Technology, Waknaghat, Himachal Pradesh, India, is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

(Signature of External Supervisor)

Dr. Geetanjali Rathee
(Assistant Professor)
Department Of Computer Science & Engineering,

Netaji Subhas University Of Technology, India

ACKNOWLEDGEMENT

Indeed, the words at my command are inadequate, either in the form of spirit or express the depth of my humility and kindness before the Almighty one, without whose endless benevolence and blessings this tedious task could not have been accomplished. I would first like to thank Dr.Pradeep Kumar Gupta, Professor, Department of Computer Science Engineering and Information Technology, my supervisor, for giving me this opportunity to conduct my Ph.D. research work under his supervision at the Department of Computer Science and Information Technology. He has constantly encouraged and supported me throughout my stay at JUIT. His vision of science and technology has motivated me to do high-quality work. Because of his active support, I can pursue research that interests me and my work on time. He gave me helpful advice about research, career, and life, which will be more valuable in my future professional career. Because of his availability, I never felt a lack of guidance. It has been a wonderful experience to work under his supervision. I found that the best are efficient, effective, and results-driven in them, but at "core, they are persons with the best qualities as human beings." Humanity, scientific potential, kindness, coolness, simplicity, and the novelty of ideas are the arms of their great personality. I am grateful and indebted to Dr.HemrajSaini for his invaluable, painstaking efforts towards my study and for showing me the absolute path of sincerity and dedication.

I am also grateful to Prof. (Dr.) Vivek Kumar Sehgal, Head of the Department of Computer Science and Engineering & IT, for his insightful comments and administrative help on various occasions. I extend my sincere thanks to my DPMC members, Prof. (Dr.) Vivek Kumar Sehgal, Prof. (Dr.) Sunil Khah, and

Dr. Rakesh Kanji, for their stimulating questions and valuable feedback. I thank the other department faculty members for their valuable feedback and support.

I would like to express my gratitude to Dr.Geetanjali, Dr.Ravinder Bhatt, Dr. Rajinder Sandhu, and other faculty members for their feedback, moral support, valuable suggestions, and necessary facilities during my research work.

A formal acknowledgment of my emotions is inadequate to convey the depth of my love and affection for my reverend father, Late Shri R. P. Tripathi, and mother, Shm. Kamla Tripathi, for their prudent persuasion, selfless sacrifice, and heartfelt blessings, have enabled me to translate their dreams into reality.

Despite all these, I can never forget to mention my adorable son, Aditya Tripathi, and daughter, Aradhya Tripathi, for their love and affection, which have given me constant strength throughout my studies. I am grateful to my wife, Usha Tripathi, for being patient and prioritizing my research.

This note of acknowledgment will always be complete with the mention of my Sisters and Brothers-in-law, Dr. Sanjay Pandey, Mrs. Shashi Pandey, Mr. Arun Mishra, and Mrs. Sushmita Mishra, for their encouragement and never-ending help during the entire course of I thank my younger brother, Ashutosh Tripathi, for his love and care. He always remained by my side during happy and challenging times to motivate me.

Finally, I acknowledge my friend, Mr. Amit Kumar Shrivastava, and his wife, Mrs. Sweta Shrivastava, for supporting me. It will never be possible for me to fully repay the price of my friends' sacrifices, encouragement, and never-ending help throughout the entire course of study.

Finally, I thank anyone else whose contribution I could have forgotten.

TABLE OF CONTENTS

DECLARATION BY THE SCHOLAR……………………………………iii

SUPERVISOR'S CERTIFICATE…………………………………...……..iv

EXTERNAL SUPERVISOR 1 CERTIFICATE……………………………v

EXTERNAL SUPERVISOR 2 CERTIFICATE……………………………vi

ACKNOWLEDGEMENT……………………………………………………vii

LIST OF TABLES……………………………………………………………xii

LIST OF FIGURES……………………………………………………..…xiii

LIST OF ACRONYMS…………………………………………………....xv

ABSTRACT……………………………………………………………..…xviii

Chapter-1 Introduction 1

1.1 Introduction to Missing Values Imputation……………………………1

1.2 Analysis of Missing Value Imputation and their Assessment………….3

1.3 Missing data mechanisms……………………………………………...…4

1.4 Why do we need to care about handling Missing data?...................... ...5

1.5 Process of missing data imputation…………………………………….....7

1.6 How to Handle Missing Values…………………………………….…8

1.7 Introduction to Deep Learning Techniques…………………………....9

1.8 Problem Statement…………………………………………………10

1.9 Challenges………………………………………………………………..12

1.10 Objectives of the Research……………………………………………13

1.11 Contributions……………………………………………………………14

1.12 Thesis Organization……………………………………….……..…15

Chapter-2 Literature Review 18

2.1 Introduction…………………………………………………….....18

2.2 Data Mining……………………………………………………....18

2.2.1Descriptive data mining……………………………………...…19

2.2.2 Predictive data mining……………………………………...…21

2.3 History of data mining…………………………………………...….22

2.4 Techniques of data mining…………………………………….…24

2.5 Data Mining Process……………………………………...……26

2.6 Missing Values Imputation is an important tool for data quality……27

2.7 Missing Values Imputation Methodologies……………………….....29

2.8 Present Trends in Missing Value Imputation…………………...……32

2.9 Summary of the Chapter…………………………………….……34

Chapter-3 Missing Values Imputation Predictive Modelling 36

3.1 Introduction…………………………………………………...36

3.2 Approach of Missing Data Mechanism……………………………36

3.2.1 Structurally Missing Data……………………………………...37

3.2.2 Missing Completely at Random (MCAR)……………………….37

3.2.3 Missing at Random (MAR)……………………………………38

3.2.4 Missing not at random (nonignorable)…………………………39

3.3 Exploring Data Missingness…………………………………………40

3.4 Treating Missing Values…………………………………………….43

3.5 Impact of Missing Data on Predictive Models…………………..…45

3.6 How missing data can affect model performance and accuracy……..46

3.7 Model Implementation Process……………………………………..…47

3.8 Bias and potential pitfalls introduced by missing data…………..…48

3.9 Machine Learning-Based Imputation…………………………..……50

3.9.1 Process of machine learning-based imputation…………………52

3.9.2 Leveraging predictive modeling for imputing missing values.…53

3.10 Deep Learning-Based Imputation:……………………………………54

3.11 Introduction to using neural networks for imputation………………57

3.12 Auto encoders and their role in imputing missing values……………59

3.13 Summary of the Chapter…………………………………………..…..61

Chapter-4 Futuristic Prediction of Food Consumption with Missing Value Imputation Methods using Extended ANN **62**

4.1 Introduction……………………………………………………………….62

4.2 Conventional techniques for borrowing food consumption databases….62

4.3 Proposed Phenomenon……………………………………………………63

4.3.1 Data searching and analysis……………………………………...65

4.4 Multivariate imputation by chained equations (MICE)……………….…….68

4.4.1 Here are the key roles of MICE in missing values imputation…...69

4.4.2 MICE has several advantages…………………………………..…72

4.4.3 Experiment and Performance Analysis…………………………..…73

4.4.4 Results and Comparison…………………………………………82

4.5 Summary of the Chapter………………………………………………..…..85

Chapter-5 MVI and Forecast Precision Upgrade of Time Series Precipitation **86**

5.1 Information for Ubiquitous Computing………………………………...86

5.2 Missing Value Imputation and Prediction of Rainfall as a Case Study...89

5.3 Types of Missing Valuea - Literature Review…………………………94

5.4 Proposed Mechanism to Handle a Mission Value…………………..…95

5.4.1 Feature Engineering…………………………………………….......99

5.5 Result Analysis of Proposed Mechanism……………………………..100

5.6 Summary of the Chapter………………………………………………108

Chapter-6 Conclusion And Future Directions **110**

6.1 Conclusion…………………………………………………………...…110

6.2 Future Directions………………………………………………………..112

List of Publications **113**

References **114**

LIST OF TABLES

Table 2.1	Overview of Related Techniques Missing Data Imputatio	**29**
Table 3.1	Missing data of a structural nature	**37**
Table 3.2	Data missing completely at random (MCAR)	**38**
Table-3.3	Missing at random (MAR)	**39**
Table 4.1	Data containing potassium values for various foods sourced from multiple national FCDBs	**66**
Table 4.2	Dataset with missing values	**73**
Table 4.3	Dataset with true values to verify the model output	**74**
Table 4.4	Dataset with missing values to verify the model output	**74**
Table 4.5	Dataset with applying the mean imputation method	**75**
Table 4.6	“ Zeroth” dataset for iteration 1	**77**
Table 4.7	Missing value dataset for iteration1	**78**
Table 4.8	Difference between the first two datasets	**79**
Table 4.9	After the first iteration, predict values	**80**
Table 4.10	First Dataset, Second Dataset, and Difference Matrix	**80**
Table 4.11	Second Dataset, Third Dataset, and Difference Matrix	**81**
Table 4.12	Third Dataset, Forth Dataset and Difference Matrix	**81**
Table 4.13	Final imputed values	**81**

LIST OF FIGURES

Figure 1.1	A standard experimental setup for Missing Imputation (MVI) procedures involves filling in missing values within various attributes	**2**
Figure 1.2	The organized tree exhibition of the commonly used MVI methods	**3**
Figure 1.3	Analysis of Missing Value Imputation and assessment	**4**
Figure 2.1	Architecture of a data mining system	**19**
Figure 2.2	Data Mining Technique	**24**
Figure 2.3	Data Mining Process Diagram	**27**
Figure 3.1	The Python Notebook 3 code merged both datasets and generated a plot illustrating the missing value matrix	**40**
Figure 3.2	The Missingno. bar chart visually represents the nullity within the dataset	**41**
Figure 3.3	Nullity matrix provides a data-dense display to identify the dataset's missing data patterns	**41**
Figure 3.4	Y-axis data to enhance the visualization of characteristics with significantly high missing values	**42**
Figure 3.5	A basic heatmap of correlation that illustrates the degree of nullity in the correlation between the various features	**42**
Figure 3.6	A hierarchical diagram illustrating various ML models categorized into five groups based on their similarities	**50**
Figure 4.1	Imputation of missing values using ANN and KNN	**83**
Figure 4.2	Forecasting the consumption of red coffee across various countries	**84**
Figure 4.3	Values imputed using KNN, extended KNN, and RNN algorithms	**84**
Figure 4.4	Imputation value using extended ANN and ANN	**85**
Figure 5.1	Monthly Rainfall Status in India	**92**
Figure 5.2	Display Missing Values in India rainfall dataset	**93**
Figure 5.3	Process of Extended Kalman for MVI	**98**

Figure 5.4	illustrates the variations in rainfall amounts across diverse locations over six years spanning from 2007 to 2012.	**100**
Figure 5.5	Presents a comparative analysis, year by year, of missing values imputation using original values alongside the Kalman filter and extended Kalman filter techniques (a) the focus is on the year 2007; (b) it shifts to 2008; (c) centers on 2009; (d) highlights 2010; (e) centers around 2011, while (f) is directed towards 2012. This analysis spans from 2007 to 2012 in Bihar state, examining various locations	**101**
Figure 5.6	Predicted values compare with the original values	**102**
Figure 5.7	Predicted values compare with original values using Optimizer	**104**
Figure 5.8	Predicted values compare with original values using RMSprop optimizer	**105**
Figure 5.9	Predicted values compare with original values using ADAM optimizer	**106**
Figure 5.10	Predicted values compare with the original values	**107**
Figure 5.11	Predicted ADAM Optimizer values compare with the original values	**108**

LIST OF ACRONYMS

EKF	ExtendedKalman Filter
KF	Kalman Filter
ANN	Artificial Neural Network
MI	Multiple Imputation
EANN	Extended Artificial Neural Network
CNN	Convolutional Neural Networks
MV	Missing Values
CD	Case Deletion
MCI	Most Common Imputation
MVI	Missing Values Imputation
EMMVI	Expectation Maximization-based MVI
MMVI	Mean/ Median/ Mode-based MVI
LLSMVI	Local Least Square-based MVI
BPCAMVI	Bayesian Principal Component Analysis based MVI
LRMVI	Linear/ Logistic Regression-based MVI
AI	Artificial Intelligence
KDD	Knowledge Discovery in Databases
ML	Machine Learning
IOT	Internet of Things

SVM	Support Vector Machines
GSP	Generalized Sequential Pattern
VAE	Variational Auto Encoders
GAN	Generative Adversarial Networks
RNN	Recurrent Neural Networks
LSTM	Long Short-Term Memory
FNN	Feedforward Neural Networks
KNN	K-Nearest Neighbour
FCDBs	Food Composition Databases
RMSE	Root Mean Squared Error
MICE	Multivariate Imputation by Chained Equations
FCS	Fully Conditional Specification
NDRRMC	National Disaster Risk Reduction and Management
AWS	Automated Weather Stations
IPCC	Intergovernmental Panel on Climate Change
EDA	Exploratory Data Analysis
SGD	Stochastic Gradient Descent
RSMProp	Root Mean Squared Propagation
ADAM.	Adaptive Moment Estimation

MCAR	Missing Completely at Random
MAR	Missing at Random
MNAR	Missing Not at Random
EM	Expectation Maximization
MLE	Maximum Likelihood Estimation
MLP	Multilayer Perceptron

LIST OF TABLES

Table 2.1	Overview of Related Techniques Missing Data Imputatio	**29**
Table 3.1	Missing data of a structural nature	**37**
Table 3.2	Data missing completely at random (MCAR)	**38**
Table-3.3	Missing at random (MAR)	**39**
Table 4.1	Data containing potassium values for various foods sourced from multiple national FCDBs	**66**
Table 4.2	Dataset with missing values	**73**
Table 4.3	Dataset with true values to verify the model output	**74**
Table 4.4	Dataset with missing values to verify the model output	**74**
Table 4.5	Dataset with applying the mean imputation method	**75**
Table 4.6	" Zeroth" dataset for iteration 1	**77**
Table 4.7	Missing value dataset for iteration1	**78**
Table 4.8	Difference between the first two datasets	**79**
Table 4.9	After the first iteration, predict values	**80**
Table 4.10	First Dataset, Second Dataset, and Difference Matrix	**80**
Table 4.11	Second Dataset, Third Dataset, and Difference Matrix	**81**
Table 4.12	Third Dataset, Forth Dataset and Difference Matrix	**81**
Table 4.13	Final imputed values	**81**

ABSTRACT

In data analysis, missing values pose a significant challenge, potentially leading to biased results and reduced statistical power. Various methods for imputing missing values have been developed to address this issue, ranging from simple imputation techniques to sophisticated algorithms based on machine learning. This thesis comprehensively reviews the existing literature on missing values imputation, discussing the advantages and limitations of different approaches. Additionally, it explores recent advancements in the field and identifies promising directions for future research. Researchers and practitioners can make informed decisions when handling missing data in their analyses by understanding the strengths and weaknesses of various imputation methods.

The primary aim of the first objective is to formulate a framework for an extensive period; data mining has persisted as a pivotal and compelling realm of research, accompanied by numerous challenges. Among these challenges, missing values within datasets emerge as a significant hurdle. This objective delves into the taxonomy of missing data, exploring diverse handling techniques.

The second objective is to tackle the challenges associated with Missing data, which is pervasive across various research fields, introducing uncertainty into data analysis. It can arise from diverse sources such as mishandling of samples, unavailability of observations, measurement errors, deletion of outliers, or simply gaps in the study. The realm of nutrition is no exception to this issue. Due to gaps in food consumption data, knowledge still needs to be completed, limiting its utility for dietary assessment, which typically requires complete datasets. Commonly, this challenge is addressed through manipulative techniques or

borrowing data from similar databases, introducing significant errors. Our study explores missing data imputation methods, including Recurrent Neural Networks, Iterative KNN imputation, K-nearest neighbors, and Artificial Neural Networks. It compares them with traditional techniques such as mean and median imputation. We utilize datasets from national Food Composition Databases collected by OpenMV.net.

The study's third and final objective aims to provide missing values, which poses a significant challenge in time-series datasets and profoundly impacts dataset analysis. Effective handling of missing values is crucial for robust analysis in ubiquitous computing. Typically, missing values are approximated using Non-linear Principal Component Analysis, with room for improvement. Utilizing the Kalman filter with the ARIMA model for imputation presents a promising approach, which can be further enhanced through Extended Kalman filtering.

Additionally, rainfall prediction employing LSTM with various optimizers, including stochastic gradient descent (SGD), RMSProp, and ADAM, is conducted. Comparative predictions demonstrate that the combination of Extended Kalman imputation, LSTM, and ADAM optimizer outperforms others. This research proposes an enhanced Extended Kalman Filter (EKF) for missing values imputation, leveraging its robust predictive capabilities initially developed for the Apollo Mission. The proposed EKF accurately estimates rainfall patterns even without data, aiding in weather prediction.

Chapter-1 Introduction

1.1 Introduction to Missing Values Imputation

Missing Values Imputation is a method used in data mining and investigation to handle missing or imperfect data in the dataset. Missing values can arise for several reasons, such as mistakes in the data collection time, malfunctions in equipment, or simply due to the nature of the data itself. When dealing with missing data, addressing them appropriately is imperative if ignoring or mishandling missing values can lead to discriminatory or inaccurate results.

Missing values can disrupt the analysis and modeling process since many algorithms and statistical methods require complete data to function effectively. Imputation refers to the process of estimating or imputing missing values with plausible values based on the available facts. The goal is to create a more complete dataset while reducing the effect of the missing values on the examination.

Data mining and data analysis and processing are recognized as essential and stimulating accountability for various applications in daily life, where a specific database aimed at a preferred problem is accrued to conduct such as examination. A database is essential to any understandable decision-making system for automated regression or classification tasks. However, real-world databases often come with challenges, such as a notable proportion of missing values, redundancy in one or more attributes, and irregular patterns (outliers). These issues need to be addressed to enhance the effectiveness of generic trained techniques. The missing values within the dataset can manifest as NaNs, blank cells, 'nan,' or occasionally as placeholders like '-999,' undefined, null, among other conventions.

Numerous factors contribute to the presence of missing values stemming from diverse sources within datasets. These may include inappropriate and erroneous data entries, data unavailability, challenges in data gathering, lost sequences, imperfect features, missing files, incomplete information, and various other sources. Addressing these challenges is crucial to ensuring the robustness and reliability of data for meaningful observations and analyses.

Figure 1.1 displays the experimental block diagram illustrating the generic Missing value Imputation (MVI) technique. This technique separates each incomplete dataset into complete and missing sets. The full dataset is then employed for parameter learning and estimation, using one of several MVI techniques (refer to Figure 1.2), to substitute missing values in the incomplete dataset. Subsequently, a straightforward assessment involves estimating the disparities between the actual and imputed values to evaluate the imputation methods. An alternative approach employs the resulting complete dataset for tasks such as classification or clustering, followed by examining the attained metrics. The literature review reveals numerous MVI methods, broadly categorized into statistical and Machine Learning (ML)-based techniques, as illustrated in Figure 1.2. These categories are further subdivided into various algorithm types (as seen in Figure 1.2), facilitating comparative discussions.

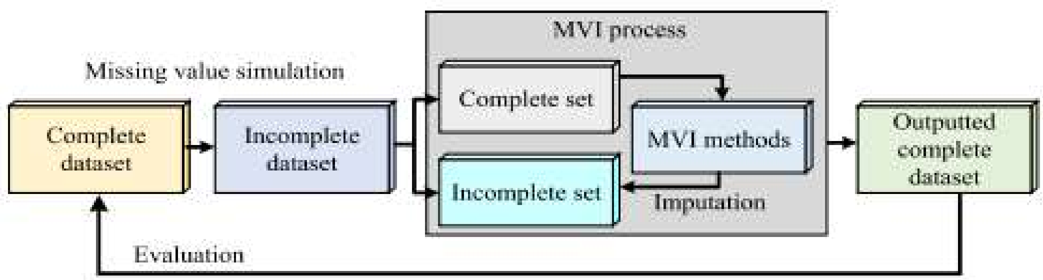


Fig. 1.1 A standard experimental setup for Missing Imputation (MVI) procedures involves filling in missing values within various attributes.

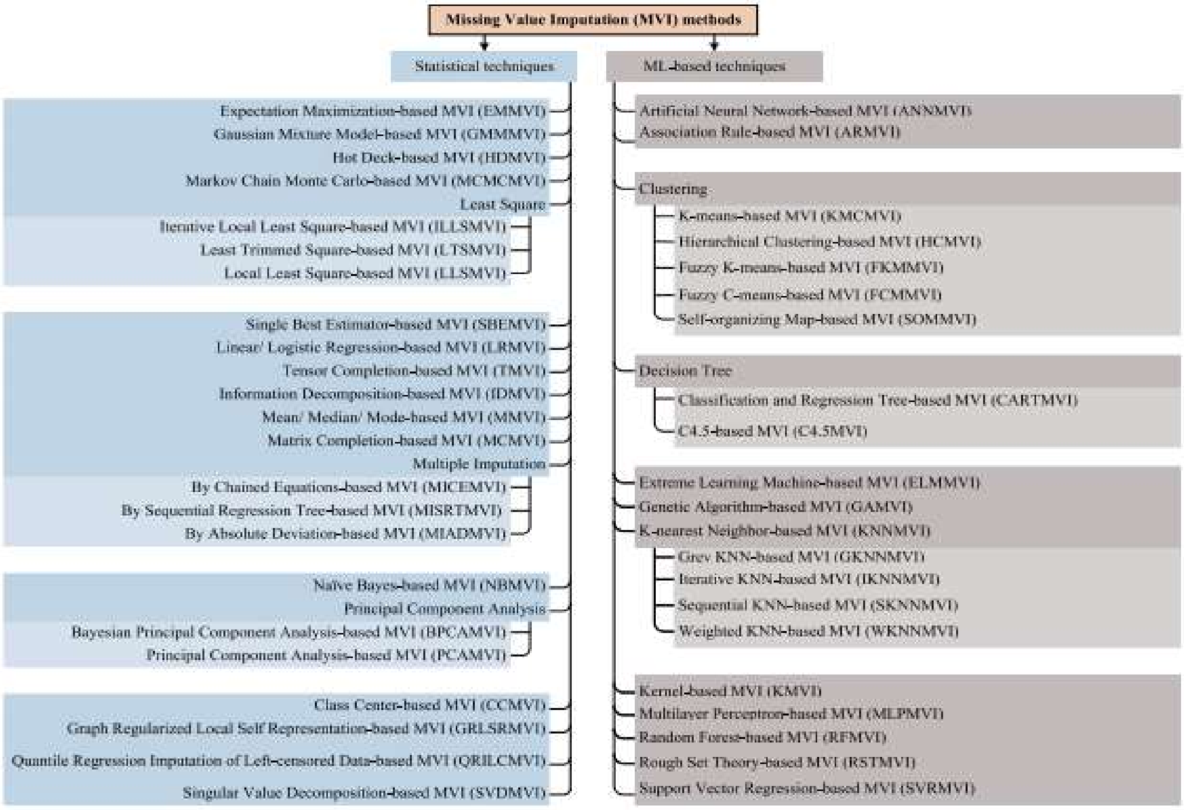


Figure1.2 The organized tree exhibition of the commonly used MVI methods.

1.2 Analysis of Missing Value Imputation and their Assessment

Figure 1.3 shows the top twelve statistical and ML-based methods for Missing Value Imputation (MVI), highly applied in literature from 2010 to August 2021. EMMVI, MMVI, LLSMVI, BPCAMVI, and LRMVI consistently emerge as the top-5 statistical MVI techniques, featured in 34, 34, 12, 11, and 11 articles, respectively. Notably, EMMVI and MMVI are the most heavily employed, with their usage approximately three times higher than the third-ranked LLSMVI. These methods are favored for their ease of implementation, memory efficiency, resilience to outlier imputation, and minimal time requirements for missing value prediction. They operate independently of prior data knowledge and maintain unbiased attribute means.

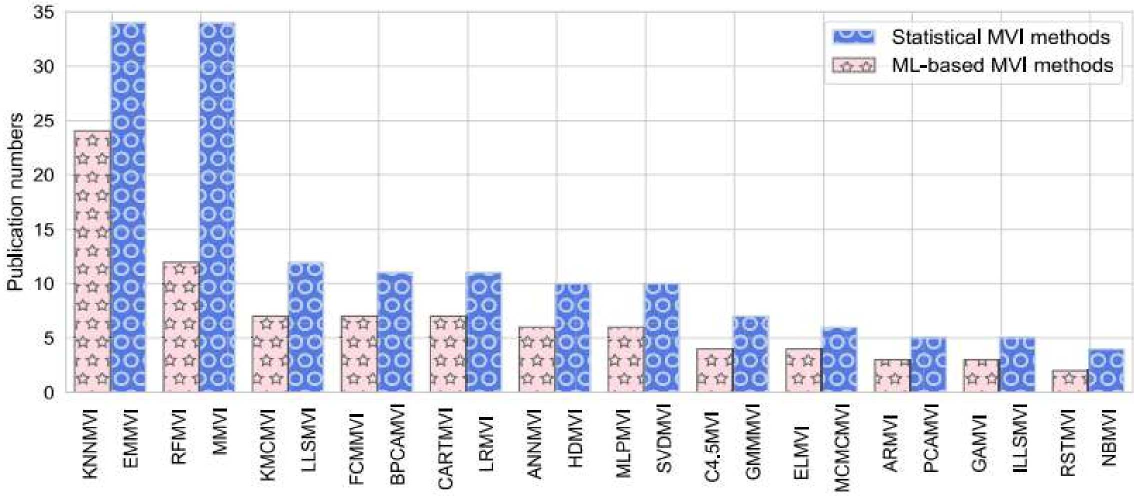


Figure 1.3 Analysis of Missing Value Imputation and their assessment

1.3 Missing data mechanisms

The term "missing data" indicates the absence of values in a dataset, and this can happen for various reasons, including errors during data collection, incomplete surveys, participant dropouts, or technical issues. Understanding and handling missing data is crucial for correct and meaningful data examination. Several mechanisms describe how missing data can occur:

a) **Missing Completely at Random (MCAR):** This happens when the missingness is unconnected to observed and unobserved data. In other words, the probability of missing data is the same for all observations, regardless of different variables. It implies that the missingness is random and not influenced by any underlying factors. MCAR is ideal but rarely occurs in practice.

b) **Missing at Random (MAR):** In this scenario, the lack of data is associated with the observed data rather than the specific missing values. The probability of missing data is contingent upon observed variables, making it potentially predictable. While MAR might introduce bias, it can often be adjusted if the variables causing the missingness are included in the analysis.

c) **Missing Not at Random (MNAR)** is the most problematic scenario. Missing data is not random and is related to unobserved data or reasons not included in the dataset. If not appropriately handled, this can introduce significant bias into analyses. Addressing MNAR requires careful consideration and, in some cases, specialized techniques.

1.4 Why do we need to care about handling hissing data?

Handling missing data is a crucial aspect of data analysis and modeling for several reasons:

a) **Preserving Data Integrity:** Missing data can introduce bias and distort the authentic relationships within the data. Ignoring missing data can lead to inaccurate results and faulty conclusions. Proper handling of disappeared data helps maintain the integrity of the dataset and ensures that analyses and models are based on accurate information.
b) **Accurate Statistical Analysis:** Many statistical analyses require complete data for accurate and meaningful results. Missing data can lead to skewed distributions, incorrect estimates of variability, and biased parameter estimates. Addressing missing data ensures that the statistical analyses are valid and reliable.
c) **Avoiding Biased Results:** If the missing data are not appropriately handled, the observed patterns in the remaining data can be misleading. Certain groups or variables might be disproportionately affected by missing data, leading to biased results that do not accurately represent the population or phenomenon being studied.
d) **Effective Modeling:** Models built using incomplete data can be less accurate and less robust. Whether you're creating predictive models, machine learning algorithms, or simulations, the quality of the model's

predictions and generalizations depends on the quality of the input data. Proper handling of missing data helps improve the performance of models.

e) **Ethical Considerations:** In some cases, missing data might not be MAR, which means that the reason for the missingness could be related to the underlying characteristics being studied. Failing to account for this can result in unfair and discriminatory outcomes. Proper handling of missing data helps mitigate ethical concerns and ensures fairness in analyses.

f) **Complete Information:** Missing data can lead to losing valuable information that might be important for decision-making. Properly handling missing data allows you to make more informed decisions based on a more complete understanding of the data.

g) **Regulatory and Compliance Requirements:** In certain domains, such as healthcare, finance, and research, there are strict regulatory and compliance requirements regarding data quality and integrity. Properly handling missing data is necessary to meet these standards.

Numerous techniques for addressing missing data include imputation approaches (where missing values are replaced with estimated values), deletion methods (which involve removing instances with missing values), and advanced approaches like employing machine learning algorithms to predict missing values. The selection of a specific method hinges on factors such as the characteristics of the data, the root causes of missing values, and the objectives of the analysis or modeling process.

1.5 Process of Missing Data Imputation

Imputing missing data in a dataset involves filling in or estimating the missing values to ensure the dataset is complete and suitable for analysis. Here's a general step-by-step guide on how to impute missing data in a dataset:

a) **Identify Missing Values:** Begin by identifying which variables in your dataset contain missing values. This can be done by examining summary statistics or using functions in programming languages like Python (e.g., is null() in pandas).
b) **Understand the Nature of Missingness:** Determine if the missing data is missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Understanding the nature of missingness can help guide the imputation approach.
c) **Select Imputation Method:** Choose an appropriate imputation method based on the nature of the data, the extent of missingness, and assumptions about the missing data mechanism. Common methods include mean/median imputation, mode imputation, regression imputation, KNN imputation, multiple imputation, and deep learning imputation.
d) **Preprocess Data:** Before imputing missing values, preprocess the data as needed. This may involve scaling numerical variables, encoding categorical variables, or performing other transformations.
e) **Impute Missing Values:** Apply the chosen imputation method to fill in the missing values in the dataset. This can be done using built-in functions in data analysis libraries or custom code.
f) **Validate Imputed Data:** After imputation, it's crucial to validate the imputed data to ensure that the imputation process has not introduced bias or affected the distribution of the variables. This can involve visualizing the

imputed data, comparing summary statistics before and after imputation, or conducting sensitivity analyses.

g) **Perform Analysis:** Once the missing values have been imputed and validated, you can proceed with your data analysis as usual.

h) **Document Imputation Process:** It's essential to document the imputation process, including the methods used, any assumptions made, and any decisions taken during the imputation process. This documentation helps ensure the analysis's transparency and reproducibility.

i) **Consider Sensitivity Analyses:** In some cases, it may be appropriate to perform sensitivity analyses to assess the robustness of the results to different imputation methods or assumptions about the missing data mechanism.

j) **Report Results:** When reporting the results of your analysis, clearly state the imputation methods used and any potential limitations associated with the imputed data.

By following these steps, one can effectively impute missing data in a dataset and ensure that the analysis is based on complete and reliable data.

1.6 How to Handle Missing Values

Handling missing values involves various techniques that aim to address the gaps in the data caused by missing observations. The selection of a technique depends on factors like the data's characteristics, the causes of missing values, and the objectives of the analysis.

Deep learning is one of the critical machine learning active research fields; it is also a subset of machine learning and has achieved great success in the spectrum of scientific and technological domains, including image classification, speech recognition, language processing, missing data imputation, big data analytics, and many more. Deep learning techniques are at the forefront of artificial intelligence research and application, revolutionizing various fields such as computer vision, natural language processing, and robotics. At its core, deep learning is a subset of machine learning that involves training artificial neural networks with vast amounts of data to learn and make predictions or decisions.

Here's a brief introduction to some vital deep-learning techniques:

a) **Artificial Neural Networks (ANNs):** ANNs are the building blocks of deep learning. They are inspired by the structure and function of the human brain, consisting of interconnected nodes (neurons) organized in layers. Each neuron receives input signals, processes them, and passes the output to the next layer.
b) **Convolutional Neural Networks (CNNs):** CNNs are specialized neural networks designed for processing grid-like data, such as images. They use convolutional layers to apply filters to input data, capturing spatial patterns and hierarchies of features. CNNs are widely used in tasks like image classification, object detection, and image segmentation.
c) **Recurrent Neural Networks (RNNs):** RNNs are designed to handle sequential data by maintaining a memory of previous inputs. Their connections form directed cycles, allowing information to persist over time. RNNs are effective in language modeling, speech recognition, and time series prediction.
d) **Long Short-Term Memory (LSTM) Networks:** LSTMs are a type of RNN architecture designed to address the vanishing gradient problem, which hinders the training of deep networks on long sequences. LSTMs use a

gating mechanism to regulate the flow of information, enabling them to learn long-term dependencies in sequential data.

e) **Generative Adversarial Networks (GANs):** GANs consist of two neural networks, a generator and a discriminator, which are trained simultaneously through a competitive process. The generator learns to generate synthetic data samples that are indistinguishable from accurate data, while the discriminator learns to differentiate between real and fake samples. GANs have applications in image generation, data augmentation, and style transfer.

f) **Autoencoders:** Autoencoders are neural networks trained to reconstruct input data, typically used for unsupervised learning and dimensionality reduction. They consist of an encoder, which compresses the input into a latent representation, and a decoder, which reconstructs the input from the latent representation. Variants like denoising autoencoders and variational autoencoders (VAEs) have been developed for various applications.

These are just a few examples of deep learning techniques, and the field is continuously evolving with new architectures, algorithms, and applications. Deep learning has shown remarkable success in various domains, driving advancements in technology and reshaping industries across the globe.

1.7 Problem Statement

Missing data is a prevalent issue in various real-world datasets across industries, and its presence can significantly impact the quality and accuracy of data mining tasks. The problem of missing values imputation involves devising effective strategies and algorithms to replace missing data with estimated or predicted values, thereby enhancing the overall integrity of the dataset and facilitating more reliable data analysis.

Detecting issues with missing data can be challenging, as it is often unpredictable. Data professionals may find it difficult to determine when missing data will impact results, and it is only sometimes clear when it will pose a problem. While each variable or question might have only a few missing responses individually, the cumulative effect of missing values can be significant. Assessing the impact of missing data has traditionally been time-consuming and error-prone, requiring systematic analysis.

Machine learning (ML) and data mining algorithms are widely employed to predict results from extensive datasets. While these procedures often generate accurate predictions, their effectiveness hinges on the quality of the dataset used for training. An integral step in the data analysis and mining process involves refining the data that will serve as the training foundation for the algorithms. This data mining process is known as data preprocessing, which is known as the most challenging part for data analysts. In many cases, data must either be included or correctly entered by humans, resulting in incorrect predictions. One of the main problems regarding data quality is that values need to be included. Missing values in the dataset may significantly increase computational cost, skew the outcome, and frustrate researchers.

In data analytics, missing data poses a challenge that can impair performance. Erroneous imputation of missing values has the potential to result in inaccurate predictions. In the current era of big data, where a colossal amount of data is generated every second and stakeholders emphasize optimizing the utilization of this data, effective handling of missing values becomes increasingly critical. This research introduces a novel technique for missing data imputation, presenting a hybrid approach that combines multiple imputation techniques. Additionally, we propose extensions for imputing categorical and numeric data, encompassing two variations.

Missing data is an issue that lowers performance in data analytics. An erroneous prediction could result from an incorrect imputation of missing values. Effectively addressing missing values becomes more crucial in the significant data era, where enormous amounts of data are produced each second. Exploiting these data is a substantial problem for the stakeholders. We have developed a novel technique for the imputation of missing data in this study that combines many imputation strategies. We have presented an extension of two versions for imputed categorical and numerical data.

We have associated the presentation of our suggested algorithm with the existing methods and found that our proposed algorithm produces higher accuracy than the existing algorithm. I am putting my effort here, hoping it will be helpful to any data practitioner or enthusiast—aims and Objectives of the Research.

1.8 Challenges:

The challenges associated with missing values imputation in data mining are multifaceted and require comprehensive solutions:

a) **Accuracy Preservation:** The imputed values should be as close to the actual values as possible to ensure the accuracy of downstream data analysis, modeling, and decision-making.
b) **Data Distribution:** Imputed values should reflect the data's central tendency and preserve its underlying distribution characteristics.
c) **Feature Interactions:** Some features may have complex relationships with one another, and imputing missing values should consider these interactions to avoid introducing unrealistic patterns.

d) **Dimensionality:** High-dimensional datasets pose challenges in terms of selecting appropriate imputation techniques that can effectively handle various data types and relationships.

e) **Bias and Outliers:** Imputed values should not introduce bias or amplify the presence of outliers in the dataset.

f) **Temporal and Contextual Information:** For time-series or context-dependent data, missing values imputation should consider temporal and contextual factors to ensure accurate representation.

g) **Scalability:** The proposed imputation methods should be scalable to handle large datasets efficiently without sacrificing imputation quality.

h) **Incorporating Domain Knowledge:** Imputation techniques should allow for integrating domain-specific knowledge or constraints to ensure that imputed values align with expert insights.

1.9 Objectives of the Research

The research objectives are to develop advanced and effective missing values imputation techniques for enhancing the quality and reliability of data mining processes. There by permitting more accurate analysis and decision-making in various domains. The research aims to achieve the following objectives:

The primary objectives of addressing the missing values imputation problem are:

- Identification of significant attributes to deal with missing values handling.
- To Analyze existing datasets for missing values using various ML approaches
- To develop a time series-based model for handling missing values.
    - ➢ Utilization of optimization technique to address the challenge of MVI

1.10 Contributions

In this thesis, the researcher's contribution to the current research work is absorbed in two aspects, as deliberated below.

This research employs multiple imputation techniques, generating numerous values for imputing a single missing value through various simulation models. These techniques introduce various imputed data types to capture a diverse range of acceptable responses. Despite their complexity, multiple imputation techniques offer an advantage over single imputation by avoiding bias values.

The multiple imputation process replaces each missing data point with n values derived from n iterations. Researchers opt for a multiple imputation approach to impute missing values, assuming that the data are missing at random (MAR). This algorithm predicts the likelihood of missing values based on observed data, providing multiple values for a single missing value through a series of regression models, each dependent on its technique parameter. In this approach, each missing variable serves as a dependent variable, with other data in the record acting as independent variables.

The proposed algorithm predicts missing data by leveraging existing data from other variables, subsequently replacing missing values with the expected values to create an imputed dataset. The iterative technique generates multiply imputed datasets; each analyzed using standard statistical methods, yielding multiple analysis results. The research introduces a technique that seamlessly imputes missing values in a dataset by examining values from other columns and estimating the best forecast for each missing value.

In this research researcher has developed a new approach with integration of mice and ANN algorithm. To understand dietary patterns and addressing challenges regarding food security, the prediction of food consumption is essential. The integration of missing value imputation methods with an Extended Artificial

Neural Network (EANN) for future predictions is a novel approach. Such a study could provide valuable insights into food consumption trends while addressing the issue of missing data, potentially leading to more accurate and reliable predictions in this vital field.

Our proposed algorithm employs linear calculations to approximate a nonlinear function. The result of this approximation is an Extended Kalman Filter (EKF). Specifically, we select a point and execute a cluster of derivatives on it. In the context of an EKF, we compute the mean of the Gaussian distribution on the nonlinear curve and conduct multiple derivatives to estimate it.

The Extended Kalman Filter (EKF) is based on the premise that a local linear approximation of the system adequately captures nonlinearities. Consequently, the linearized model is utilized instead of the original nonlinear function. These calculations are notably straightforward, contributing to the filter's widespread use. Nevertheless, when confronted with highly nonlinear systems, the EKF estimates encounter significant challenges, including unstable and rapidly divergent behaviors, suboptimal linearization, and erratic responses.

The Kalman filter is used to determine optimal approximations and is anticipated to conform to a normal distribution. Its critical function is to calculate the conditional mean and variance of the distribution for observed conditions up to a given time. This research aims to enhance the Kalman Filter (KF) by introducing an adaptive structure that seeks neighboring derivative outcomes and multiplies them by the rate of change in the extended Kalman filter.

1.11 Thesis Organization

That sounds like a comprehensive approach, but addressing missing value imputation and prediction through machine learning and deep learning techniques can enhance the robustness of data analysis. This thesis consists of the following chapters:

Chapter-1 Introductory Information

This chapter explores the study's background regarding missing values, including imputation techniques and data mechanisms, which sets a solid foundation. Defining the problem statement and outlining the contributions of your study helps focus the research objectives.

Chapter-2 Related Literature

This chapter presents the literature surrounding missing values imputation, which is rich and diverse, covering various techniques and approaches. Here are some key areas and methods often explored in related literature. When delving into the related literature, it's beneficial to examine recent publications, comparative studies, and research papers focusing on the study's specific context.

Chapter-3 Missing Values Imputation Predictive Modelling

This chapter highlights the challenge missing values pose in predictive modeling and emphasizes the pivotal role of imputation techniques in setting the stage well. Emphasizing the importance of selecting an imputation method that optimizes predictive power while minimizing information loss aligns with best practices in this field. Exploring the impact of imputation on model performance and comparing various strategies could provide valuable insights.

Chapter-4 Futuristic Prediction of Food Consumption with Missing Value Imputation Methods Using Extended ANN

This chapter argues that predicting food consumption is crucial to understanding dietary patterns and addressing food security challenges. Integrating missing value imputation methods with an Extended Artificial Neural Network (EANN) for futuristic prediction sounds innovative. Such a study could offer valuable insights into predicting food consumption trends while addressing the challenges of missing data, potentially contributing to more accurate and reliable futuristic predictions in this critical domain.

Chapter-5 MVI and Forecast Precision Upgrade of Time Series Precipitation

This chapter presents how the study can significantly contribute to advancing precipitation forecasting techniques by addressing missing values and improving the reliability and accuracy of predictions, which is crucial for informed decision-making in various sectors dependent on weather forecasts.

Forecasting precipitation is vital for various sectors, especially agriculture, water resource management, and disaster preparedness. Combining missing values imputation with an Extended Kalman Filter (EKF) to enhance the precision of time series precipitation forecasting is fascinating.

Chapter-6 Conclusion and Future work

This chapter provides the conclusion and future directions to the research scholars for carrying out the work in future.

Chapter-2 Literature Review

2.1 Introduction

Data mining is extracting valuable patterns and knowledge from large datasets. It encompasses a range of techniques, including classification, clustering, regression, association rule mining, and anomaly detection [1]. For instance, in the field of marketing, data mining can be used to identify customer segments and tailor marketing strategies accordingly. In finance, it can help predict stock market trends. In healthcare, it can aid in disease diagnosis and treatment planning. And in cyber security, it can detect and prevent potential threats. These techniques, powered by machine learning algorithms and statistical methods, enable organizations to uncover insights, make informed decisions, and gain competitive advantages in various fields [2].

Data mining techniques are invaluable in missing value imputation, aiding in predicting and estimating absent data points. Through pattern recognition, regression analysis, clustering, and neural networks, these methods enhance data completeness, ensuring the reliability of datasets for subsequent analysis and decision-making

2.2 **Data mining**

Data mining is a transformative process that turns raw data into meaningful and actionable insights. It involves exploring patterns, relationships, and valuable insights within large and complex datasets. Employing various techniques, algorithms, and methodologies, it aims to unearth meaningful information and knowledge from data sets that might otherwise remain obscure or challenging to uncover [3]. The primary goal of data mining is to transform raw data into meaningful and actionable insights that can inform decision-making for predicting future outcomes and optimizing various domains. By exploring and analyzing data from multiple perspectives, often using advanced computational and statistical methods, data mining opens up a world of possibilities and potential for organizations [4].

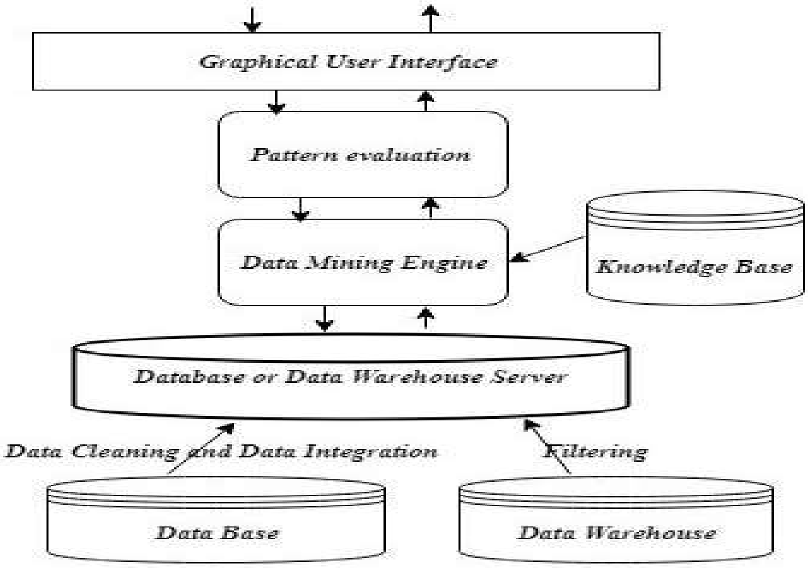


Figure 2.1: Architecture of a data mining system.

The data mining methods can be applied to various kinds of data, including databases, text documents, data warehouses, social media data, multimedia files, etc [5]. Here are some standard data mining techniques are included:

Data mining is usually divided into two parts:

(i) Descriptive data mining

(ii) Predictive data mining.

2.2.1**Descriptive data mining**

Descriptive data mining, also known as descriptive analytics, focuses on presenting and summarizing existing data to provide insights and understanding about the patterns and characteristics within the data[6]. This form of data mining does not involve making predictions or extrapolations about future consequences; instead, it proposes to describe and visualize the data in a meaningful technique. On the other hand, predictive data mining, also called predictive analytics, involves using historical datasets and statistical procedures to predict forthcoming events or results. It is widely used in numerous fields to help organizations make informed decisions, optimize procedures, and anticipate upcoming trends [7].

There are some critical aspects of descriptive data mining:

a) **Data Visualization:** Data Visualizations, such as diagrams, graphs, and plots, are commonly used in descriptive data mining to present data patterns in a visually appealing and informative way. Examples include histograms, bar charts, pie charts, scatter plots, and line graphs [8].

b) **Data Summarization:** Descriptive data mining involves summarizing large and complex datasets into more manageable and understandable forms. This can include calculating basic statistics like mean, median, mode, standard deviation, and range for numerical variables. Categorical variables involve calculating frequencies and proportions [9].

c) **Data Exploration:** Exploring the data involves interacting with it to discover interesting patterns or relationships that might not be immediately obvious. This could involve interactive visualizations, filtering, and drilling down into subsets of the data [10].

d) **Data Profiling:** Data profiling involves examining the structure and content of the data to understand its quality, completeness, and integrity. This can help identify missing values, outliers, and inconsistencies [11].

e) **Pattern Recognition:** In descriptive data mining, pattern recognition identifies recurring trends, anomalies, and patterns within the data. This can be particularly useful for understanding customer behavior, market trends, or other regularities in the data [12].

f) Segmentation: Descriptive data mining often involves segmenting the data into meaningful groups based on specific characteristics. This segmentation can help businesses tailor their strategies to different customer segments [13].

g) Data Presentation: The results of descriptive data mining are typically presented in reports, dashboards, or presentations. These presentations can help stakeholders understand the current state of the data and make informed decisions based on the insights.

h) Data Cleaning: While not the primary focus, data cleaning is often a part of descriptive data mining. It's crucial to ensure that the data used for analysis is accurate, complete, and reliable [14].

Descriptive data mining is the foundation upon which further analyses, such as predictive modeling or prescriptive analytics, can be built. By thoroughly understanding the data's characteristics, patterns, and distributions, analysts can make informed choices about how to proceed with more advanced analyses to address specific business questions or objectives.

2.2.2 **Predictive data mining**

Predictive data mining, also called predictive analytics, involves using historical datasets and statistical procedures to predict forthcoming events or results. It is widely used in numerous fields to help organizations make informed decisions, optimize procedures, and anticipate upcoming trends.

This kind of data mining drives beyond descriptive analysis, which focuses on understanding and summarizing existing data and purposes of predicting what might occur based on patterns and relations identified in the past data. Here are some key features of predictive data mining [15]:

a) **Classification and Regression:** Predictive data mining involves classification tasks (categorizing data into predefined classes) and regression tasks (predicting continuous numeric values) [16-17].

Predictive data mining can provide valuable insights into customer behavior, market trends, risk assessment, and more. It empowers organizations to anticipate potential outcomes and make proactive decisions, leading to improved efficiency, better resource allocation, and enhanced strategic planning.

2.3 History of data mining

The history of data mining can be traced back to the 1960s and 1970s when early attempts were made to extract knowledge from large datasets.[18,19] Here is a brief overview of the key milestones and developments in the history of data mining:

- Early Origins (1960s-1970s): The foundations of data mining were laid during this period with the emergence of techniques such as clustering, regression analysis, and exploratory data analysis. Researchers and statisticians began exploring ways to extract useful information from large datasets [20].
- Birth of Artificial Intelligence (1980s): In the 1980s, the field of Artificial Intelligence (AI) experienced significant advancements, directly impacting data mining. Researchers started developing algorithms and techniques to discover patterns and relationships within data automatically [21, 22].
- Knowledge Discovery in Databases (KDD) (1990s): The term "Knowledge Discovery in Databases" (KDD) was coined to describe the process of

extracting knowledge from data. KDD encompassed various stages, including data cleaning, integration, selection, transformation, mining, pattern evaluation, and knowledge presentation. It became a popular framework for data mining research [23, 24].

- Machine Learning and Data Mining Convergence (1990s): Machine learning and data mining began to converge in the 1990s. Machine learning algorithms, such as decision trees, neural networks, and support vector machines, were adapted and applied to data mining problems [25].
- Rapid Growth and Commercialization (2000s): In the early 2000s, data mining gained significant momentum as technologies advanced, computational power increased, and data storage became more affordable. Data mining tools and platforms became more accessible, increasing adoption across industries [26].
- Big Data and Advanced Techniques (2010s): With the exponential growth of data generated from various sources, including social media, sensors, and the Internet of Things (IoT), the focus shifted to handling and mining big data. Advanced techniques, such as deep learning, ensemble methods, and natural language processing, were developed to extract insights from vast and complex datasets[27,28].
- Current Trends: In recent years, data mining has become an integral part of numerous domains, including finance, healthcare, marketing, cybersecurity, and more. Techniques like data visualization, text, and graph mining have gained prominence. Additionally, ethical considerations and privacy concerns associated with data mining have received increased attention [29].

Data mining continues to evolve as technology advances and new challenges emerge. It plays a crucial role in leveraging today's vast data, enabling organizations to gain valuable insights and make data-driven decisions.

2.4 Techniques of data mining

Data mining encompasses various techniques and algorithms to extract patterns, relationships, and insights from large datasets. Figure 2.2 explores the diverse array of data mining techniques applicable across various fields. Here are some commonly used data mining techniques [30]:

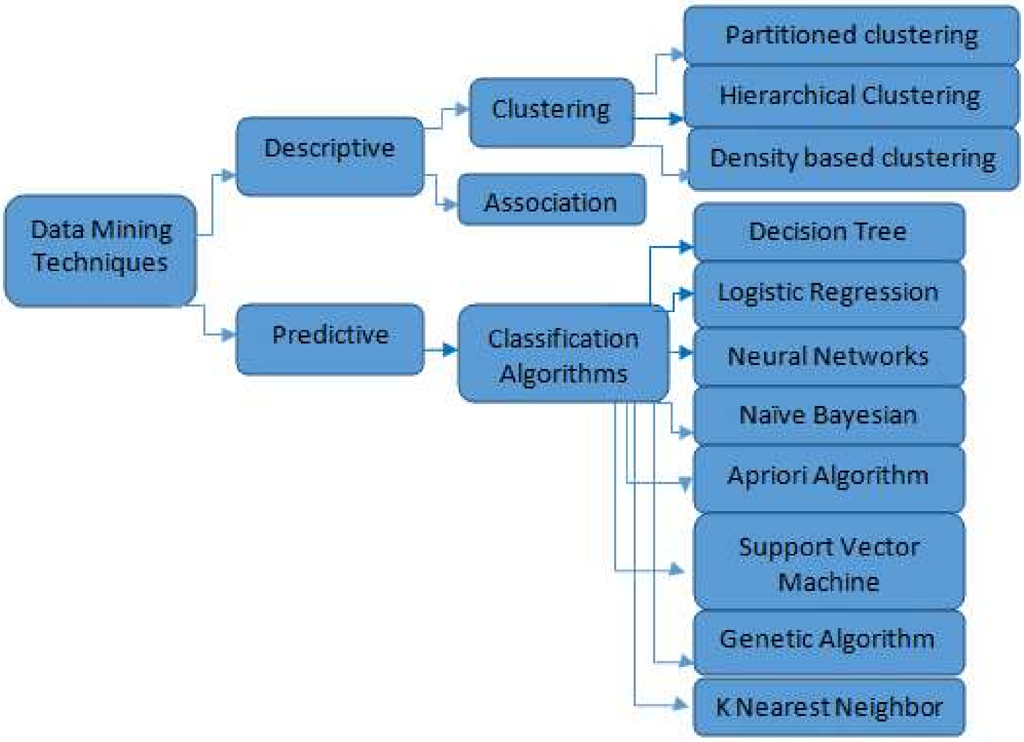


Figure 2.2 Data Mining Technique

a) Association Rule Mining: This technique is used to identify relationships and associations between items in a dataset. It helps discover co-occurrence patterns, such as "people who buy X also tend to buy Y." The widely used algorithm for association rule mining is the Apriori algorithm.

b) Classification: Classification involves building models that can classify data into predefined classes or groups based on specific attributes or features. Algorithms like Decision Trees, Naive Bayes, Random Forests, and Support Vector Machines (SVM) are commonly used for classification tasks [31, 32].

c) Clustering: Clustering groups similar data points based on their characteristics or attributes. It helps identify natural clusters or patterns within a dataset. Popular clustering algorithms include K-means, Hierarchical Clustering, and DBSCAN [33, 34].

d) Regression Analysis: Regression analysis identifies and models the relationships between variables to predict numerical values. It helps in understanding how one variable affects another. Linear regression, logistic regression, and polynomial regression are standard regression techniques [35, 36].

e) Anomaly Detection: Anomaly detection focuses on identifying unusual patterns or outliers in the data that deviate significantly from the norm. It helps detect anomalies or abnormalities that may indicate fraudulent activities, system failures, or anomalies in sensor readings. Techniques like statistical methods, clustering-based approaches, and outlier detection algorithms are employed for anomaly detection [37, 38].

f) Text Mining: Text mining involves extracting helpful information and insights from unstructured text data, such as documents, emails, social media posts, and customer reviews. It includes techniques like text classification, sentiment analysis, named entity recognition, and topic modelling [39, 40].

g) Sequential Pattern Mining: This technique is used to discover sequential patterns or frequent sequences of events in data that occur over time. It finds patterns in temporal or sequential data, such as customer browsing behavior, market basket analysis, or DNA sequences. The GSP (Generalized Sequential Pattern) algorithm is widely used for sequential pattern mining [41].

h) Decision Trees: Decision trees are graphical models that represent decisions and their possible consequences. They are widely used for classification and regression tasks. Decision trees recursively split the dataset based on different attributes to create a tree-like structure that can be used for decision-making [42, 43].

These are just a few examples of data mining techniques. Depending on the specific requirements and nature of the data, many more advanced and specialized algorithms and methods are available.

2.5 Data Mining Process

Data mining architecture refers to the overall structure or framework of systems and processes used to perform data mining tasks effectively. It encompasses various components, including Data integration, Data Collection, data pre-treatment, data translation, data mining method collection, data mining algorithm collection, data mining information expression and analysis, and decision [44, 45]. Figure 2.4 is an overview of the typical data mining process.

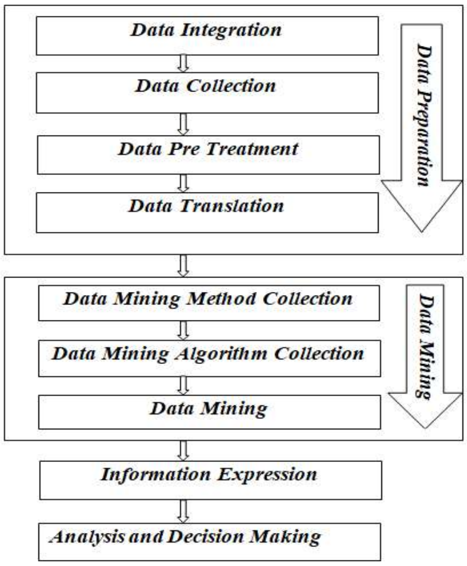


Figure 2.3: Data Mining Process Diagram

2.6 Missing Values Imputation is an essential tool for data quality

Missing values imputation is a fundamental technique for enhancing data quality and is crucial for accurate and reliable data analysis [46-48]. Here are some reasons why missing values imputation is an essential tool for maintaining data quality:

a) Prevents Data Loss: Incomplete datasets with missing values may exclude valuable information, which could potentially impact the accuracy and comprehensiveness of analyses [49].

b) Maintains Data Integrity: Imputing missing values helps maintain the dataset's overall integrity by ensuring that it is complete and representative of the real-world context it aims to model[51,52].

c) Supports Reliable Analysis: Data analysis and modeling techniques often require complete datasets. Imputation ensures that these techniques can be applied without bias or underrepresentation due to missing data [52].

d) Reduces Bias: If missing values are not handled properly, they can introduce bias into analyses, leading to incorrect conclusions and predictions. Imputation aims to minimize this bias [53].

e) Enhances Predictive Modeling: Missing values in predictor variables can adversely affect the accuracy of predictive models. Imputation helps build more reliable models by preserving relationships between variables [54, 55].

f) Preserves Data Relationships: Many datasets contain interrelated features. Imputing missing values while considering these relationships helps maintain the consistency and accuracy of data patterns [56, 57].

g) Improves Statistical Power: Imputation increases the sample size by filling in missing values, leading to better statistical power and more reliable results [58, 59].

h) Facilitates Comparative Analysis: When multiple datasets are compared or merged, imputation ensures compatibility and consistency by making them more comparable [60].

i) Supports Longitudinal Studies: In research involving time-series or longitudinal data, missing values imputation helps maintain the continuity of the data over time [61-63].

j) Enables Cross-Domain Utilization: High-quality imputed datasets can be more easily shared and utilized across different research or application domains [64].

k) Enhances Data Mining and Machine Learning: Complete datasets are essential for accurate feature selection, pattern recognition, and building robust machine learning models [65-67].

l) Contributes to Decision-Making: Accurate and complete data are vital for informed decision-making in various domains, such as healthcare, finance, and marketing [68,69].

However, it is essential to note that imputation methods should be chosen carefully, as poorly executed imputation can introduce unintended biases or distort the underlying data distribution. The choice of imputation technique should align with the nature of the data, the patterns of missingness, and the specific analysis goals.

2.7 Missing Values Imputation Methodologies

In this part of the research, we provide an overview of previous research in data analysis, focusing on areas such as missing data, imputation methods, and clustering algorithms [70]. The significance of data quality in both Data Mining and Business Intelligence cannot be overstated because incomplete or noisy data can significantly hamper analysis and result in flawed statistical conclusions and decision-making processes. Consequently, addressing missing values prior to analysis is paramount [71, 72]. We present a summary of prevalent techniques for handling missing values in Table 2.1

Table 2.1 Overview of Related Techniques Missing Data Imputation

Summary of Related Approaches Missing Data Imputation	
Reference Related Work	Description
[73]	Joreskog's classification of missing data types
[74]	Factors contributing to missing values in surveys
[75]	Imputation techniques for improving data analysis
[76]	Challenges posed by missing data in analysis
[77]	Impact of missing data on decision-making
[78]	Imputation as a solution for replacing missing dat
[79]	Uninformed patterns of absent data in real-world

	datasets
[80]	UCI Machine Learning Repository for benchmark datasets
[81]	Critique of discarding data vectors with missing values
[82]	R-language as an open-source tool for data analysis
[83]	Artificial Neural Networks for model-based imputation
[84]	Evaluation of k-Nearest Neighbors imputation method
[85]	Missing data imputation using predictive models
[86]	Impact of removing samples with high missing values proportion
[87]	Determining the acceptable threshold for missing values
[88]	Limitations of simple statistical imputation approaches
[90]	Systematic review of data imputation methods in data mining
[91]	Data imputation for missing values and ensuring data integrity
[92]	Regression techniques for model-based imputation
[93]	Deep Learning-based imputation approaches
[94]	Comparison of dynamic imputation techniques

Various techniques have been developed for addressing missing values in numeric datasets, though their applicability to ordinal data sets may differ [94]. Several frequently used techniques are worth noting when addressing missing values.

The Case Deletion (CD) must remove records containing missing data, thereby generating a revised dataset for subsequent analysis. Nevertheless, this approach may not be suitable for datasets with a substantial proportion of missing values. Even in cases with fewer missing entries, it is essential to assess the potential bias introduced by the modified dataset [95]

Random Value Imputation is a method for filling in missing values, resulting in a complete dataset. Although simple, this approach does not utilize figures from the dataset and may introduce randomness that impacts further analysis[96].

The **Mean Imputation (MI)** involves filling in missing values with the mean value of the respective feature or attribute from the complete dataset [97]. This method, also referred to as complete mean imputation, has limitations [98]. It may not be ideal for datasets with many missing values, as it reduces variance and can inflate the apparent sample size [99].

The **Most Common Imputation (MCI)** technique replaces missing values with the most frequently occurring value in the dataset [100,101]. This approach assumes that the most common value symbolizes a plausible estimate for the missing data.

The Median Imputation fills in missing values by using the dataset's median value of the respective feature [102]. Alternatively, class median imputation replaces missing values with the median of the feature within a particular class [103]. The class should correspond to the class variable of the vector containing the missing value.

Various strategies have been suggested for managing missing data in ordinal datasets. Decision trees have proven effective in classifying ordinal data by organizing data into splits or branches [104[. Traditional methods for handling missing values might introduce bias and diminish or amplify statistical power. Removing missing instances is often favored for simplicity and is commonly the default procedure in statistical data analysis tools. However, this approach may lead to the loss of a substantial portion of the data in practical scenarios [105,106.]

Neural networks provide a viable solution for handling missing values in ordinal data. They offer a classifier-based imputation method inspired by the functioning of the brain. Neural networks usually comprise an input, hidden, and output layer, with training algorithms enabling them to tackle intricate mathematical problems. The hidden layer adjusts dynamically during the training process [107,108]. Another classifier-based approach for addressing missing data is Support Vector Machines (SVM) [109,110].

Classification-based imputation methods have emerged as effective strategies for estimating and filling missing values within datasets [111, 112]. These approaches utilize diverse classification techniques, including neural networks, decision trees, and similar methods, to address the challenge of missing values. However, previous research has primarily concentrated on managing numerical or nominal missing data values [113]. In contrast, this study seeks to address this research gap by focusing on the treatment of missing values in ordinal data and examining the effects of these treatments on unsupervised learning techniques, particularly clustering

2.8 Present Trends in Missing Value Imputation

As of my last knowledge update in January 2024, I can provide some insights into emerging trends in the field of missing value imputation [114]. Keep in mind that these trends have evolved since then. Here are some trends that were observed:

a) **Machine Learning-Based Imputation:** Machine learning techniques, such as intense learning, were being applied to missing value imputation [115]. Neural networks and advanced algorithms were used to learn complex relationships within the data and impute missing values more accurately [116].

b) **Multiple Imputations:** Multiple imputation techniques were gaining popularity due to their ability to account for uncertainty in imputed values. They were used to generate multiple plausible imputed datasets, which were combined to provide more reliable results [117].

c) **Integration of Domain Knowledge:** Researchers focused on incorporating domain knowledge into the imputation process. This involves using domain-specific information to guide imputation decisions and make the imputed values more contextually relevant [118].

d) **Sequential Data Imputation:** With the increasing prevalence of time-series and sequential data in various fields, there was a growing interest in developing imputation techniques that consider such data's temporal dependencies and patterns [119].

e) **Nonparametric Imputation:** Traditional imputation methods often assume specific distributions or relationships. Nonparametric approaches were gaining attention for their ability to handle a wide range of data types without solid assumptions [120].

f) **Imputation for Big Data:** As more organizations deal with massive datasets, there was a focus on developing scalable and efficient imputation techniques to handle big data settings without compromising accuracy [121].

g) **Missing Data in Deep Learning:** Researchers were exploring ways to make deep learning models more robust to missing data, allowing the models to learn from partially observed data effectively [122].

h) **Evaluation Metrics:** It was becoming more important to develop new metrics to assess the quality of imputed data and the impact of imputation on downstream tasks. Researchers were looking beyond simple imputation accuracy to capture the practical utility of imputed data [123].

i) **Imputation Uncertainty:** Researchers were working on methods to quantify and communicate the uncertainty associated with imputed values. This was particularly relevant for decision-making scenarios where understanding the reliability of imputed data was crucial [124].

j) **Automated Imputation Pipelines:** As data analysis workflows became increasingly complex, a trend toward developing automated pipelines that seamlessly integrated missing value imputation with other data pre-processing and analysis steps emerged [124].

It is essential to consult more recent literature and resources to understand the current trends in missing values imputation as the field evolves with new research and advancements.

2.9 Summary of the Chapter

Data mining is the process of analyzing large data sets to discover patterns, trends, correlations, or other valuable information. It involves using various statistics, machine learning, and artificial intelligence techniques to sift through vast amounts of data and extract meaningful insights.

Data mining is commonly used in various fields, such as marketing, finance, healthcare, and scientific research. For example, companies might use data mining techniques to identify customer purchasing behaviors and preferences to improve targeted advertising campaigns. In healthcare, data mining can help identify patterns in patient data to aid in diagnosis, treatment planning, and predicting outcomes.

Some standard data mining techniques include clustering, classification, regression, association rule mining, and anomaly detection. These techniques can be applied to structured data (e.g., databases) and unstructured data (e.g., text documents, images) to uncover hidden patterns and relationships that can inform decision-making and drive innovation.

Present trends in missing value imputation encompass the evolution towards more advanced techniques tailored to specific domains. Traditional methods like mean or median imputation are supplemented by sophisticated approaches such as K-nearest neighbours and deep learning. Integration with machine learning models is growing, ensuring seamless incorporation of imputation into analysis pipelines. Uncertainty estimation is gaining importance, providing insights into the reliability of imputed values. Sequential techniques and integration of auxiliary information are improving accuracy by capturing complex dependencies and incorporating external context. Moreover, modern imputation methods are designed to handle different missingness mechanisms, enhancing their applicability in various scenarios. These trends collectively aim to address the challenges posed by complex datasets and elevate the robustness and effectiveness of missing value imputation processes.

Chapter-3 Missing Values Imputation Predictive Modelling

3.1 Introduction

Missing values imputation in predictive modeling refers to filling in or estimating missing data points in a dataset before using it to build predictive models. Missing values are typically encountered in real-world datasets, often stemming from incomplete data collection, errors, or intentional omissions.

Predictive modeling involves creating models that can make predictions or decisions based on input data. These models learn patterns and relationships from historical data to predict new, unseen data. However, most machine learning algorithms and statistical techniques require complete data to work effectively. When missing values are present in the dataset, they can lead to biased, incomplete, or inaccurate predictions if not handled properly.

Missing values can be handled using various techniques, including imputation. Imputing missing values entails replacing those missing data points with estimated or predicted values derived from the available information within the dataset. The goal is to make the dataset more complete and representative of the real-world scenario so that the predictive model can learn meaningful patterns and relationships.

3.2 Approach of Missing Data Mechanism

Understanding these mechanisms helps in selecting the appropriate methods to handle missing data and ensure valid statistical inferences. In statistics, dealing with missing data is crucial as it can significantly impact the results of an analysis. The way missing data is handled often depends on the mechanism causing the data to be missing.

The most straightforward approach to infer a missing data mechanism from the data is by understanding the data collection process and leveraging substantive scientific knowledge, which is crucial for assessing randomness in missing data. Statistical testing is the method to ascertain the type of missing data mechanism. This approach is primarily employed when determining whether the mechanism is Structurally Missing Data, MCAR, MAR, and MNAR.

3.2.1 Structurally Missing Data

This type of missing data arises when certain variables only apply to all observations. For instance, if a survey asks about marital status but only to respondents above a certain age, marital status would be structurally missing for younger respondents.

Information that is missing because it makes sense logically should not be there and is referred to as structurally missing. The first and third entries in Table 3.1 do not have any data for the "Age of the youngest child" characteristic. These people do not have any kids, so their absence makes sense. Similarly, more examples of structurally missing information are in the "Number of soft drinks consumed in the past 12 hours" column. Under such circumstances, assuming that the correct answer is 0 seems sensible. For our analysis, we should thus replace these missing values with 0.

TABLE 3.1: MISSING DATA OF A STRUCTURAL NATURE

SL_No.	Kids	Age of the youngest child	Have you consumed a soft drink in the past 12 hours?	How many soft drinks have you consumed in the last 12 hours?
1	X		X	
2	√	18	√	4
3	X		X	
4	√	13	X	
5	√	8	√	3

3.2.2 Missing Completely at Random (MCAR): We must think about the tenable income of the fourth observation, as shown in Table 3.2 below. A simple way to start is to see that half of the other people have high earnings and the other half have poor incomes. It is reasonable to infer that she has a 50/50 chance of having a high or low income. Treating the missing value as missing entirely at random (MCAR) is the term used to describe this presumption. Little's MCAR test is a useful technique for verifying this assumption.

The MCAR assumption is only sometimes reliable. It is only likely to hold when the missing data occurs due to genuinely random phenomena (e.g. if survey respondents were randomly asked 10 out of 15 questions). In such cases, there is no discernible pattern in the missing data across various factors. This represents the best-case scenario in terms of our confidence in the assumption.

Table 3.2 Data are missing completely at random (MCAR).

SL. No.	Gender	Age	Earnings
I	ME	Less then30	L
II	FE	Less than 30	L
III	FE	Greater than or equal to 30	H
IV	FE	Greater than or equal to 30	
V	FE	Greater than or equal to 30	H

*ME signifies Male, FE indicates Female, L denotes Low, and H signifies High.

3.2.3 Missing at Random (MAR)

The assumption in the case of missing completely at random was that there was no discernible pattern. An alternative assumption, somewhat confusingly termed missing at random (MAR), instead posits that we can predict the missing value based on other data.

We apply this assumption to revisit the issue of estimating the value of income for the fourth observation. A straightforward predictive model assumes that income can be predicted based on gender and age. Looking at Table 3.3 below, which mirrors the one above, we observe that our missing data pertains to a Female aged 30 or older and other females in the same age group with high income. Therefore, we predict that the missing data should be categorized as High. It is important to note that the probability of prediction does not guarantee perfect accuracy in predicting a relationship. All that is required is a probabilistic relationship (i.e., that we have a better than random chance of predicting the actual value of the missing data)

Table-3.3 missing at random (MAR)

ID	Gender	Age	Earnings
I	ME	Less then30	L
II	FE	Less than 30	L
III	FE	Greater than or equal to 30	H
IV	FE	Greater than or equal to 30	
V	FE	Greater than or equal to 30	H

ME signifies Male, FE indicates Female, L denotes Low, and H signifies High.

3.2.4 Missing not at random (non-ignorable)

We may be unable to confidently make any assumptions about the potential value of missing data. For instance, individuals with low and exceptionally high incomes tend not to respond, or there could be some other unknown reason. This scenario is termed missing not at random data or non-ignorable missing data.

Consider the following study on homelessness. Data was collected from 31 women, of whom 14 were located six months later. Among them, three were found to have exited homelessness, resulting in an estimated proportion of 3/14 = 21%. Since there is no data available for the remaining 17 women who could not be reached, it is possible that none, some, or all of these 17 individuals may have also exited homelessness. This implies that the proportion of women who have exited homelessness in the sample could range from 3/31 = 10% to 20/31 = 65%. Therefore, reporting 21% as the correct outcome would be misleading.

A pattern in the missing data impacts the vital demographic factors. For instance, lower-income individuals are less likely to respond, which could skew the findings regarding income and likelihood to recommend. In this scenario, missing not at random represents our worst-case scenario.

3.3 Exploring Data Missingness

In this analysis, I employed the datasets from Zillow's Home Value Prediction Competition on Kaggle, offering a $1.2 million prize. Utilizing the Python package named missing, a versatile missing data visualization tool integrated with matplotlib, I visualized missing data patterns. This tool seamlessly handles any pandas data frame provided. The Kaggle/Zillow dataset comprises a training set and a properties dataset detailing the properties of all homes. By merging these datasets, I created a missing value matrix plot using Python Notebook 3, following the code outlined in Figure 3.1.

```
import numpy as np
import pandas as pd
import matplotlib
import missingno as msno
%matplotlib inlinetrain_df=pd.read_csv('train_2016_v2.csv',
parse_dates=["transactiondate"])
properties_df=pd.read_csv('properties_2016.csv')
merged_df=pd.merge(train_df.properties_df)
missingdata_df=
merged_df.columns[merged_df.isnull().any()].tolist()msno.matrix(merged_df[missingdata_df])
```

Figure 3.1 The Python Notebook 3 code merged both datasets and generated a plot illustrating the missing value matrix.

The nullity matrix provides a small-scale representation that helps quickly find patterns in the dataset that contain missing data. Furthermore, the sparkling situated on the right offers a summary of the completeness of the data and indicates the rows that have the greatest and least number of entries.

```
msno.bar(merged_df[missingdata_df], color="blue", log=True,
figsize=(30,18))
```

Fig 3.2 The Missingno. bar chart visually represents the nullity within the dataset.

Figure 3.2 the missing no. bar chart visually represents the nullity within the dataset. The nullity in the dataset is shown visually in the bar chart produced by the code in Figure 3.2. To improve the visibility of features with large missing values, we transformed the data on the y-axis using a logarithmic function, as shown in Figure 3.3.

A simple correlation heatmap produced by the code in Figure 3.2 is shown in Figure 3.4. The degree of nullity association between various features is displayed in this heatmap, which ranges from -1 to 1 ($-1 \leq R \leq 1$). Features in the heatmap that have no missing data are not included. No correlation data is shown when the nullity correlation is near zero ($-0.05 < R < 0.05$). While a perfect negative nullity correlation ($R=-1$) implies one feature is missing more than the other, a perfect positive nullity correlation ($R=1$) indicates both features have comparable missing values in Figure 3.5

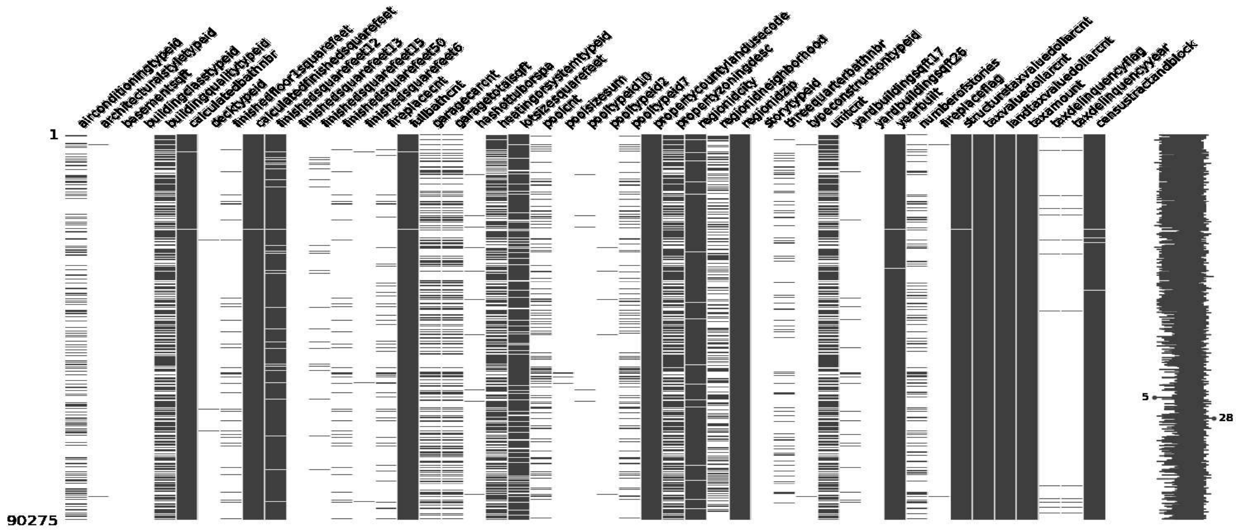


Figure 3.3: The Nullity matrix provides a data-dense display to identify the dataset's missing patterns.

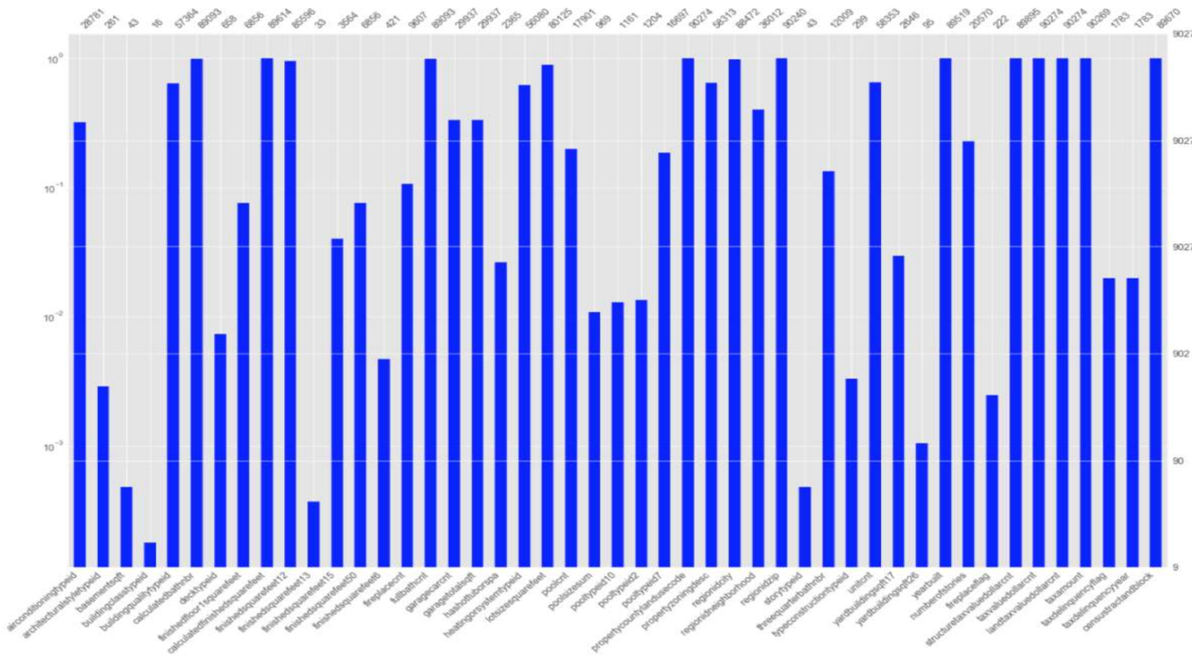


Figure 3.4: Y-axis data to enhance the visualization of characteristics with significantly high missing values

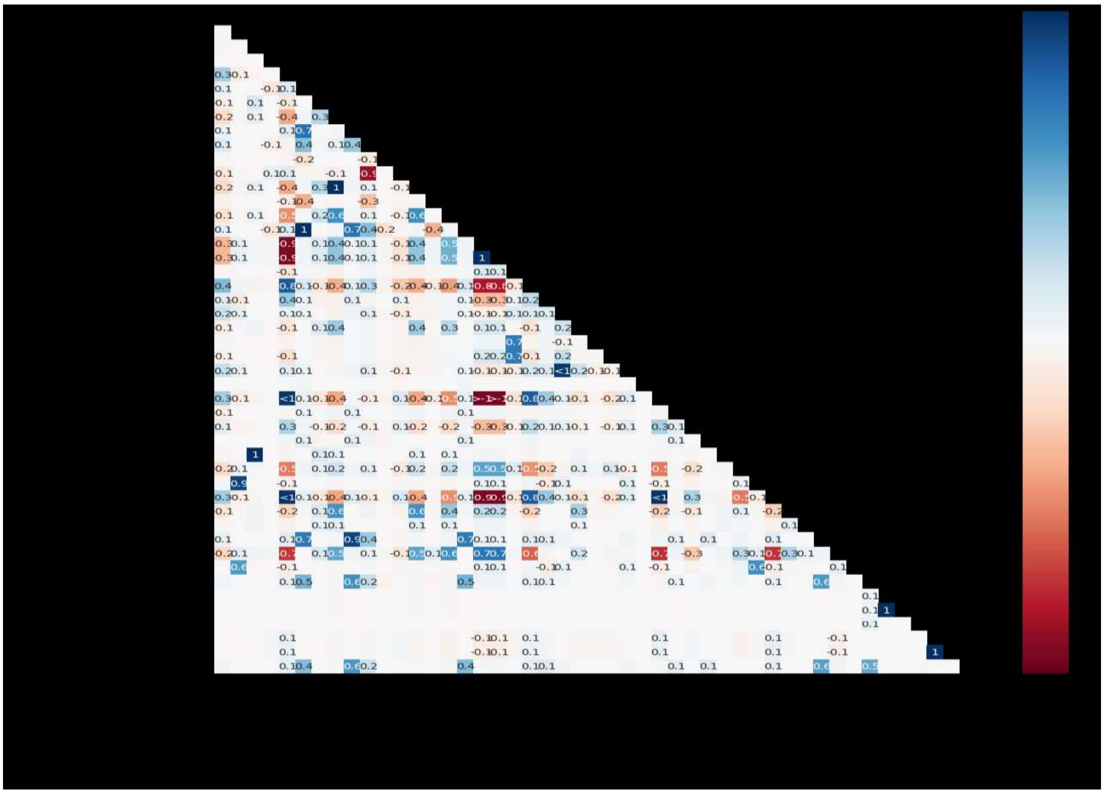


Figure 3.5: A basic heatmap of correlation that illustrates the degree of nullity in the correlation between the various features

3.4 Treating Missing Values

Various techniques have been developed for addressing missing values in numeric datasets, though their applicability to ordinal data sets may differ [11]. When addressing missing values, it is worth noting several frequently used techniques. There are several strategies for imputing missing values in predictive modeling:

- **The Case Deletion (CD)** must remove records containing missing data, thereby generating a revised dataset for subsequent analysis. Nevertheless, this approach may not be suitable for datasets with a substantial proportion of missing values. Even in cases with fewer missing entries, it is essential to assess the potential bias introduced by the modified dataset [9]
- **Random Value Imputation** is a method for filling in missing values, resulting in a complete dataset. Although simple, this approach does not

utilize figures from the dataset and may introduce randomness that impacts further analysis.

- **The Mean Imputation (MI)** involves filling in missing values with the mean value of the respective feature or attribute from the complete dataset. This method, also referred to as complete mean imputation, has limitations. It may not be ideal for datasets with many missing values, as it reduces variance and can inflate the apparent sample size.
- **The Most Common Imputation (MCI)** technique replaces missing values with the most frequently occurring value in the dataset. This approach assumes that the most common value symbolizes a plausible estimate for the missing data.
- **The Median Imputation** fills in missing values by using the dataset's median value of the respective feature. Alternatively, class median imputation replaces missing values with the median of the feature within a particular class. The class should correspond to the class variable of the vector containing the missing value.

Various strategies have been suggested for managing missing data in ordinal datasets. Decision trees have proven effective in classifying ordinal data by organizing data into splits or branches. Traditional methods for handling missing values might introduce bias and diminish or amplify statistical power. Removing missing instances is often favored for simplicity and is commonly the default procedure in statistical data analysis tools. However, this approach may lead to the loss of a substantial portion of the data in practical scenarios.

Neural networks provide a viable solution for handling missing values in ordinal data. They offer a classifier-based imputation method inspired by the functioning of the brain. Neural networks usually comprise an input, hidden, and output layer, with training algorithms enabling them to tackle intricate mathematical problems. The hidden layer adjusts dynamically during the training process. Another classifier-based approach for addressing missing data is Support Vector Machines (SVM).

Classification-based imputation methods have emerged as effective strategies for estimating and filling missing values within datasets. These approaches utilize diverse classification techniques, including neural networks, decision trees, and similar methods, to address the challenge of missing values. However, previous research has primarily concentrated on managing numerical or nominal missing data values. In contrast, this study seeks to address this research gap by focusing on the treatment of missing values in ordinal data and examining the effects of these treatments on unsupervised learning techniques, particularly clustering.

a) **Mean/Median/Mode Imputation:** This involves replacing missing values with the mean, median, or mode of the observed values for that variable. This is a simple approach, but it might not capture complex relationships in the data.
b) **Regression Imputation:** This technique involves using other variables in the dataset to predict and fill in the missing values. Regression models can be used for numerical variables, while classification models can be employed for categorical variables.
c) **K-Nearest Neighbors (KNN) Imputation:** KNN imputation entails identifying the k-nearest data points with complete information and then averaging or employing their values to fill in missing data points.
d) **Multiple Imputations:** Multiple imputations generates several imputed datasets, each with a different set of imputed values, to account for

uncertainty in imputation. These datasets are then used to build models, and their predictions are combined for more robust results.

e) **Interpolation and Extrapolation:** For time-series data, missing values can be estimated using interpolation (estimating values within the observed range) or extrapolation (estimating values outside the observed range).

It is important to note that the choice of imputation method depends on the nature of the data, the extent of missing values, and the specific goals of the predictive modeling task. Incorrect imputation can introduce biases or distort the relationships in the data, leading to poor model performance.

3.5 Impact of Missing Data on Predictive Models

The issue of missing values is pervasive across all data-driven domains, posing numerous challenges such as diminished performance, analytical hurdles, and the risk of biased outcomes due to disparities between incomplete and complete data. Enhancements are necessary in managing and disclosing missing data within predictive modeling studies. A widely endorsed approach to mitigate bias involves employing multiple imputation techniques. Additionally, exploring machine learning algorithms equipped with innate mechanisms for addressing missing data presents another promising avenue for consideration.

3.6 How missing Data can affect model Performance and accuracy

Missing values can significantly impact the performance and accuracy of predictive models. Here is how missing data can affect models:

a) **Bias in Model Parameters:** When data is missing not at random (MNAR), meaning that the probability of missing data depends on unobserved factors, imputing missing values with simple methods like mean or median can introduce bias in the model's parameters. This bias can lead the model to misrepresent the relationships between variables.

b) **Reduced Sample Size:** Missing data effectively reduces the dataset size available for model training. With fewer data points, the model may need more information to learn complex patterns, leading to overfitting the available data.
c) **Loss of Information:** Missing data can result in the loss of valuable information and insights that those variables could have provided. This loss of information can lead to less informed decisions and predictions.
d) **Incorrect Relationships:** Imputing missing values improperly can distort the relationships between variables. If the imputed values are systematically different from the true values, the model will learn relationships that don't accurately reflect reality.
e) **Increased Variability:** Imputing missing values can artificially increase the variability in the dataset, as imputed values often introduce variability that might not exist in the actual data distribution. This can lead to increased uncertainty in model predictions.
f) **Inflated Significance:** In some cases, imputed values might introduce noise that gets incorporated into the model. This noise can lead to falsely significant relationships and features in the model, impacting its generalization to new data.
g) **Incorrect Outcomes:** In specific scenarios, missing data can cause incorrect outcomes. For example, in medical studies, missing data on treatment outcomes could lead to incorrect medical decisions.
h) **Model Instability:** Models built on datasets with missing values can be less stable. Small changes in the imputed values or the handling of missing data can lead to different model outcomes, reducing the model's reliability.

To mitigate these issues and minimize the impact of missing data on model performance and accuracy:

3.7 Model Implementation Process

a) **Choose Appropriate Imputation Methods:** Select imputation methods suitable for the data type and context. More advanced imputation techniques that incorporate relationships between variables can help mitigate biases and inaccuracies.

b) **Evaluate Imputation Quality:** When available, assess the quality of imputed values by comparing them to the observed values. If imputed values deviate significantly from observed values, they might not accurately represent reality.

c) **Consider Multiple Imputations:** Implement multiple imputations to account for the uncertainty introduced by imputed values. This involves creating several imputed datasets and aggregating results for more robust model training and evaluation.

d) **Use Feature Engineering:** Create additional features that capture the missing data for a particular variable. This can help the model learn from the missingness pattern.

e) **Analyse Missing Data Mechanism:** Understand why data is missing and whether the missing data mechanism is random, missing at random (MAR), or MNAR. This can guide the imputation strategy.

f) **Assess Model Sensitivity:** Test the model's sensitivity to missing data by analyzing its performance on complete cases and different imputation scenarios. This can give the insights into the model's robustness to missing data.

3.8 Bias and potential pitfalls introduced by missing data

Missing data can introduce bias and pitfalls into data analysis and modeling processes. Here are some ways bias can arise and potential pitfalls associated with missing data:

a) **Selection Bias:** When the missing data is not completely random, it can lead to selection bias. This occurs when the probability of data being missing depends on the unobserved value itself or another variable. This bias can distort the relationships between variables and lead to incorrect conclusions.

b) **Underestimation or Overestimation of Relationships:** If missing data is not handled correctly, it can lead to underestimation or overestimation of relationships between variables. Imputed values might not accurately represent the actual values, leading to incorrect correlations, coefficients, and impact assessments.

c) **Bias in Imputation Methods:** Imputing missing values using inappropriate methods or models can introduce bias. For instance, using a linear regression model to impute missing data in a nonlinear relationship can lead to biased imputations.

d) **Impact on Model Generalization:** Models trained on datasets with imputed values can need help to generalize to new, unseen data. If imputed values introduce noise or distort relationships, the model's performance on real-world data may suffer.

e) **Artificial Inflation of Statistical Significance:** Imputing missing values can introduce variability, and this artificially increased variability can lead to falsely significant p-values in statistical tests, making some relationships appear statistically significant when they are not.

f) **Misleading Interpretations:** Missing data can lead to misleading interpretations of results. Analysts might draw conclusions based on incomplete or biased data, leading to decisions that must be grounded in reality.

g) **Inaccurate Insights:** Missing data can lead to accurate insights and informed decisions. Models and analyses based on complete data may provide reliable recommendations or predictions.

h) **Inconsistent Comparisons:** If different subsets of data have varying levels of missingness, comparisons between groups can become problematic. The missingness can disproportionately affect one group, leading to incorrect comparisons.

i) **Propagation of Error:** If imputed values are used as inputs for subsequent analyses, any errors in imputation can propagate throughout the analysis pipeline, leading to further inaccuracies.

j) **Model Instability:** Models built on datasets with missing values can be less stable. Small changes in imputed values or handling of missing data can lead to different model outcomes, reducing confidence in the results.

Adopting careful data preprocessing and imputation strategies is essential to mitigate these biases and pitfalls associated with missing data.

3.9 Machine Learning-Based Imputation

Machine learning-based imputation techniques leverage machine learning algorithms' power to predict and fill in missing values in datasets. These techniques can be instrumental when dealing with complex and high-dimensional datasets where traditional imputation methods may need to perform better. Machine learning has been the angled stone in studying and extracting information from data, and often, a problem of missing values is encountered. Instead of using traditional statistical methods or simple imputation strategies like mean or median, machine learning models are used to predict the missing values based on the patterns and relationships present in the data.

After thorough discussions, it has been determined that Machine Learning (ML) models can be categorized into five different groups: Classification, Regression, Clustering, Association rule learning, and Reinforcement learning. Figure 3.6 shows the ML models featured in the selected articles for this review.

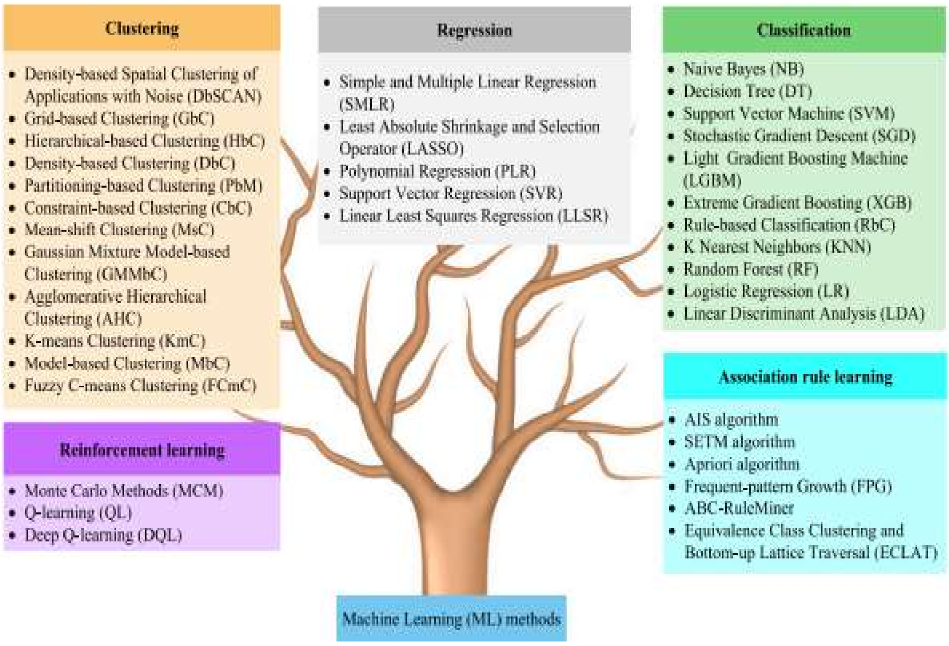


Figure 3.6 A hierarchical diagram illustrating various ML models categorized into five groups based on their similarities.

Here is the description of five different groups of machine learning-based imputation techniques:

a) Clustering based imputation

Clustering-based imputation is a machine-learning technique for filling in missing values in datasets by grouping similar data points together and imputing missing values within each cluster. This approach assumes that data points within the same cluster share similar characteristics, and missing values can be estimated based on the values of other data points within the same cluster.

b) Regression-based imputation

Regression-based imputation is a technique used to estimate and fill in missing values in a dataset by creating regression models to predict the missing values based on the relationship between the variable with missing data and other relevant variables. This approach is beneficial when there is a meaningful correlation or association between the variable with missing values and other variables in the dataset.

c) Classification-based imputation

Classification-based imputation is a technique for filling in missing values in a dataset by predicting and assigning class labels to the instances with missing data. This method is precious when dealing with categorical or discrete variables. It involves building a classification model to forecast the missing categorical values based on other obtainable features.

d) Reinforcement learning-based imputation

Reinforcement learning-based imputation is a novel approach that uses reinforcement learning techniques to fill in missing values in a dataset. Unlike traditional imputation methods that rely on statistical or machine learning models, reinforcement-based imputation treats imputation as a sequential decision-making problem. It uses reinforcement learning agents to learn how to impute missing values by interacting with the dataset over multiple iterations.

e) Association rule learning-based imputation.

Association rule learning-based imputation fills in missing values in a dataset by leveraging association rule mining algorithms. This approach is particularly suitable for categorical data and relies on identifying associations between variables to impute missing values.

3.9.1 Process of machine learning-based imputation

a) **Feature Selection and Engineering:** The first step is identifying relevant features (variables) to help predict the missing values. Sometimes, additional features must be created or engineered to capture the relationship between variables better.

b) **Data Splitting:** The dataset is split into two parts: one with complete data and another with missing values. The complete data portion is used to train the machine learning model, and the portion with missing values is used to test the imputation.

c) **Model Training:** A machine learning model is chosen based on the characteristics of the data. Common choices include regression models (linear regression, decision trees, random forests), k-nearest neighbors (KNN), support vector machines (SVM), or even neural networks.

d) **Prediction:** The model is trained on the complete data, using the available features as inputs and the variable with missing values as the target. Once trained, the model is used to predict the missing values in the test portion of the dataset.

e) **Evaluation:** The imputed values are compared to the observed values for the missing entries. Various metrics, such as mean squared error, mean absolute error, or correlation coefficients, can be used to assess the quality of the imputations.

f) **Application to Entire Dataset:** If the imputation model performs well on the test portion, it can be applied to the entire dataset to fill in all the missing values.

3.9.2 Leveraging predictive modeling for imputing missing values

Leveraging predictive modeling for imputing missing values involves using machine learning algorithms to predict missing values based on the observed data. This approach is beneficial when there are complex relationships between variables, and traditional imputation methods may need to capture these relationships more effectively. Here is how predictive modeling can be applied for missing value imputation:

a) **Feature Selection:** Identify relevant features (variables) correlated with the variable containing missing values. Feature selection techniques such as correlation analysis or feature importance from tree-based models can help identify these variables.

b) **Model Training:** Based on the characteristics and nature of the missing values' data, select an appropriate machine learning algorithm. Common choices include regression models (e.g., linear regression, decision trees, and random forests), support vector machines (SVM), or deep learning models (e.g., neural networks).

c) **Train-Test Split:** Split the dataset into a training set (with non-missing values) and a test set (with missing values). The training set is used to train the predictive model, while the test set is used to evaluate its performance.

d) **Model Training and Evaluation:** Train the predictive model using the training data, treating the variable with missing values as the target variable. Evaluate the model's performance using appropriate metrics (e.g., mean squared error for regression models) on the test set to ensure its accuracy and generalization capability.

e) **Prediction:** Based on the observed data, use the trained model to predict missing values in the test set. These predictions serve as imputed values for the missing entries.

f) **Model Validation:** Validate the imputation results by comparing the predicted values with the true values (if available) or by assessing the impact of imputation on downstream analyses or predictive tasks.

g) **Integration with Imputation Pipelines:** Integrate predictive modeling into imputation pipelines alongside other imputation techniques. Ensemble methods or multiple imputations can combine predictions from multiple predictive models for enhanced imputation accuracy.

Predictive modeling for missing value imputation offers several advantages, including the ability to capture complex relationships and interactions between variables, handle nonlinearity, and adapt to the dataset's specific characteristics. However, careful model selection, feature engineering, and validation are required to ensure reliable imputation results.

3.10 Deep Learning-Based Imputation:

Deep learning-based imputation is a technique for filling in missing data in a dataset using deep neural networks. Deep learning models, a subset of machine learning models, have gained popularity for their ability to learn complex patterns and relationships in data, making them suitable for handling missing data in various applications.

In recent years, deep learning (DL) methods have increasingly addressed missing value challenges and showcased improved imputation accuracy [10, 11]. These DL models can be modified to handle complicated missing patterns [12, 13] and diverse data structures, such as time-series data with sequential arrangements and image data with spatial characteristics. Their superior performance and adaptable design have propelled their adoption across various domains, including in-patient mortality prediction [14, 15] and early Alzheimer's disease detection [12, 16]. Despite existing reviews on missing value imputation, many either focus on non-DL methods or treat neural networks as a monolithic approach, needing more specificity to guide researchers in applying DL models to their unique datasets [17-19]. To fill this void, we introduce a systematic review encompassing DL-based missing value imputation methods across diverse datasets [20-24]. Our evidence map analysis examines model usage based on data types, offering valuable insights and guidance for researchers leveraging DL methodologies to address missing data challenges effectively [25].

Here are the key steps and considerations when using deep learning-based imputation.

a) **Autoencoders:** Autoencoders are a type of neural network architecture used for unsupervised learning. They consist of an encoder network that maps input data to a lower-dimensional representation (encoding) and a decoder network that reconstructs the input data from this encoding. Autoencoders can be used for imputation by training them on the complete data and then using the decoder to generate imputed values for missing data points.
b) **Variational Autoencoders (VAEs):** VAEs are variants of autoencoders that can capture the probabilistic distribution of data in the encoding space. They are helpful for imputation because they provide not only a point estimate for the missing values but also an estimate of uncertainty. This can be particularly valuable when dealing with noisy or uncertain data.
c) **GANs):** GANs consist of a generator network that produces data samples and a discriminator network that distinguishes between accurate and generated data. GANs can be adapted for imputation by training them on the complete data and using the generator to generate imputed values for missing data points. GANs can produce realistic imputations that are consistent with the underlying data distribution.
d) **Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM):** RNNs and LSTMs are specialized neural network architectures for sequence data. They can be used for imputation in time series or sequential data by learning temporal dependencies and filling in missing values based on the context of nearby observations.

e) **Transformers:** Transformers, known for their success in natural language processing tasks, can also be adapted for imputation. They excel at capturing long-range dependencies in data and can be used for imputation in various structured or unstructured datasets.

f) **Data Augmentation:** Deep learning models can also be used to perform data augmentation, where synthetic data is generated to supplement the existing dataset. This can help balance class distributions or increase the amount of training data for imputation tasks.

Deep learning-based imputation methods have the advantage of being able to capture intricate patterns and dependencies in the data. However, they also require large amounts of data and computational resources for training and hyperparameter tuning to achieve optimal performance. Additionally, depending on the dataset size and complexity, they may only sometimes be necessary or suitable for all imputation tasks;

3.11 Introduction to using neural networks for imputation.

Using neural networks for imputation is a practical and powerful approach to fill in missing data in datasets. Neural networks, a subset of deep learning techniques, can capture complex patterns and relationships in data, making them well-suited for imputation tasks. Here is an introduction to using neural networks for imputation:

a) **Understanding the Imputation Problem:**

Imputation is the Process of filling in missing values in a dataset. Missing data can arise for various reasons, including sensor errors, data collection issues, or incomplete records.

Accurate imputation is crucial because missing data can lead to biased analyses, reduced model performance, and incomplete insights.

b) **Data Preprocessing:**

Before using neural networks for imputation, it's essential to preprocess the data:
Identify missing values: Determine which features or variables have missing data.
Handling categorical data: Convert categorical variables into numerical representations, such as one-hot encoding.
Normalize/standardize numerical features: Scaling numerical features can help neural networks converge faster.

c) **Choosing the Right Neural Network Architecture:**

The choice of neural network architecture depends on the nature of the data and the imputation task. Common architectures include:
Feed forward Neural Networks (FNN): Suitable for tabular data and structured datasets.
Recurrent Neural Networks (RNN) or Long-Short-Term Memory (LSTM) are ideal for sequential data with temporal dependencies.
Autoencoders: Effective for capturing complex relationships in data for imputation.

d) **Data Splitting:**

Divide the dataset into training, validation, and test sets. The training set is used to train the neural network, the validation set is used to tune hyperparameters and monitor model performance, and the test set is used for final evaluation.

e) **Training the Neural Network:**

During training, the neural network learns to predict missing values based on the available data. Define a loss function quantifying the error between predicted imputations and actual values. Choose an optimization algorithm (e.g., Adam, SGD) to minimize the loss function. Train the neural network on the training data for sufficient epochs.

f) **Hyperparameter Tuning:**

Experiment with various hyperparameters, such as the number of layers, neurons per layer, learning rate, batch size, and activation functions. Use the validation set to fine-tune hyperparameters and avoid overfitting.

g) **Evaluation:**

Assess the imputation performance using appropriate metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or others relevant to the specific problem. Evaluate the model on the test set to ensure it generalizes well to new, unseen data.

h) **Post-processing:**

Depending on the application, one may need to post-process the imputed values, e.g., by rounding to integers, converting back to categorical values, or applying domain-specific rules.

i) **Deployment:**

Once the neural network model demonstrates satisfactory imputation performance, it can be deployed to handle missing data in real-world applications.

Using neural networks for imputation can be a valuable data preprocessing and analysis tool. However, when applying these techniques, it's essential to choose the right architecture, perform rigorous validation, and consider the specific characteristics of the dataset and problem.

3.12 Auto encoders and their role in imputing missing values

Autoencoders are a type of neural network architecture that can be valuable in imputing missing values in datasets. They are instrumental when dealing with high-dimensional data or data with complex dependencies between variables. Here is how autoencoders work and their role in imputing missing values:

a) Autoencoder Architecture:

An autoencoder consists of two main parts: an encoder and a decoder.

Encoder: The encoder network takes the input data and maps it to a lower-dimensional representation, often referred to as the "encoding" or "latent space." This Process reduces the data's dimensionality while capturing its essential features.

Decoder: The decoder network takes the encoded representation and attempts to reconstruct the original input data.

b) **Training an Autoencoder:**

Autoencoders are typically trained unsupervised, meaning they learn to encode and decode data without explicit labels. During training, the goal is to minimize the reconstruction error, which measures how well the autoencoder can reconstruct its input data.

The loss function used for training is often a measure of dissimilarity between the input and reconstructed data, such as Mean Squared Error (MSE).

c) Role in Imputing Missing Values:

Autoencoders can be used for imputation by leveraging their ability to capture data patterns and relationships. When dealing with a dataset containing missing values, one can use the autoencoder to predict or impute the missing values based on the available data.

Here is a common approach to using autoencoders for imputation:

- Input Preparation: Encode the dataset with missing values such that missing values are replaced with zeros or other placeholders.

- Training: Train the autoencoder using the complete data (i.e., data without missing values). The encoder learns a compact representation of the data, and the decoder learns to reconstruct it.
- . Encoding: Use the trained encoder to encode the dataset with missing values, generating the encoded representations for all data points.
- Imputation: Replace the encoded values corresponding to missing entries with zeros or placeholders.
- Decoding: Pass the modified encoded data through the decoder to obtain the imputed data, including the values for previously missing entries.

d) Benefits of Autoencoders for Imputation:

Autoencoders can capture complex relationships and patterns in the data, making them suitable for imputing missing values even in high-dimensional datasets. They can learn a lower-dimensional data representation that often highlights the most critical features, helping reduce dimensionality. Autoencoders can handle various data types, including numerical, categorical, and mixed data.

e) Considerations:

When using autoencoders for imputation, it is essential to preprocess the data appropriately and choose an appropriate loss function and hyperparameters. The choice of architecture (e.g., the number of layers and units) can significantly impact imputation performance. Autoencoders may need help imputing missing values for extremely rare or novel patterns in the data.

Autoencoders can be a powerful tool in the data preprocessing toolbox for imputing missing values, especially when dealing with complex datasets where traditional imputation methods may be less effective. However, they require careful tuning and validation to provide accurate and meaningful imputations.

3.13 Summary of the Chapter

In summary, missing values imputation in predictive modeling is a crucial step to ensure that the model learns from a complete and representative dataset. This, in turn, leads to more accurate and reliable predictions of new, unseen data.

Missing data can introduce various challenges and biases to predictive modeling. Proper handling and imputation strategies are essential to maintain the model's predictions' integrity, accuracy, and reliability.

Understanding the potential biases and pitfalls introduced by missing data and taking appropriate measures to handle them is crucial for producing reliable and accurate analyses and models.

Machine learning-based imputation leverages the power of machine learning algorithms to predict missing values in a dataset. While it can offer improved accuracy over traditional methods, it also requires proper data preprocessing, feature engineering, and model selection to ensure effective results.

Chapter-4 Futuristic Prediction of Food Consumption with Missing Value Imputation Methods Using Extended ANN

4.1 Introduction

Missing data is a pervasive challenge encountered across various research domains, contributing to the approach of uncertainty in data analysis. Missing data imputation strategies concentrate on arithmetical prediction algorithms to replace missing values. Recurrent Neural Networks, Iterative KNN Imputation, K-Nearest Neighbors, and Artificial Neural Networks are some methods used in research for missing value imputation and prediction. We compared these methods by filling in the mean and median. We used a dataset from national FCDBs that OpenMV.net collected. The dataset shows different food items used in Scandinavian and European countries. The results show that state-of-the-art imputation methodologies produce significantly better results than traditional methods. Predicting futuristic food consumption with our proposed missing value imputation methods using an Extended Artificial Neural Network (EANN) is an exciting application of machine learning and data analysis. Extended ANN refers to neural network architectures that handle more complex data and tasks.

4.2 Conventional techniques for borrowing food consumption databases

Conventional techniques for borrowing food consumption databases typically involve several steps:

a) Data Collection and Preprocessing:

Gather historical food consumption data, including time, location, demographics, and dietary habits. Researchers used data from national FCDBs that collected

OpenMV.Net. In this thesis, the researchers try to handle missing values in the dataset. The missing value imputation methods one can use include:

b) Mean/Median Imputation: Replace missing values with the mean or median of the respective feature.

c) RNN Imputation: Imputing missing values using Recurrent Neural Networks can be a powerful technique when dealing with sequential data where the missing values depend on previous values.

d) KNN Imputation: Imputing missing values using the K-Nearest Neighbors (KNN) algorithm is a straightforward and effective technique, mainly when dealing with tabular data. KNN imputation involves estimating missing values by considering the values of the nearest neighbors in the dataset.

e) ANN Imputation: Train a neural network model for imputing missing values. Normalize or standardize the data to ensure all features have the same scale. This can be part of the Extended ANN.

4.3 Proposed Phenomenon

Create relevant features that influence food consumption. This could include economic indicators, weather, cultural events, and more. The researcher considers using techniques like one-hot encoding for categorical variables.

a) Data Splitting:

Here, we split our dataset into training, validation, and test sets. The training set should contain historical data, the validation set can be used for hyperparameter tuning, and the test set can be used to evaluate the final model.

b) Extended Artificial Neural Network (EANN):

Design an Extended ANN architecture that can capture the complexity of the food consumption prediction task. Researchers may consider using deep neural networks with multiple layers and different activation functions, including dropout layers, to prevent overfitting. Experiment with various imputation techniques and neural network architectures, such as K-Nearest Neighbors (KNNs), artificial neural networks (ANNs), and recurrent neural networks (RNNs), for missing value imputation and prediction in food consumption datasets.

c) Training and Hyperparameter Tuning:

Train the Extended ANN on the training dataset using appropriate loss functions (e.g., mean squared error for regression). Tune hyperparameters like learning rate, batch size, number of layers, and neurons per layer using the validation set.

d) Missing Value Imputation within the ANN:

Implement a specific component within the ANN to handle missing value imputation. This can be completed using methods like autoencoders. A neural network architecture where the input and output layers are the same and an intermediate bottleneck layer encrypts information about the missing values.

ANN trains a generator network to produce imputed values, while a discriminator network assesses imputation quality. Moreover, use attention mechanisms within the neural network to focus on relevant information when imputing missing values.

Evaluate the Extended ANN's performance on the test set using appropriate evaluation metrics (RMSE for regression tasks). Analyze the quality of the imputed missing values and their impact on the overall predictions.

e) Futuristic Predictions:

Once we have a well-trained model, we can use it to predict future food consumption by providing relevant input data for the future period. Continuously monitor and retrain the model's performance as new data becomes available. Be aware of concept drift, where the relationships between features and food consumption may change. Consider techniques to make the model's predictions more interpretable, significantly if the results impact decision-making.

f) Ethical Considerations:

Be mindful of ethical considerations related to food consumption predictions, such as privacy and data biases. This approach leverages the power of extended ANN architectures to handle complex patterns and relationships in data while addressing missing values. However, remember that building and training such models can be computationally intensive, and we may need access to substantial computing resources for successful implementation.

4.3.1 Data searching and analysis

Data searching and data analysis for missing values imputation is exploring, classifying, and preparing food consumption datasets to handle the missing data effectively. This critical step sets the foundation for choosing the most suitable imputation technique and ensuring the data is accurate and reliable. Here are the essential aspects of data searching and data analysis for missing values imputation. The collection of the national FCDBs dataset of countrywide constituents used to determine potassium supply is shown in Table 4.1.

Table 4.1: Data containing potassium values for various foods sourced from multiple national FCDBs

Country	Germany	Italy	France	Holland	Belgium	Luxembourg	England	Portugal	Austria	Switzerland	Sweden	Denmark	Norway	Finland	Spain	Ireland
Real coffee	91	81	87	98	93	98	28	73	56	74	98	97	91	98	70	31
Instant coffee	48	11	41	61	39	62	87	27	32	73	14	18	18	13	41	51
Tea	89	61	63	97	49	87	98	78	62	86	94	91	84	85	44	98
Sweetener	19	4	5	31	12	29	23	3	16	26	32	36	14	21	62	12
Biscuits	56	54	76	63	75	80	92	23	30	32	44	67	63	65	41	81
Powder soup	52	42	51	68	38	72	54	35	33	70	44	31	52	28	3	76
Tin soup	20	4	12	42	24	13	75	2	2	11	38	18	5	11	15	19
Potatoes	21	3	23	8	10	8	18	6	6	18	55	12	18	9	24	3
Frozen fish	27	4	12	15	14	27	21	21	16	20	46	52	31	19	8	6
Frozen veggies	22	3	6	15	13	24	25	4	12	16	55	41	16	13	58	4
Apples	82	68	88	84	77	86	77	21	50	80	79	82	62	51	78	58
Oranges	76	72	85	90	77	95	90	52	43	71	54	72	71	58	31	52
Tinned fruit	45	10	41	60	43	82	89	9	15	47	76	51	35	21	39	47
Jam	72	47	46	82	58	21	92	17	42	63	10	65	51	38	87	90
Garlic	21	81	87	16	28	93	12	88	52	65	69	12	12	16	45	6
Butter	92	67	95	33	85	95	96	66	53	81	33	91	64	96	52	98
Margarine	86	25	48	98	81	95	95	79	74	49	49	92	95	95	92	26
Olive oil	75	95	37	14	84	85	58	93	29	63	3	31	29	18	17	32
Yogurt	31	6	58	54	22	32	12	7	14	49	94	12	3	65	14	4
Crispbread	27	19	4	16	6	25	30	8	12	31		35	61			10

a) Identify Missing Values:

First, the process starts by identifying which columns or features in the dataset contain the missing values. Depending on the dataset format, missing values can be represented as "NaN," "null," "NA," blank, or any other placeholder.

Here, we describe and understand the Nature of Missing Data. We analyze why the data is missing. Missing data can be categorized mainly into three categories:

Missing completely at Random (MCAR): The Data is missing randomly across observations. Missing at Random (MAR): The Data is missing depending on the values of other variables. Missing not at Random (MNAR): The Data is missing based on unobserved factors, and the mechanism is not related to other variables.

a) Determine Missing Data Patterns:

Identify any patterns or dependences in the missing data. For instance, do certain variables tend to have missing values together?

b) Data Exploration and Visualization:

Explore the dataset visually and statistically. Generate summary statistics, histograms, box plots, and correlation matrices to understand the distribution and relationships among variables.

c) Imputation Strategy Selection:

Choose an appropriate imputation strategy based on the nature of the missing data and the specific goals of analysis. Standard imputation methods include mean imputation, median imputation, mode imputation, K-nearest neighbors imputation, MICE imputation, RNN imputation, ANN imputation, or Consider Multiple Imputation and other machine learning-based imputation techniques. In some cases required, multiple imputation techniques, such as MICE (Multiple Imputation by Chained Equations), can be beneficial. This involves creating multiple imputed datasets and combining their results to obtain more accurate estimates.

d) Evaluate the Impact of Imputation:

Before and after the imputation, assess how the imputation process affects the data distribution, statistical properties, and any relationships between variables. This helps safeguard that imputed values are reasonable and do not introduce bias.

e) Assess Imputation Quality:

If we have access to the ground truth (e.g., by using synthetic missing data for testing), evaluate the quality of the imputed data using appropriate metrics like RMSE (Root Mean Squared Error).

f) Document the Process:

Keep detailed documentation of the data searching, analysis, and imputation steps. This documentation helps in transparency, reproducibility, and sharing insights with other team members or stakeholders.

h) Iterate as Needed:

If the quality of imputed data could be more satisfactory or if the analysis reveals issues, iterate on the imputation strategy or consider alternative approaches.

Data searching and analysis for missing values imputation is crucial to the data preprocessing pipeline. It ensures that the imputed data maintains the integrity of the analysis and provides meaningful results. The choice of imputation method should align with the nature of the data and the goals of the analysis.

4.4 Multivariate imputation by chained equations (MICE)

Multivariate Imputation by Chained Equations (MICE) is a statistical technique for handling missing data in multivariate datasets. Its primary role is to provide a systematic and statistically sound approach to impute missing values in a way that preserves the relationships and structure within the data. It is also known as Fully Conditional Specification (FCS) or Sequential Regression Multiple Imputation. MICE is a flexible and widely used method for imputing missing values in a dataset by iteratively imputing one variable at a time while considering the relationships between variables.

4.4.1 Here are the critical roles of MICE in missing values imputation:

Preserving Multivariate Relationships: MICE recognizes that variables in a dataset are often interrelated, and it leverages this information to impute missing values. By considering the relationships between variables, MICE can produce more accurate imputations than univariate methods that treat each variable in isolation.

Initialization: The process begins by initializing the missing values with some initial estimates. This could be mean imputation, random imputation, or any other reasonable method.

Iterative Imputation: MICE operates iteratively, typically using the following steps for each variable with missing data:

Step 1- Choose a variable as the target variable for imputation.

Step 2- Treat all other variables (including those with imputed values) as predictor variables.

Step 3- Build an imputation model using the target variable as the dependent variable and the predictor variables as independent variables. This model estimates the missing values of the target variable.

Step 3- Update the imputed values for the target variable based on the model's predictions.

Step 4- Repeat these steps for each variable with missing data until Convergence is achieved.

Multiple imputed datasets are generated once all iterations are completed, each representing a plausible set of imputations. These datasets can be used for subsequent analyses. After imputing missing values for one variable, the dataset is updated, and the imputed values are incorporated into the dataset for the next iteration.

Algorithm 1 presents the method to impute the missing values in the given data set using the MICE algorithm. As depicted in Algorithm 1, the missing information is processed by computing the Difference and mean of each data for n number of rows. Upon identifying the missing values in a given dataset, the missed entries can be filled with reshaping or other missing imputation schemes.

This statement outlines the process of imputing missing values using the MICE algorithm. The missing information is processed by calculating differences and means for each data point in a specified number of rows. Once missing values are identified, they can be filled using reshaping or other imputation techniques. The MICE algorithm typically iterates through these steps until Convergence to obtain imputed values for the missing entries.

Algorithm-I:

Input: Number of rows containing missing values in a dataset.

Output: Whether the missing values are imputed using the MICE algorithm

```
MICE = []
data2 = data
diff_mat = np.subtract(data2,data1);
mean2 = diff_mat.mean()
mean1 = 1000
while (mean2<=mean1):
   mean1 = mean2
   data2 = data1;
For i in test_missing:
y_imp = process_fn(data1,i)
data1[i[0],i[1]] = y_imp
diff_mat = np.subtract(data2,data1)
   mean2 = diff_mat.mean()
For i in test_missing:
mice.append(data2[i[0]][i[1]])
```

Funcprocess()

```
defprocess_fn(data, i):
x_train = np.delete(data, (i[0]), axis=0)
x_train = np.delete(x_train, (i[1]), axis=1)
x_test = data[:,i[1]]
x_test = np.delete(x_test,i[0])
x_missing = data[i[0],:]
x_missing = (np.delete(x_missing,i[1]))
x_missing = x_missing.reshape((1,19))
y_imp = model_fn(x_train,x_train,x_missing)
returny_imp
```

In summary, Algorithm 2 combines the MICE algorithm for initial imputation and an extended version of an Artificial Neural Network for subsequent processing and imputation of missing data. The matrix operations and model fitting steps indicate a sophisticated approach to handling missing values in a dataset, leveraging statistical and machine learning techniques.

Algorithm-II: **Proposed Algorithm**

Input: Number of rows containing missing values in a dataset.

Output: Whether the missing values are imputed using an extended ANN algorithm

```
Extended_ANN→ []                              //Initialize empty array
data2 = data                                  //data is the value of the dataset before applying imputation
data1 → [0]                                   // initiaze matrix with zero value.
diff_mat = data2 – data1                      //take difference of data2 and data1
mean2 = average(diff_mat)                     //mean2 is absolute average of the matrix diff_mat
mean1 → INT_MAX                               //initialize mean1 as max integer possible
while (mean2 <= mean1):                //iterate until error is reducing.
mean1 = mean2                                 //make mean1 as the previous value for the next iteration
data2 = data1;                                // make data2 as previous value for the next iteration
for i → Missing:                       // iterate for every missing value
imputed_value→process_fn(data1,i)             //Function for fitting the model
data1[index][index2] = imputed_value          // update the current value
diff_mat→ data2-data1                         // updating difference matrix
mean2 = diff_mat.mean()                //updating current mean
```

4.4.2 MICE has several advantages:

1-It can handle both continuous and categorical variables.

2-It considers the relationships between variables, making it suitable for complex datasets with interdependencies.

3-It provides multiple imputed datasets, allowing for uncertainty estimation.

MICE also has some limitations, such as slow Convergence, especially for large datasets with complex dependencies. The imputation process may only partially capture non-linear relationships between variables. The quality of imputations depends on the appropriateness of the chosen imputation models.

In practice, MICE is a valuable tool for addressing missing data issues. However, it is essential to carefully consider the data and the imputation models used to ensure meaningful and reliable results. Additionally, diagnostics and sensitivity analyses should be performed to assess the impact of missing data and imputation choices on the final analyses.

4.4.3 Experiment and Performance Analysis

Now, let us illustrate the concept of missing values and discuss how to address them in our research. We will use an example involving food consumption patterns across European and Scandinavian countries. This study focuses on predicting missing values to analyze the supply of food products required by these regions. The dataset includes information on food names and countries for predictive analysis.

Table 4.2. Dataset with missing values

Country	DE	IT		ES	IE
Real coffee		82		70	30
Inst. Coffee	49	10	……	40	52
			.		
			.		
Tea	88			40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	

Now, the food product supplier is seeking guidance on the types of food products to target when approaching markets in European and Scandinavian countries.

Step 1: The first phase involves filling in the missing values within the collected data. In this scenario, the known actual values will be used to impute the missing values, allowing for an analysis of the data gaps. The actual known values, highlighted in green, will be utilized to complete the missing entries.

Table 4.3. Dataset with true values to verify the model output

Country	DE	IT		ES	IE
Real coffee	90	82		70	30
Inst. Coffee	49	10	......	40	52
			.		
			.		
Tea	88	60		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	3

We will reserve this data for future reference, using it to cross-verify our model's performance and assess its accuracy.

Step 2: We will solely impute missing values in Germany, Italy, and Ireland for the three feature columns - Real coffee, tea, and yogurt, within the provided matrix as illustrated below:

Table 4.4 Dataset with missing values to verify the model output

Country	DE	IT		ES	IE
Real coffee		82		70	30
Inst. Coffee	49	10	......	40	52
			.		
			.		
Tea	88			40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	

A pertinent question arises after examining the feature matrix above: Why not employ univariate methods such as mean, median, mode, frequent values, or constants to impute the missing values? Univariate methods utilize specific column statistics, such as mean, mode, or median, to fill in missing values within that column.

In contrast, Extended Artificial Neural Network (EANN) imputation goes beyond univariate approaches by considering data from other columns, providing a more comprehensive estimation for each missing value.

To address this inquiry, let us apply the mean imputation method to the feature above matrix to address the missing values. The outcome of applying mean imputation is as follows:

Table 4.5 Dataset applying the mean imputation method

Country	DE	IT		ES	IE
Real coffee	66.6	82		70	30
Inst. Coffee	49	10	……	40	52
			.		
			.		
Tea	88	75.6		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	16

Upon reviewing the results, it becomes apparent that the mean imputation has generated values that seem implausible for Italy and Spain. For instance, Italy is shown to consume 60 units of tea and 82 units of real coffee, while Spain consumes 40 units of tea and 70 units of real coffee. Such values are inconsistent with Germany's consumption, with tea at 88 units and real coffee at 66.6 units. Additionally, Germany's consumption appears higher across other products than Italy, Spain, and Ireland. This discrepancy highlights a limitation of mean imputation, demonstrating that the method is not yielding expected outcomes.

This "brute-force" approach exemplifies one of the shortcomings of single or univariate imputation techniques.

The ANN Extended algorithm effectively addresses the issue encountered with mean imputation, which takes into account other variables in the dataset to enhance predictions of missing values. In this context, to determine the missing value in the Germany column, a regression model is applied to features, with "contrary" and "food product" serving as predictors and Germany as the target. Similar procedures are followed to obtain missing values for contrary and food product features.

Now, let us delve into the implementation of the ANN Extended algorithm. This algorithm operates iteratively, and we will explore each iteration in detail to observe how missing values are computed and assess whether the predictions closely align with the actual values.

Iteration 1

Table4.6 "*Zeroth*" dataset for iteration 1

Country	DE	IT		ES	IE
Real coffee	66.6	82		70	30
Inst. coffee	49	10	……	40	52
			.		
			.		
Tea	88	75.6		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	16

Dataset: Imputed all missing values using mean imputation

Step 1: Apply mean imputation to fill in all missing values, utilizing the mean of their respective columns. This will be referred to as our "Zeroth" dataset, and the imputation process will proceed from left to right across the columns.

Table4.6 Missing value dataset for iteration1

Country	DE	IT		ES	IE
Real coffee		82		70	30
Inst. coffee	49	10	……	40	52
			.		
			.		
Tea	88	75.6		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	16

Step 2: Exclude the imputed values for "Real coffee in Germany" and retain the imputed values in the other columns, as indicated in the provided data.

Step 3: The remaining features and rows, specifically the top 4 rows of the country, constitute the feature matrix, while "Germany and Real Coffee" serves as the target variable. A linear regression model will be executed on the filled rows, where X represents the country, and Y represents Real coffee. The row with the missing value will be employed as test data to predict the missing real coffee value. Consequently, the top 5 rows function as the training data and the first row containing the missing Real coffee value in Germany is designated as the test data. Utilizing "Country" and "Foods Product" as predictors, the model is employed to predict the corresponding "Real coffee in Germany" value, yielding a prediction of 80 in my analysis.

Step 4: Update the forecasted Real coffee value in the missing cell within the "Real coffee" column. Subsequently, the imputed value for "Tea in Italy" should be eliminated. The remaining features and rows form the feature matrix, with "Tea in Italy" designated as the target variable. A linear regression model will be applied to the filled rows, utilizing X as the country and Y as Tea. To predict the absent Tea value in Italy, the row with the missing value will be employed as the test data. The forecasted value of tea in Italy is 65.

Step 5: Revise the anticipated Tea value in Italy in the absent cell within the "Tea" column. Subsequently, eliminate the imputed value for "Yoghurt in Ireland." The remaining features and rows now constitute the feature matrix, with "Yoghurt" designated as the target variable. They employ age and experience as X variables, and Y as Yoghurt, and a linear regression model will be executed on the filled rows. To gauge the absent Yoghurt value in Ireland, the row with the missing values (white cells) will serve as the test data. The forecasted value for Yoghurt in Ireland is 8.

Let us name this as the " *First*" dataset.

This is Iteration 1, done and dusted.

Step 6: We will subtract the two datasets (zeroth and first). The resultant dataset is as follows:

Table4.7 Difference between the first two datasets

7	DE	IT		ES	IE
Real coffee	60.6	82		70	30
Inst. Coffee	49	10	… …	40	52
			. .		
Tea	88	75.6		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	16

" *Zeroth*" dataset

Minim

Country	DE	IT		ES	IE
Real coffee	80	82		70	30
Inst. Coffee	49	10	… …	40	52
			. .		
Tea	88	65		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	8

First Dataset

Country	DE	IT		ES	IE
Real coffee	19.4	0		0	0
Inst. Coffee	0	0	… …	0	0
			. .		
Tea	0	-10.6		0	0
Olive oil	0	0		0	0
Yogurt	30	0		0	-8

Difference Matrix

Upon examination, it is evident that there is a notable disparity between the two datasets, particularly in specific imputed values. We aim to minimize these differences, aiming for values close to zero. To accomplish this objective, multiple iterations are required.

The process involves repeating steps 2–6 with the updated dataset initially introduced. This cycle is reiterated until we attain a stable model, meaning that the disparity between the two most recent imputed datasets becomes extremely small, approaching zero. Technically, we conclude the iterations when a predetermined threshold is reached, or a predefined maximum number of iterations is achieved.

Iteration 2:

Table 4.8 After the first iteration, predict values

Country	DE	IT		ES	IE
Real coffee	80	82		70	30
Inst. Coffee	49	10	……	40	52
			.		
			.		
Tea	88	65		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	8

Next, we will employ the "first" dataset as our foundational dataset for imputations, discarding the "Zeroth" dataset that utilized mean imputations.

Utilizing the "first" dataset as the starting point, we will again execute steps 2–6 to predict imputed values for the initial three missing values. The results of this second iteration involve taking the first dataset, performing all imputations, subtracting the new dataset values from the original dataset, and obtaining the difference matrix, which is illustrated below:

Table4.9 First Dataset

Country	DE	IT		ES	IE
Real coffee	80	82		70	30
Instant coffee	49	10	……	40	52
			.		
			.		
Tea	88	65		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	8

minus

Second Dataset

Country	DE	IT		ES	IE
Real coffee	87	82		70	30
Instant coffee	49	10	……	40	52
			.		
			.		
Tea	88	62		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	5

Difference Matrix

Country	DE	IT		ES	IE
Real coffee	7	0		0	0
Instant coffee	0	0	……	0	0
			.		
			.		
Tea	0	-3		0	0
Olive oil	0	0		0	0
Yogurt	30	0		0	-3

Now, utilizing the "Second" dataset as our foundational dataset, we will again execute steps 2–6 to predict imputed values for the initial three missing values. The outcomes of this third iteration involve taking the Second dataset, conducting all imputations, subtracting the new dataset values from the Second dataset, and deriving the difference matrix, as illustrated below:

Table4.10 Second Dataset

Country	DE	IT		ES	IE
Real coffee	87	82		70	30
Instant coffee	49	10	……	40	52
			.		
			.		
Tea	88	62		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	5

Minim

Third Dataset

Country	DE	IT		ES	IE
Real coffee	89	82		70	30
Instant coffee	49	10	……	40	52
			.		
			.		
Tea	88	61		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	4

Difference Matrix

Country	DE	IT		ES	IE
Real coffee	2	0		0	0
Instant coffee	0	0	……	0	0
			.		
			.		
Tea	0	-1		0	0
Olive oil	0	0		0	0
Yogurt	30	0		0	-1

Now, utilizing the "Third" dataset as our foundational dataset, we will again execute steps 2–6 to predict imputed values for the initial three missing values. The results of this fourth iteration involve taking the Third dataset, conducting all imputations, subtracting the new dataset values from the Third dataset, and obtaining the difference matrix, as depicted below:

Table4.11 Third Dataset

.	DE	IT		ES	IE
Real coffee	89	82		70	30
Instant coffee	49	10	……	40	52
			.		
			.		
Tea	88	61		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	4

Minim

Forth Dataset

Country	DE	IT		ES	IE
Real coffee	89.2	82		70	30
Instant coffee	49	10	……	40	52
			.		
			.		
Tea	88	60.5		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	3.7

Difference Matrix

Country	DE	IT		ES	IE
Real coffee	0.2	0		0	0
Instant coffee	0	0	……	0	0
			.		
			.		
Tea	0	-0.5		0	0
Olive oil	0	0		0	0
Yogurt	30	0		0	-0.3

After the fourth iteration, we can see that the Difference is negligible.

Table4.12 Final imputed values

Country	DE	IT		ES	IE
Real coffee	89.2	82		70	30
Instant coffee	49	10	……	40	52
			.		
			.		
Tea	88	60.5		40	99
Olive oil	74	94		16	31
Yogurt	30	5		13	3.7

The imputed values for the fourth dataset are as follows: Real coffee in Germany is 89.2, Tea in Italy is 60.5, and Yoghurt in Ireland is 3.7. When we compare these imputed values with the actual values of the missing data, which are Real coffee in Germany = 90, Tea in Italy = 60, and Yoghurt in Ireland = 3, we observe a minimal difference. The values are almost identical. We can conclude the process since we have achieved nearly identical numbers, or we can continue with additional iterations until we reach zero differences. In this particular example, we will conclude the process. Therefore, the values from the second dataset are the final imputed values for the missing data, as illustrated in the table above.

4.4.4 Results and Comparison

We are using datasets from national FCDBs, which OpenMV.net collects. The dataset shows the comparative analysis of convinced food substances in Scandinavian and European countries. The consequences show that state-of-the-art imputation strategies give way higher outcomes than conventional techniques.

Our main aim in using the food consumption dataset is to support the decision-making for food consumption and the type of food requirements in upcoming years. The essential capacities were extracted from the century and consumption of foods. Missing data imputation methods focus on based techniques for alternate missing values with arithmetical prediction.

Due to the insufficient number of observations for the food Consumption data algorithms usage, four algorithms of the imputation of the missing values for standard samples were designated for the analysis:

- Recurrent Neural Network,
- Iterative KNN imputation method,
- K-Nearest Neighbors
- Artificial Neural networks

To provide a comparison, we utilized the Mean, Median, and Iterative Imputer algorithms with the scikit-learn library in Python. The Iterative Imputer, a multivariate imputer, gauges each feature's missing values by considering all the other features. It models each feature with missing data as a cyclical function of the remaining features. While the Iterative Imputer algorithm draws inspiration from the MICE technique, it also furnishes a single imputation rather than multiple imputations.

Implementations of our proposed hybrid Mode using feather MICE and ANN and implementing hybrid algorithm Extended ANN.

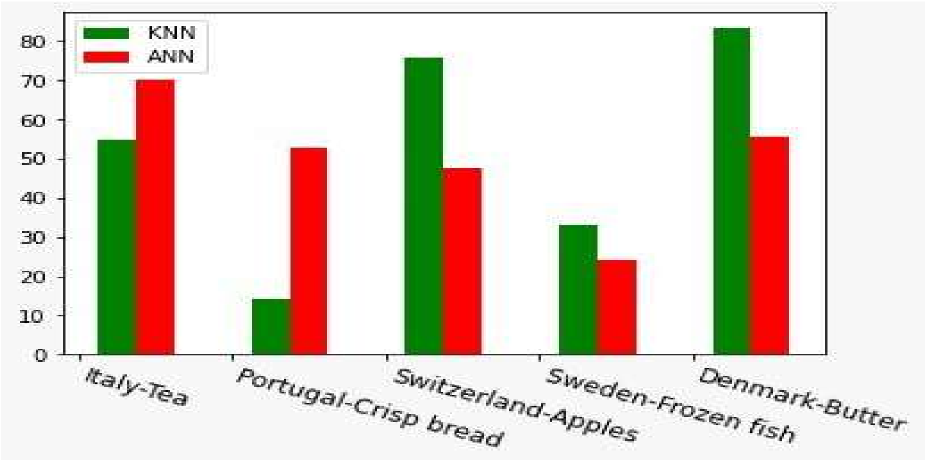


Figure 4.1: Imputation of missing values using ANN and KNN approach

In the initial step of the analysis, two distinct algorithms were applied to identical data parameters to discern which one yields superior results. Figure 4.1 illustrates the imputation of missing values using both KNN and ANN approaches. The depicted results showcase KNN's superiority over ANN, attributed to its more effective data searching and analysis, resulting in a more robust comparison than the ANN method. Additionally, the ANN approach serves as a benchmark for further comparison against alternative schemes.

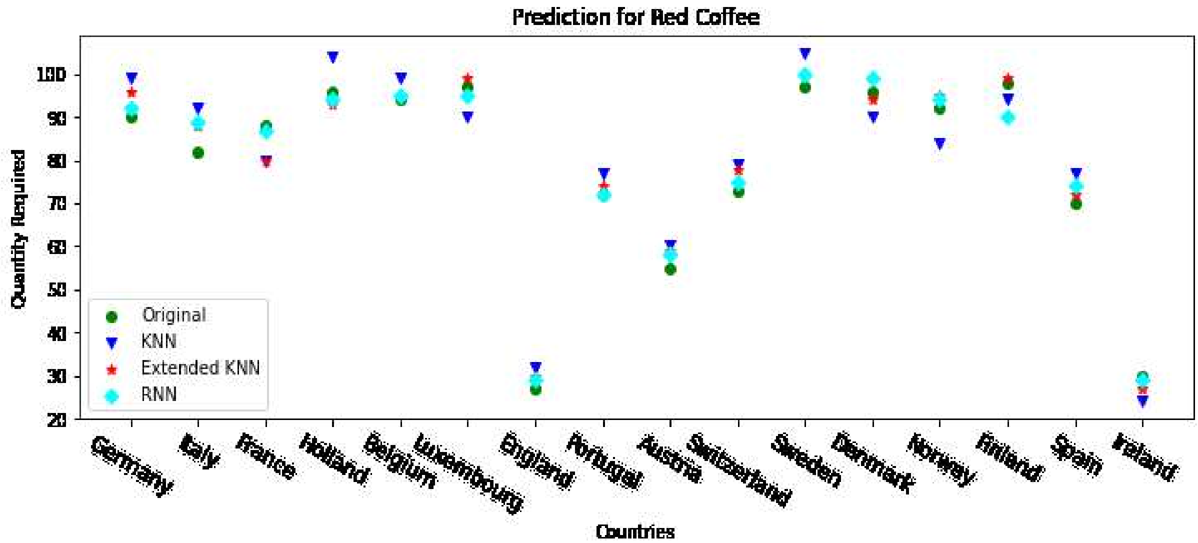


Figure 4.2: Forecasting the consumption of red coffee across various countries

When the factorization rank becomes excessively high, the NMF algorithm produces errors. To address this, we reduce the factorization rank until the NMF algorithm runs without errors. This approach is feasible because we are working with small-sized datasets, a practice applicable across all instances when dealing with FCDBs. As depicted in Figure 4.2, we can showcase the consumption of red coffee among citizens of different countries using various algorithms. Predicting data across diverse countries through different algorithms allows for analysis by comparing results before and after imputing values using KNN, extended KNN, and RNN on the original dataset.

```
Original Values:  [90, 82, 88, 96, 94, 97, 27, 72, 55, 73, 97, 96, 92, 98, 70, 30]
Imputed Values using KNN:  [99, 92, 80, 104, 99, 90, 32, 77, 60, 79, 105, 90, 84, 94, 77, 24]
Imputed Values using KNN Exyended:  [96, 88, 80, 93, 95, 99, 30, 74, 59, 78, 100, 94, 95, 99, 72, 27]
Imputed Values using RNN:  [92, 89, 87, 94, 95, 95, 29, 72, 58, 75, 100, 99, 94, 90, 74, 29]
```

Figure 4.3: Values imputed using KNN, extended KNN, and RNN algorithms.

Figure 4.3 showcases the outcome values post-imputation of missing data in columns from the original dataset, employing KNN, extended KNN, and RNN methods. As illustrated in Figure 4.2, KNN yields superior results compared to alternative methods. These outcomes can also be compared with our proposed extended ANN, a fusion of MICE and ANN techniques.

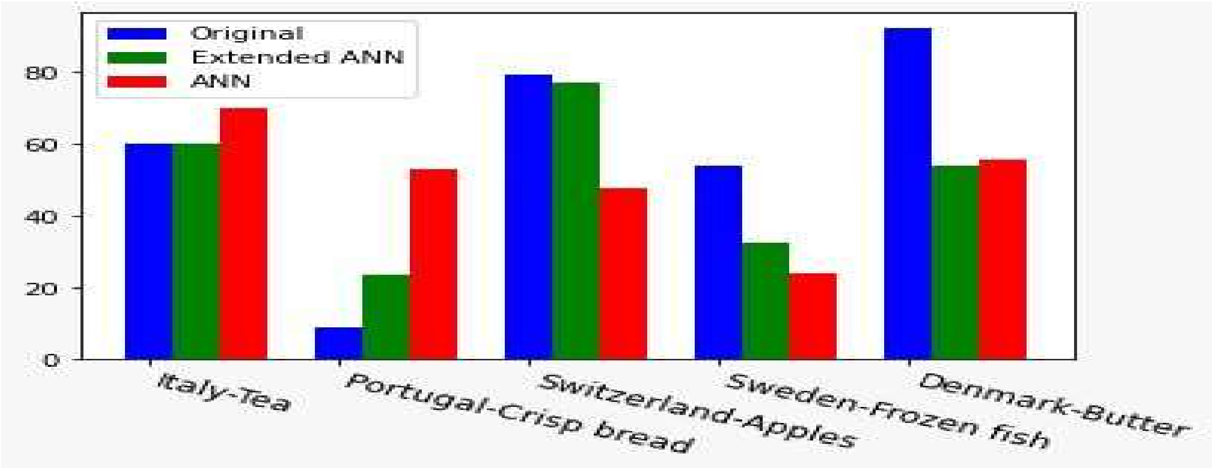


Figure 4.4: Imputation value using original, extended ANN and ANN

Figure 4.4 presents the missing values imputation on the given dataset to determine product consumption over original data with missing values, ANN, and extended ANN. Figure 4 presents better results in the case of extended ANN compared to original and ANN.

4.5 Summary of the Chapter

This thesis introduces multiple machine learning algorithms to handle missing values, offering a comparison with existing approaches. It proposes hybrid schemes combining MICE and ANN, referred to as extended ANN, designed to identify and address missing values within datasets. A thorough evaluation contrasts this approach with recent algorithms, showcasing its efficiency. Simulated results demonstrate the superior performance of the proposed mechanism through graphical representations, including predictions of red coffee consumption among citizens and imputations of missing values regarding food consumption across different countries.

Chapter-5 MVI and Forecast Precision Upgrade of Time Series Precipitation

5.1 Information for Ubiquitous Computing

Introduction

Pervasive or ubiquitous computing indeed aims to seamlessly integrate connectivity and computational capabilities into various objects and environments. This integration allows these objects to communicate, share information, and execute tasks autonomously, minimizing the need for direct human intervention. This vision involves creating environments where technology functions harmoniously in the background, enabling automation and smoother interactions between devices to enhance efficiency and convenience in our daily lives.

Mark Weiser, a prominent computer scientist, played a pivotal role in conceptualizing and popularizing the idea of ubiquitous computing. At Xerox PARC in the late 1980s, Weiser, John Seely Brown, and others began exploring the notion of ubiquitous or pervasive computing. Weiser's vision cantered around seamlessly integrating computing into the environment, making technology almost invisible yet highly functional. He emphasized the idea of computers being everywhere and anywhere, woven into the fabric of everyday life. His contributions laid the foundation for the evolution of this field and continue to influence the development of modern computing paradigms.

Integrating technologies into these settings enhances user experiences through contextual data collection, applications tailored to specific situations, and streamlined payment processes. This integration makes these environments more engaging and enhances their functionality by leveraging the seamless interaction between devices, services, and users. The ultimate goal is to create environments where technology seamlessly integrates into our lives, making tasks more convenient and efficient. The interconnectedness allows for a more holistic and immersive experience across daily life.

The aim is to have computer systems and network technology seamlessly blend into the background of users' lives, operating without demanding constant attention. The key idea is to integrate computers into our daily activities and physical environments so smoothly that they become almost imperceptible, working autonomously and reacting to their surroundings. This invisibility of technology, combined with the ability to detect and respond to their environments autonomously, defines the core principle of ubiquitous computing. It is about enabling technology to assist without intrusive, making interactions more natural and effortless for users.

Pervasive computing transcends the limitations of traditional desktop computing by allowing various devices to connect and operate seamlessly, regardless of location or device type. This means users can access data, applications, and services from any device, anywhere, and anytime, thanks to interconnected networks. The ability to transfer tasks between devices as users move from one location to another, such as from a vehicle to the workplace, is a crucial feature of pervasive computing. The array of devices encompassing laptops, smartphones, tablets, wearables, sensors, and more from the ubiquitous computing ecosystem offers diverse ways for users to interact with technology across different contexts. This flexibility and adaptability represent the fundamental shift from fixed computing environments to a more dynamic and interconnected computing paradigm.

That is an excellent illustration of how ubiquitous computing manifests in autonomous vehicles! Such a system seamlessly integrates various functionalities, such as user identification via smartphone proximity, self-docking and charging capabilities, and efficient handling of tasks like emergency responses and payments through interactions with infrastructure.

This scenario showcases the integration of computing power into everyday objects (in this case, the vehicle) to enable autonomy and intelligence. It involves embedding microprocessors and connectivity within the vehicle to communicate with the environment and infrastructure, ensuring a continuous flow of information for optimal performance. The vehicle becomes a part of the more extensive interconnected network, always available and fully connected, demonstrating the core attributes of ubiquitous computing in action.

Ubiquitous computing significantly emphasizes simplifying computer complexity and improving efficiency to seamlessly integrate technology into everyday life. It builds on the concept of utilizing computing power to enhance everyday activities without the constraints of traditional computing setups.

It is often seen as an evolution beyond mobile computing, incorporating wireless communication, networking technologies, embedded systems, wearable devices, RFID tags, middleware, and intelligent software agents. Integrating these technologies allows for a more holistic and interconnected computing experience, making interactions smoother and more intuitive for users. The overarching goal remains to augment human capabilities by embedding computational power into the fabric of our environments and activities.

Internet connectivity, speech recognition, and artificial intelligence (AI) functionalities often play pivotal roles in ubiquitous computing. These features significantly enhance the capabilities of everyday objects by enabling them to connect to the internet, recognize spoken commands, and employ AI for intelligent decision-making or automation.

By integrating computers into everyday items, ubiquitous computing aims to create a seamless environment where people can effortlessly interact with information-processing devices regardless of their location or context. This integration facilitates more accessible connections and offers more flexibility and freedom in how individuals access and utilize information. It ultimately aims to make technology more intuitive and responsive to human needs, fostering a more natural and efficient interaction between people and their surrounding environment.

Summary

Ubiquitous computing represents a significant shift from the era of large, cumbersome computers to a landscape where each individual, whether a teacher or a student, possesses their own internet-connected, private mobile computing device. This device is versatile, portable, and seamlessly integrated into home and classroom environments, allowing continuous access to information and resources.

The essence of ubiquitous computing lies in the miniaturization and integration of computer technologies into lightweight, handheld devices, making them omnipresent across various emerging contexts. This evolution has transformed computing from stationary and bulky machines to devices that can accompany individuals wherever they go. The ubiquity of these devices allows for a more personalized and flexible computing experience, enabling users to interact with technology in a manner that suits their needs and preferences.

5.2 Missing Value Imputation and Prediction of Rainfall as a Case Study

Rainfall serves as a fundamental component in various aspects of our ecosystem. Its distribution, intensity, and frequency significantly impact agriculture, water resource management, ecology, and climate patterns. Understanding rainfall patterns helps assess potential flood risks, manage water resources efficiently, study climate change effects, and predict agricultural yields. The data gathered from rainfall analysis supports decision-making processes across numerous fields, enabling informed choices for sustainability and environmental protection [1]. The missing data in rainfall datasets poses a prevalent challenge from the different resources [2]. In cases where specific weather stations hold pivotal significance for the study area or are integral to understanding local weather dynamics, alternative strategies for handling missing data become essential. Advanced imputation techniques, such as interpolation methods, statistical modeling, or machine learning algorithms, can help estimate missing values based on neighboring station data, historical patterns, or other relevant variables. These approaches allow for the preservation of critical station data while mitigating the impact of missing values on the overall dataset.

The Indian Administration has conducted several research studies to comprehend the effects of global warming, particularly focusing on shifts in climate and rainfall patterns within India. The India National Disaster Risk Reduction and Management (NDRRMC) framework depends on climate data for delivering weather-related information and services, utilizing tools such as automated weather stations (AWS) for rainfall measurement.

However, it appears that an accurate rainfall prediction model might be limited in estimating the amount of rainfall expected in a specific month and location. To improve this, historical rainfall data collected from various weather stations within the region could be utilized to develop a more robust rainfall prediction model. This model would integrate the data collected from measurement sensors and past campaigns, allowing for a more accurate estimation of expected rainfall in different months and locations.

Developing such a model could significantly enhance the ability to forecast and prepare for potential weather-related risks, aiding disaster risk reduction and better managing weather-related challenges in India.

There is a noticeable trend of increased extreme rainfall events, particularly during the June to September rainy season in some areas of India. While there might not be direct evidence linking annual or periodic rainfall changes to global warming, a growing body of evidence suggests that extreme rainfall could be attributed to global warming. The Intergovernmental Panel on Climate Change (IPCC) analysis indicates that due to global warming, the frequency of intense rainfall events might escalate in the future, specifically over India.

The Indian Monsoon system, despite being relatively stable, poses challenges for statistical models to precisely predict specific data points, especially considering the variations in average rainfall. To address this, neural networks have been employed to predict average rainfall. These networks create multiple functions that assist in predicting data points with enhanced seasonal variations, offering a more nuanced approach to forecasting rainfall patterns.

Visual representation, such as graphical formats showcasing monthly rainfall across different states in India measured in millimeters, can provide a clearer understanding of these variations and aid in analyzing and predicting future rainfall patterns more accurately.

The Indian government has compiled a comprehensive rainfall dataset spanning 115 years, from 1901 to 2015, using around 3000 rain-gauge positions distributed across the country. This dataset, available on platforms like DataWorld and Kaggle, forms the basis for various research studies. This research focuses on analyzing rainfall patterns in specific months within particular states over the years, showcasing this through graphical representations.

Dealing with missing data is crucial for the accuracy and reliability of any analysis. Proper planning and meticulous data collection are recommended to address this issue and minimize missing values in the dataset. The presence of missing data can significantly impact the outcomes of randomized experimental trials if not handled appropriately, as depicted in Figure 1.

Our research centers around the imputation and prediction of missing values based on the available rainfall dataset. By employing techniques for imputing missing values, we aim to enhance the completeness of the dataset and subsequently improve the accuracy of analysis and predictions regarding rainfall patterns. Conducting an analysis to quantify the extent of missing information within the dataset is a critical step toward understanding the dataset's integrity and ensuring the reliability of the research findings.

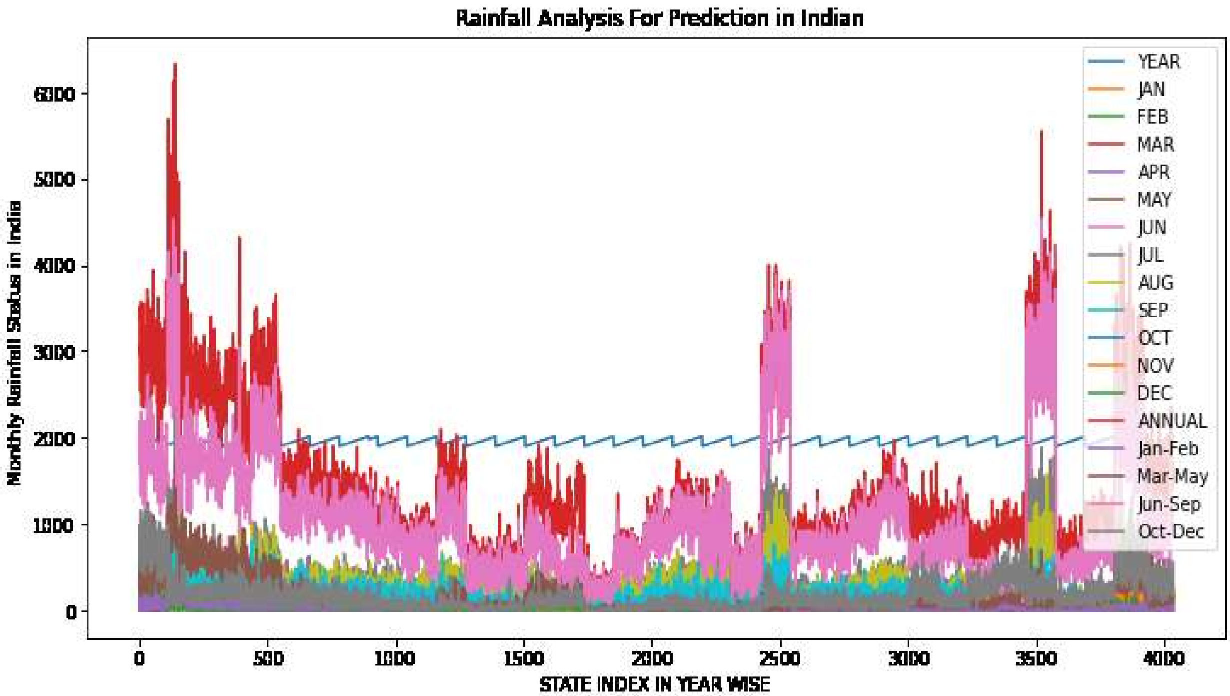


Figure5.1. Monthly Rainfall Status in India

A data analysis technique is described using a heatmap to visualize missing values in a dataset, mainly focusing on monthly missing values in different columns.

Heatmaps are a great way to represent missing data and understand the patterns of incompleteness across variables.

In Figure 5.2, using a heatmap, we can easily identify which months or columns have more missing values than others. The color intensity or shading in the heatmap typically represents the level of missingness, making it visually apparent which areas of the dataset lack information.

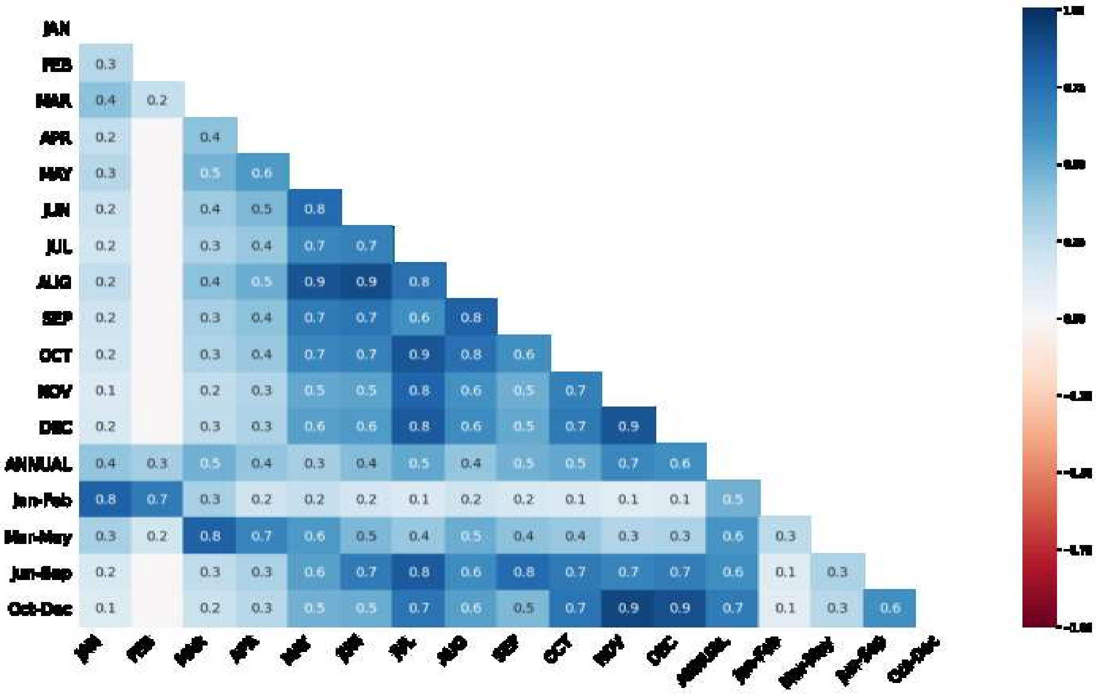


Figure 5.2. Display Missing Values in India Rainfall Dataset

This kind of visualization helps understand the overall structure of missing data, allowing for informed decisions on handling or imputing missing values in the dataset. Sorting rows or columns in a way that showcases patterns in missing data can reveal insights that might not be immediately evident in the raw dataset.

This study examines the drawbacks of prior investigations into rainfall patterns, highlighting their reliance on historical data spanning 115 years, which may overlook recent statistics and contemporary trends. In contrast, this study prioritizes a more current timeframe, spanning 35 years from 1981 to 2015.

Concentrating on this more contemporary dataset, this study aims to capture and analyze experiential rainfall patterns, trends, and variations. This updated information can provide valuable insights into the current state of rainfall and its potential implications for climate change adaptation.

5.3 Proposed Mechanism to Handle a Mission Value

This research methodology uses daily rainfall records from this period to derive month-to-month rainfall series, ultimately constructing monthly rainfall collections by averaging the rainfall values for each month across Bihar. Additionally, we have opted to compute the month-to-month rainfall series by incorporating region-weighted rainfall values from all districts within Bihar.

The decision to focus on a single state, Bihar, allows for a more localized and detailed analysis of rainfall patterns, potentially providing actionable information for state-level institutions involved in climate change adaptation and management.

Missing value imputation is crucial in data preprocessing, particularly in predictive modeling tasks like forecasting rainfall. When dealing with meteorological data, missing values are common due to factors like sensor errors, equipment malfunction, or natural conditions.

For rainfall prediction, imputation methods become essential to ensure a comprehensive dataset for accurate modeling. Here is a generalized approach using a case study:

a) Understanding the Data:

Start by examining the dataset, identifying missing values, and understanding their distribution across the variables, especially the rainfall data.

b) Exploratory Data Analysis (EDA):

Perform EDA to understand the patterns and relationships in the available data. This helps identify the nature of missingness, correlations, and potential predictors.

c) Missing Value Imputation:

In this section, we initially delve into the methodology of imputation, exploring well-established techniques and their application to the Rainfall dataset. Among these approaches are the univariate Kalman filter and the Extended Kalman filter, also called the Kalman smoother. Specifically, we employ two models within the Kalman filter framework, namely the representation of the ARIMA model and StructTS model. This technique is often considered adequate for imputing data in highly seasonal univariate datasets.

This investigation explores the Kalman filter approach in addressing reservoir sampling and histogram-based methodologies. Our findings showcase that the Extend Kalman and Kalman filters excel in imputing missing values, displaying the least pull mean squared error in most cases. We introduce a novel rationale for integrating this estimation process, explicitly addressing irregular missing values within sensor and device streams.

To tackle this issue, we propose a technique based on the Kalman filter and Extend Kalman filter. We treat data sensor streams and instrument streams as time series and utilize the Extend Kalman filter to predict and impute missing values. This thesis focuses on univariate time series, encompassing a single recorded perception over equivalent time intervals. However, it is worth noting that the methodology outlined here can be extended to multivariate time series using specialized techniques.

The Extend Kalman filter, a recursive and numerical evaluation algorithm, excels at data assimilation and predicting missing values. It leverages historical data to estimate the current values of the variables of interest. Employing a Kalman filter in various state techniques, such as the dynamic linear technique, presents advantages like simplicity, requiring certain initializations, and dynamically updating its state.

The dynamic linear technique involves differences in proceeding values over time. When coupled with this technique, the extended Kalman filter enables precise one-pass forecasts, continually predicting and updating as new data streams in. For further clarity, we provide the implemented algorithm of the Extend Kalman filter used for missing values imputation.

Proposed Algorithm:

Algorithm: **Proposed Algorithm**

Input: Number of rows containing missing values in a dataset.

Output: Whether the missing values are imputed using the Extended-KF algorithm

```
BEGIN
Extended_KF→ []                               //Initialize empty array
data2 = data                                  //data is the value of the dataset before applying imputation
na_mean(x)                                    // Impute the missing values
data1 → [0]                                   // initiaze matrix with zero value.
diff_mat = data2 – data1                      //take difference of data2 and data1
mean2 = average(diff_mat)                     //mean2 is absolute average of the matrix diff_mat
mean1 → INT_MAX                               //initialize mean1 as max integer possible
for (mean2 <= mean1):                         //iterate until error is reducing.
mean1 = mean2                                 //make mean1 as the previous value for the next iteration
data2 = data1;                                // make data2 as previous value for the next iteration
for i → Missing:                              // iterate for every missing value
imputed_value→ imp <- na_kalman(tsAirgap)     // time series provided by the imputeTS package
equation applied is Y = h(X')                 // The function h delineates the mapping of our location
data1[index][index2] = imputed_value          // update the current value
diff_mat→ data2-data1                         // updating difference matrix
mean2 = diff_mat.mean()                       //updating current value
for i in range(len(data) - seq_length)        //Create input-output sequences
LSTM(units=50, input_shape=(seq_length, 1))   //Model Building
model.compile(optimizer='adam', loss='mean_squared_error') // Optimizer and Calculate evaluation metrics  RMSE
END
```

Flow Chart of Proposed Algorithm:

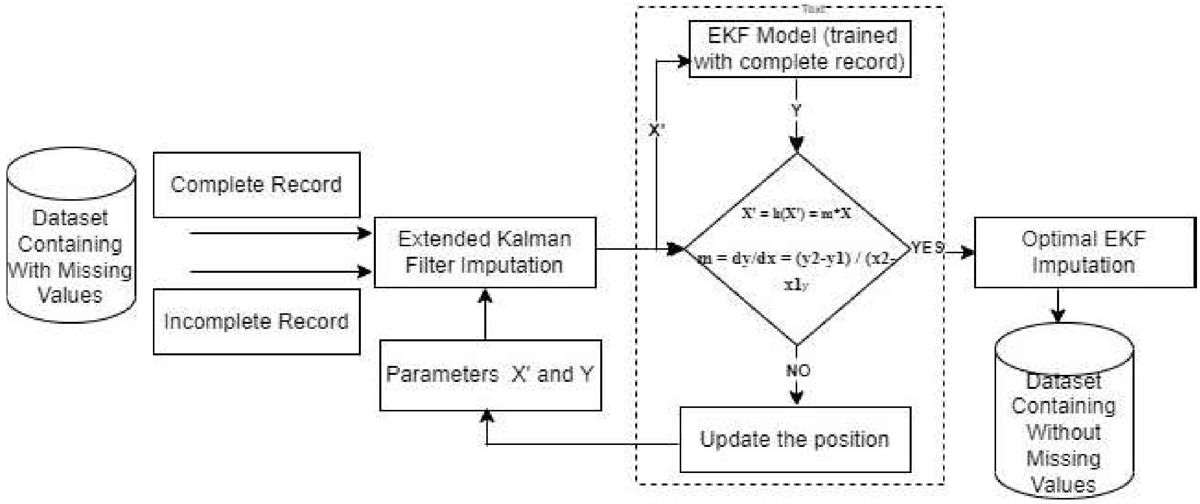


Figure 5.3 Process of Extended Kalman for Missing Value Imputation

Figure 5.3 illustrates the flow diagram detailing the algorithm's process. This program starts with a dataset containing missing values and implements the Extended Kalman filter. Two parameters, X' and Y, are utilized within this process, and the equation applied is Y = h(X'). The function h delineates the mapping of our location to polar coordinates, where x′ represents predicted values, while y denotes the difference between the measured and actual values.

This mapping function aims to define how our predicted values, initially in Cartesian coordinates, are translated into Polar coordinates. This translation is crucial since our predictions are made in Cartesian coordinates, while the sensor's measurements are provided in Polar coordinates. Therefore, this mapping facilitates the conversion between these coordinate systems, ensuring compatibility between predicted values and sensor measurements.

The primary role of the Kalman filter is to determine optimal estimates, operating under the assumption of normality. The Kalman filter calculates the conditional mean and modifies the distribution based on observations up to a specific time.

In this research, an adaptive structure is proposed to enhance the Extended Kalman Filter (EKF). This improved EKF takes derivative results that are close to each other and multiplies them with the rate of change in the extended Kalman filter. This adaptation aims to refine the estimation process by incorporating derivative information and optimizing the filter's performance and adaptability.

5.3.1 Feature Engineering

Engineer relevant features that might influence rainfall, such as temperature, humidity, wind speed, geographical location, etc. Consider lag features or moving averages of rainfall to capture temporal patterns.

a) Modeling and Prediction:

Utilize machine learning models (e.g., regression, time series models like ARIMA and LSTM for deep learning) to predict rainfall based on the available features.

Split the dataset into training and testing sets, considering the temporal aspect.

Evaluate the model performance using appropriate metrics (RMSE, MAE, etc.) on the test set to assess the accuracy of rainfall predictions.

b) Validation and Iteration:

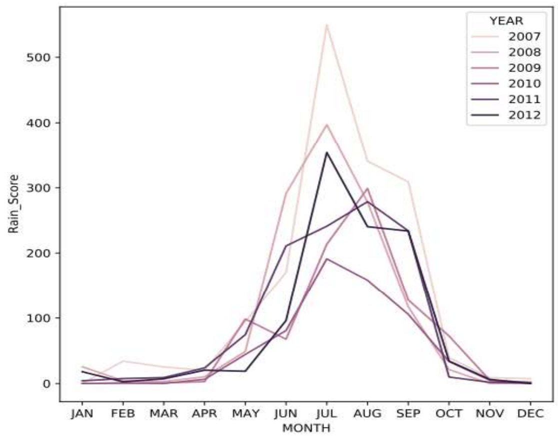
Validate the model's performance over different periods or geographical regions if applicable.

Iterate through the process by refining feature selection, trying different imputation methods, or tuning model parameters to improve accuracy.

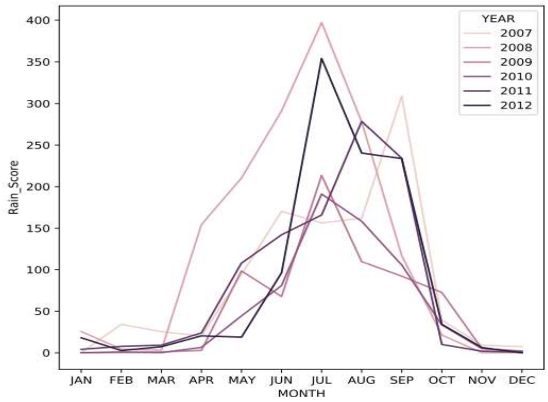
Remember, the choice of imputation method and predictive model may vary based on the dataset characteristics, domain knowledge, and the specific requirements of the rainfall prediction task.

5.4 Result Analysis of Proposed Mechanism

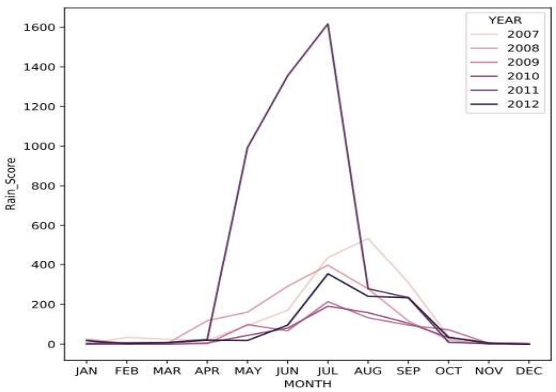
This section delineates several graphs employed to assess the proposed mechanism's superior performance compared to various existing approaches. Figure 5.4 specifically represents six years (2007 to 2012) of observed rainfall data in Bihar state. The X-axis portrays the measured rainfall amount in millimeters at different locations, while the Y-axis illustrates the corresponding monthly rainfall produced.



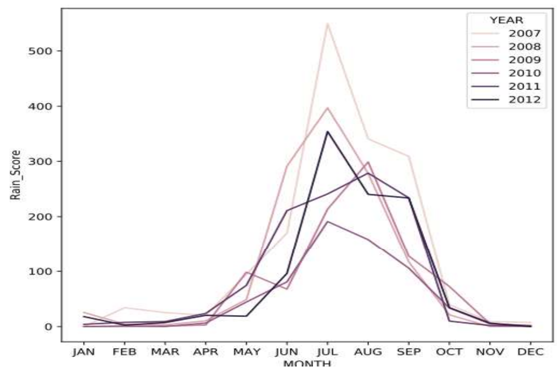
(a)



(b)



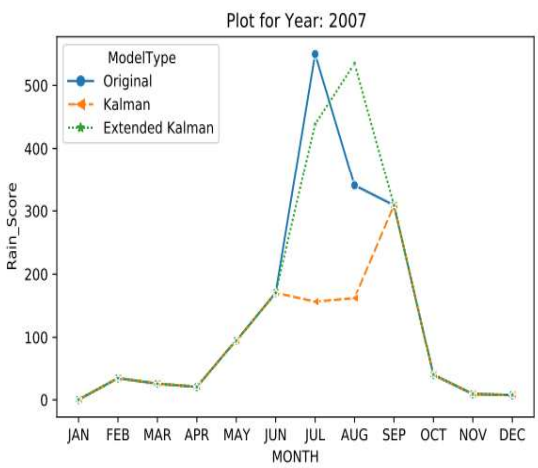
(c)



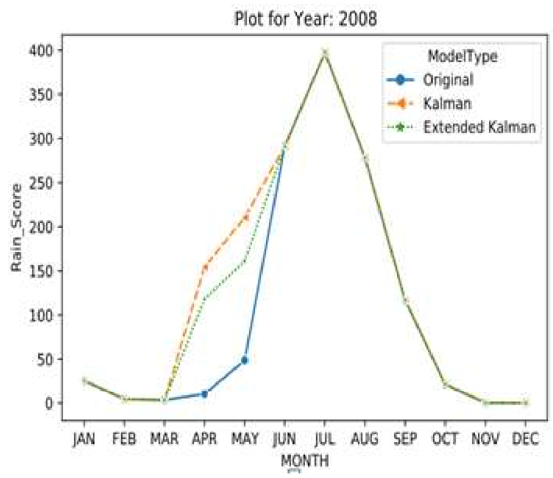
(d)

Figure 5.4 illustrates the variations in rainfall amounts across diverse locations over six years spanning from 2007 to 2012. Each location depicted in Figures a, b, c, and d exhibits distinct rainfall patterns, while all other parameters remain consistent.

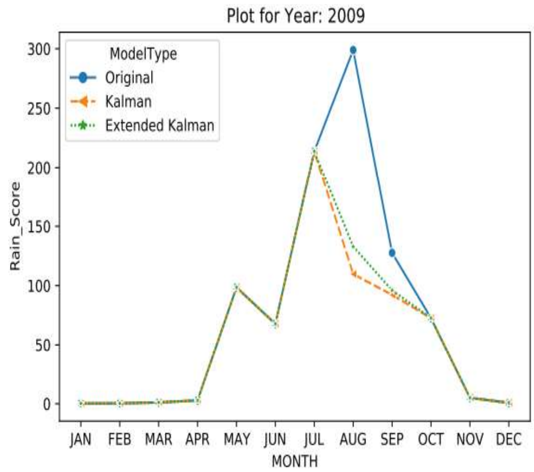
Figure 5.5 presents a graphical comparison of the result analysis showcasing the missing values imputation using two approaches with original values, the Kalman filter and our proposed extended Kalman filter algorithm. Figure 5.5 presents a comparative analysis, year by year, of missing values imputation using original values alongside the Kalman filter and extended Kalman filter techniques. This analysis spans from 2007 (denoted as A) to 2012 (denoted as F) in Bihar state, examining various locations. The imputed results using original values are depicted in blue. Predicted values generated by the Kalman filter are represented in orange, while those from the extended Kalman filter are displayed in green.



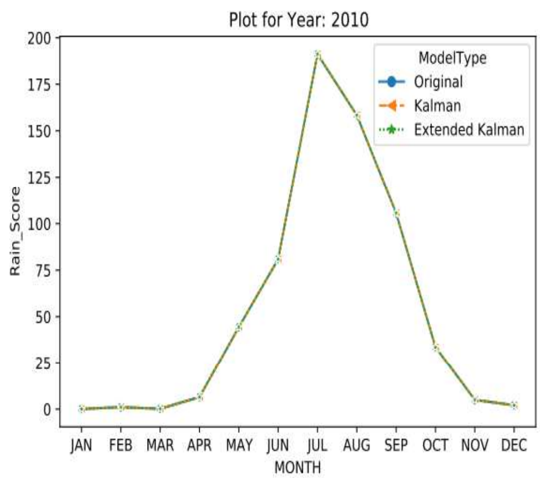
(a)



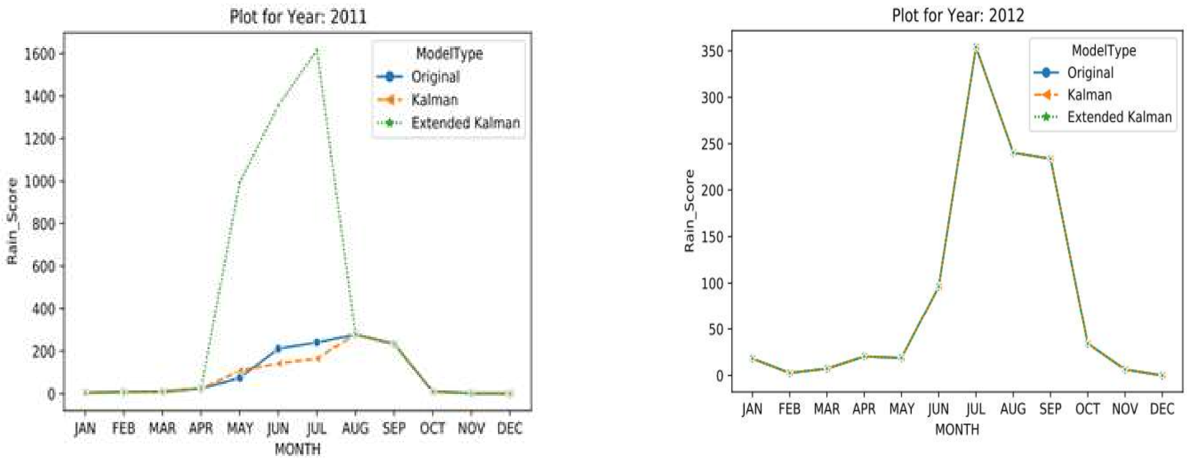
(b)



(c)



(d)



(e) (f)

Figure 5.5 presents a comparative analysis, year by year, of missing values imputation using original values alongside the Kalman filter and extended Kalman filter techniques (a) the focus is on the year 2007; (b) it shifts to 2008; (c) centers on 2009; (d) highlights 2010; (e) centers around 2011, while (f) is directed towards 2012. This analysis spans from 2007 to 2012 in Bihar state, examining various locations.

In Figure 5.6, researchers display the comparative result of the proposed algorithm's predicted value with the original value and find that the accuracy of predicted values is not good. At that stage, optimization techniques are required.

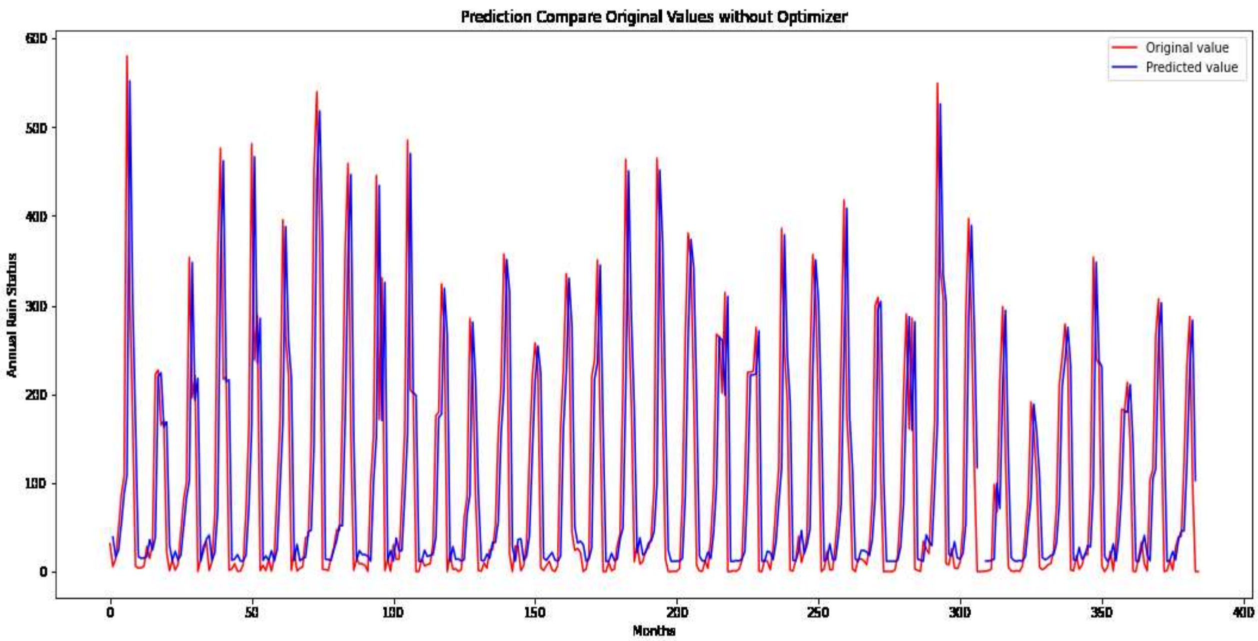


Figure 5.6. Predicted values compare with the original values

Here, we are using a stochastic gradient descent (SGD) optimizer. To enhance results, our proposed algorithm leverages the SGD optimizer to refine the achieved outcomes.

Moreover, in Figure 5.7, SGD addresses the Gradient Descent problem by employing individual records for structure updates. However, despite this improvement, SGD's convergence results remain slow. Its iterative nature, involving forward and backward propagation for each record, results in a noisy path toward reaching the global minima, hindering swift convergence.

Gradient Descent is a primary optimization method in machine learning and deep learning, and it is widely applicable across various learning procedures. It operates by utilizing the gradient, which represents the slope of a function, to measure how a variable changes relative to changes in another variable. Mathematically, Gradient Descent involves navigating a curved function to iteratively adjust a set of parameters to minimize the function's output.

If a researcher fails to achieve the desired results using the stochastic gradient descent (SGD) optimizer, an alternative approach involves substituting it with the RSMProp optimizer. Upon execution, this change yields superior results compared to the SGD optimizer. The subsequent graph depicts a comparison between the original values, demonstrating that the accuracy achieved with the new optimizer surpasses that of the previous one.

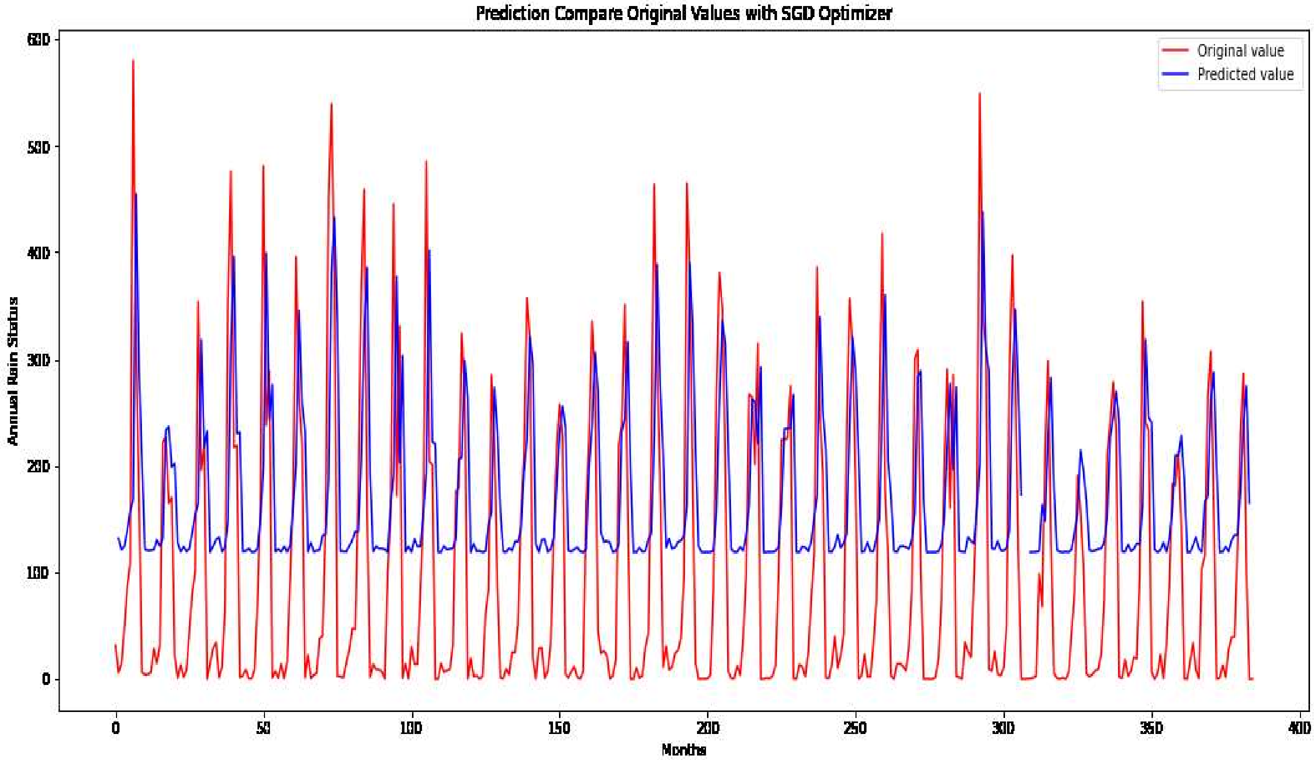


Figure 5.7. Predicted values compare with original values using SGD optimizer

The RMSprop optimizer, illustrated in Figure 5.8, bears similarities to the gradient descent procedure with momentum. Unlike standard gradient descent, RMSprop curtails fluctuations in a perpendicular direction. This increases the learning rate, facilitating more substantial strides in the horizontal direction and faster convergence. The critical distinction between RMSprop and gradient descent lies in how gradients are computed. The subsequent computation outlines the gradient computation processes for RMSprop and gradient descent with momentum. The momentum value, denoted by beta, is frequently used. For those less interested in the intricacies behind the optimizer, feel free to skip this technical detail.

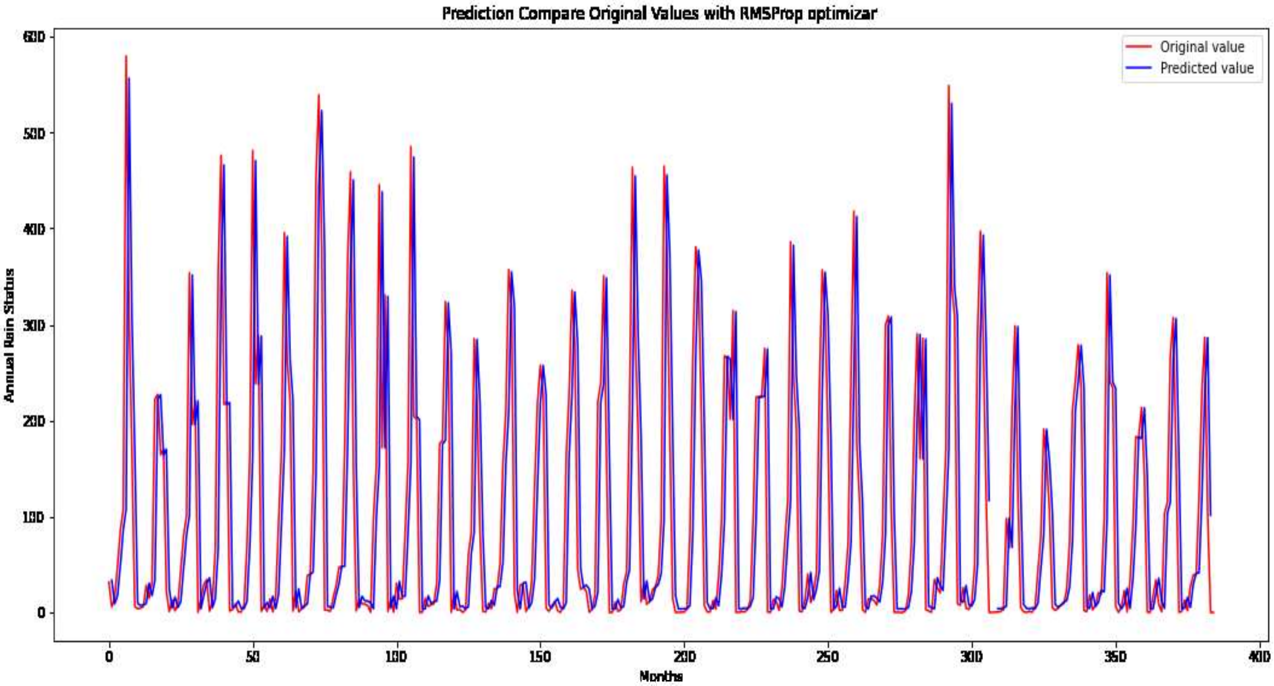


Figure 5.8. Predicted values compare with original values using RMSprop optimizer

The RMSprop optimizer is a gradient-based optimization technique primarily utilized in training neural networks. Initially proposed by Geoffrey Hinton, a pioneer in back-propagation, gradients within complex functions like neural networks tend to either vanish or explode as data traverses through the network.

RMSprop was designed as a solution for this issue within mini-batch learning scenarios. It tackles the problem by employing a moving average of squared gradients, effectively normalizing the gradient. This regularization process adjusts the step size, reducing it for large gradients to prevent explosion and increasing it for small gradients to prevent vanishing.

Moreover, Figure 5.9 showcases the comparison between the precise outcomes of our proposed algorithm and the performance achieved by the RSMProp optimizer. Upon further optimization attempts, employing the ADAM optimizer yielded significantly superior results compared to all prior optimization methods.

Integrating the ADAM optimizer with our proposed algorithm led to predicted values aligning closely with actual values, depicted graphically compared to the original values. This amalgamation showcased notably improved accuracy over the previous optimization techniques.

The ADAM optimization technique enhances stochastic gradient descent, offering more efficient parameter updates. It computes individual learning rates for each parameter, a method developed by its creators to perform effectively in practical applications and demonstrate favorable evaluations compared to other adaptive learning algorithms.

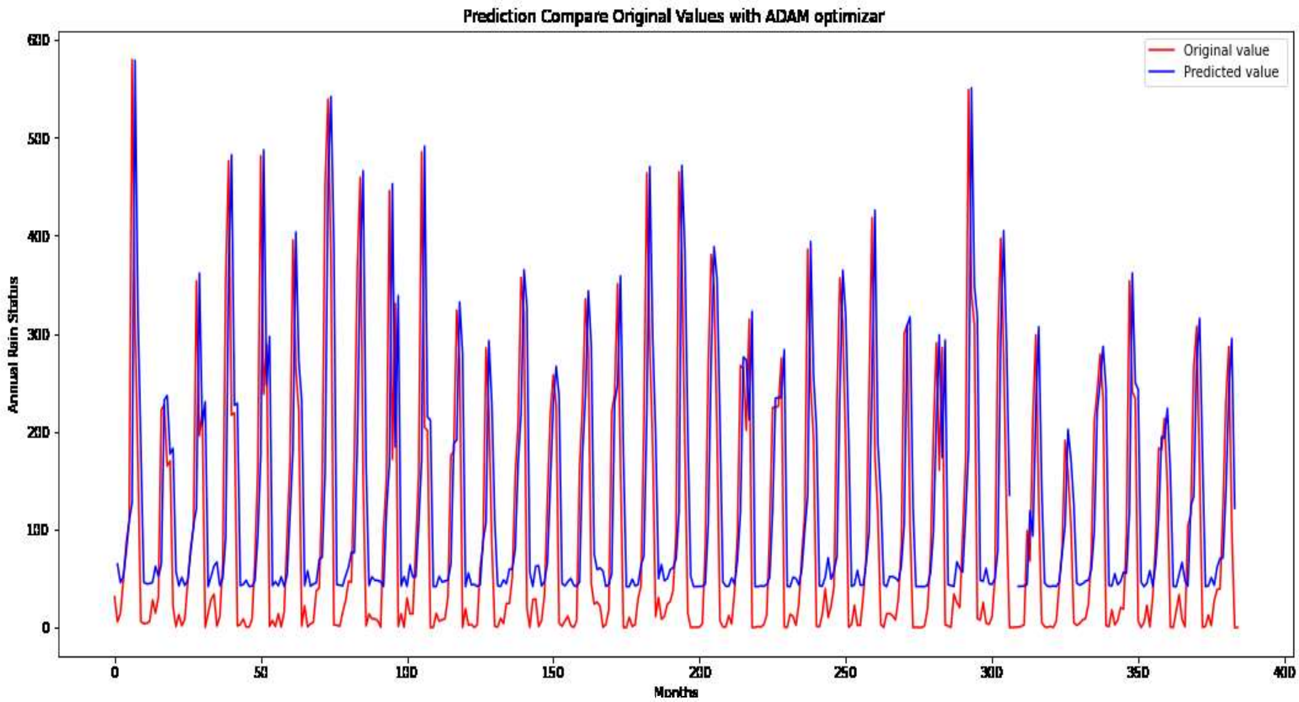


Figure 5.9. Predicted values compare with original values using ADAM optimizer

In this figure, our research focuses on missing data imputation through arithmetic prediction methods, encompassing four distinct prediction approaches. One is our proposed algorithm, while the other three utilize optimization methods. The comparison is made against the original values, showcased in Figure 5.10. This study aims to conduct a comprehensive comparative analysis, evaluating for the first time the performance of SGD optimizer, RSMProp optimizer, and ADAM optimizer in addition to our proposed algorithm for missing data imputation. The figure provides a graphical representation depicting the performance of these optimizers alongside our proposed algorithm concerning the imputed values compared to the original dataset. Use optimization algorithms like stochastic gradient descent (SGD), RSMProp optimizer and Adam to minimize the loss function.

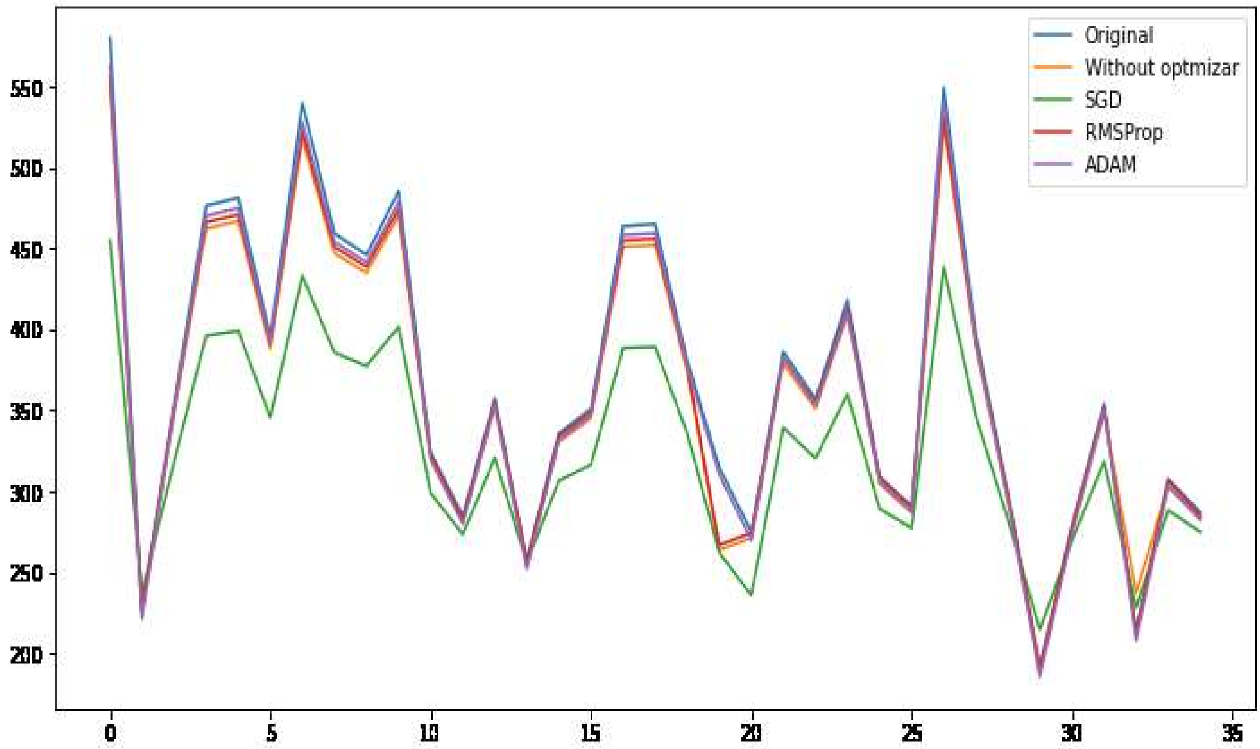


Figure 5.10. Predicted values compare with the original values

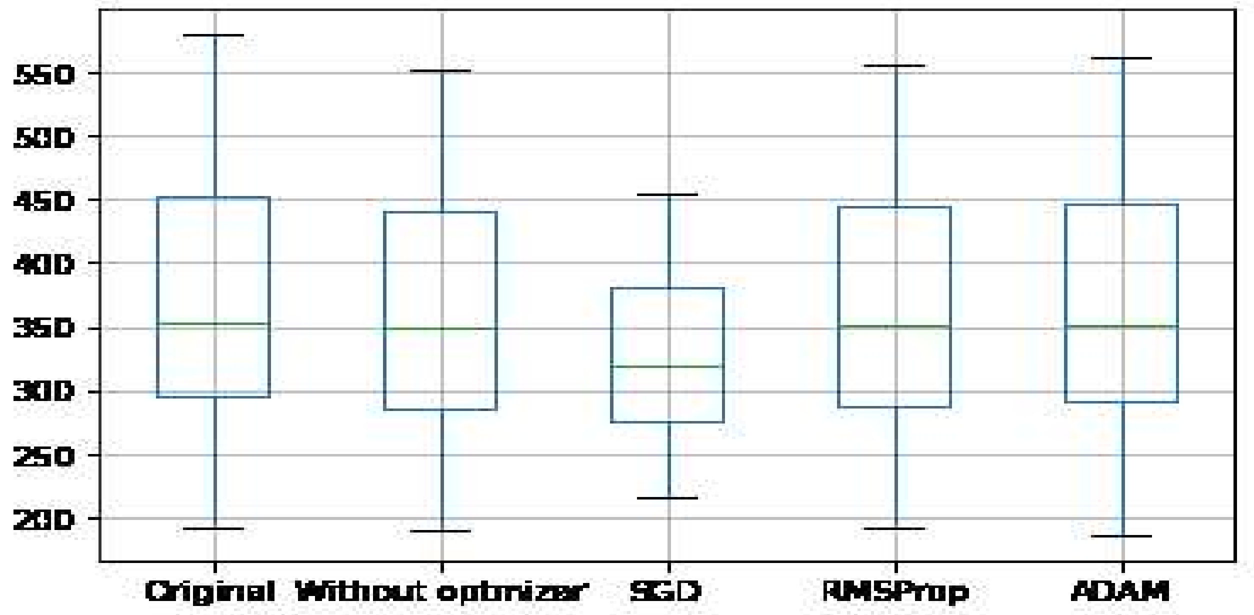


Figure 5.11. Predicted ADAM Optimizer values compare with the original values

Furthermore, Figure 5.11 illustrates the accuracy achieved across all scenarios examined in our research. The box plot graph presents the best results obtained by different optimizers with algorithms, notably showcasing the ADAM optimizer achieving results closely aligned with the actual values. This graph serves as the conclusive depiction of the accuracy attained through the various optimization techniques employed in our study.

5.5 Summary of the Chapter

This thesis introduces a novel approach, employing extended Kalman filter methods for missing values imputation. This method utilizes linear relations in the imputation process. Assessing the predictive capabilities of LSTM-based models holds promise for advancing research in deep learning methodologies tailored for addressing rainfall prediction challenges in ubiquitous computing scenarios. Our study proposes a comprehensive bidirectional and unidirectional LSTM architecture tailored for network-wide rainfall forecasting. Also, the optimization technique should be used to improve accuracy.

Furthermore, evaluating the forecasting performance of LSTM-based models presents numerous opportunities to refine neural network strategies for accurate rainfall prediction, ensuring efficacy in ubiquitous computing applications. The research utilizes practical rainfall data from Bihar State on the specified date. Experiments conducted on a real-world dataset with various missing values demonstrate that the proposed architecture achieves exceptional results in imputation and prediction tasks.

Chapter-6 Conclusion

In conclusion, data mining and analysis are indispensable tools in numerous real-world applications, relying on databases tailored to specific needs. Handling challenges like missing values is crucial for accurate results and effective decision-making systems. This thesis delves into the taxonomy of missing data types and proposes structured handling methods to mitigate their impact. Statistical testing and substantive knowledge aid in identifying missing data mechanisms, and guiding appropriate handling strategies. Techniques like mean imputation and advanced methods including deep learning algorithms enhance data completeness and analysis accuracy. In ubiquitous computing, precise forecast models heavily rely on effective missing data imputation techniques such as machine learning and deep learning approaches, ensuring continuity and reliability. Leveraging these methods enhances decision-making across diverse domains, emphasizing the importance of robust data-handling strategies in maximizing the utility of datasets for practical applications.

In this thesis, we have suggested some improved methods of missing value imputation and introduced a range of machine learning algorithms designed to address missing data imputation, contrasting them with existing methods. It proposes a novel hybrid approach, combining Multiple Imputation by Chained Equations (MICE) with Artificial Neural Networks (ANN) into an extended ANN model. This hybrid scheme is devised to identify and address missing values within datasets. To evaluate its efficacy, the proposed mechanism is benchmarked against contemporary algorithms. Simulated results demonstrate the superior performance of the proposed method, showcased through graphical analyses. Notably, the model excels in scenarios such as predicting red coffee consumption across diverse demographics within a nation and imputing missing values about food consumption across various countries. Through these comparisons, this thesis underscores the effectiveness of the extended ANN model in addressing missing data challenges, offering a promising avenue for data imputation in diverse contexts.

In this thesis, we embark on a pioneering journey employing the extended Kalman methodology for the imputation of missing values—a novel approach in the realm of data completion. Leveraging linear relationships, we aim to fortify the imputation process, enriching its efficacy and reliability.

Furthermore, the evaluation of LSTM-based models holds promise in expediting advancements in deep learning methodologies tailored for addressing the challenges in rainfall prediction within the context of ubiquitous computing. Our proposal introduces a sophisticated architecture, integrating both loaded bidirectional and unidirectional LSTM networks, tailored for comprehensive rainfall forecasting across network domains.

Moreover, the appraisal of LSTM-based forecasting models presents myriad avenues for refining neural network strategies geared towards precipitation prediction, thus fostering optimal performance in ubiquitous computing scenarios. Notably, we conduct empirical analyses using real-world rainfall data sourced from Bihar State, ensuring the relevance and applicability of our research findings.

Our experiments, conducted on a genuine dataset featuring varying degrees of missing data, underscore the efficacy of the proposed architecture, showcasing its prowess in both imputation and prediction tasks

In summary, while the Extended Kalman Filter offers a flexible framework for handling non-linear systems, its application to missing values imputation is not without limitations. Careful consideration of the assumptions, initialization procedures, and computational challenges is necessary when using the EKF for imputing missing data.

6.2 Future Directions

In the current landscape, there arises a pressing need for missing data imputation, a practice gaining significant traction in recent years, particularly in the domain of time sensor data management. Consequently, this research endeavor necessitates the exploration of novel approaches to address this imperative.

Firstly, an exhaustive comparative analysis is essential, juxtaposing various imputation techniques with deep learning models to ascertain their efficacy and applicability in diverse scenarios.

Secondly, there is a call for the development of a monitoring and warning system aimed at mitigating sensor malfunctions, thereby ensuring the integrity and reliability of the data collected.

Lastly, there is a critical need for the implementation of a real-time Missing Value Imputation (MVI) technique within a comprehensive predictive framework. Such an approach will empower real-time data-driven decision-making strategies, particularly crucial for optimizing the energy-efficient operations of marine machinery.

These proposed avenues for future research not only address the current gaps but also pave the way for enhanced efficiency and reliability in time series data management practices.

List of Publications

Journal

1. Tripathi, Ashok Kumar, Hemraj Saini, and Geetanjali Rathee. "Futuristic prediction of missing value imputation methods using extended ANN." *International Journal of Business Analytics (IJBAN),* vol.9, no.3, 2022, pp.1-12.
2. Tripathi, Ashok Kumar, P.K. Gupta, Hemraj Saini, and Geetanjali Rathee "MVI and Forecast Precision Upgrade of Time Series Precipitation Information for Ubiquitous Computing." *Informatica,* vol.47, no.5, 2023, pp.83-94.

Conferences

1. Tripathi, Ashok Kumar, Geetanjali Rathee, and Hemraj Saini. "Taxonomy of missing data along with their handling methods." *2019 Fifth International Conference on Image Information Processing (ICIIP),* pp. 463-468, IEEE, 2019.

2. Tripathi, Ashok Kumar, Hemraj Saini, and Geetanjali Rathee. "Missing Values Imputation in Food Consumption: An Analytical Study." *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC),* pp.303-307, IEEE, 2021.

References

[1] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

[2] Zhu, Jinlin, et al. "Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data." *Annual Reviews in Control* 46 (2018): 107-133.

[3] Olaniyi, Oluwaseun, et al. "Harnessing predictive analytics for strategic foresight: a comprehensive review of techniques and applications in transforming raw data to actionable insights." *Available at SSRN 4635189* (2023).

[4] Plotnikova, Veronika, Marlon Dumas, and Fredrik Milani. "Adaptations of data mining methodologies: a systematic literature review." *PeerJ Computer Science* 6 (2020): e267.

[5] Gul, Sumeer, Shohar Bano, and Taseen Shah. "Exploring data mining: facets and emerging trends." *Digital Library Perspectives* 37.4 (2021): 429-448.

[6] Wu, Wen-Tao, et al. "Data mining in clinical big data: the frequently used databases, steps, and methodological models." *Military Medical Research* 8 (2021): 1-12.

[7] Gupta, Manoj Kumar, and Pravin Chandra. "A comprehensive survey of data mining." *International Journal of Information Technology* 12.4 (2020): 1243-1257.

[8] Wu, Aoyu, et al. "Ai4vis: Survey on artificial intelligence approaches for data visualization." *IEEE Transactions on Visualization and Computer Graphics* 28.12 (2021): 5049-5070.

[9] Gupta, Manoj Kumar, and Pravin Chandra. "A comprehensive survey of data mining." *International Journal of Information Technology* 12.4 (2020): 1243-1257.

[10] Shmueli, Galit, et al. *Machine learning for business analytics: Concepts, techniques, and applications with analytic solver data mining*. John Wiley & Sons, 2023.

[11] Onikoyi, Babatunde, Nonso Nnamoko, and Ioannis Korkontzelos. "Gender prediction with descriptive textual data using a Machine Learning approach." *Natural Language Processing Journal* 4 (2023): 100018.

[12] Gorokhovatskyi, Volodymyr, et al. "Search for visual objects by request in the form of a cluster representation for the structural image description." *Advances in Electrical and Electronic Engineering* 21.1 (2023): 19-27.

[13] Pynadath, Minnu F., T. M. Rofin, and Sam Thomas. "Evolution of customer relationship management to data mining-based customer relationship management: a scientometric analysis." *Quality & Quantity* 57.4 (2023): 3241-3272.

[14] Chang, Qingqing, and Jincheng Hu. "Research and Application of the Data Mining Technology in Economic Intelligence System." *Computational Intelligence and Neuroscience* 2022 (2022).

[15] Delen, Dursun. *Predictive Analytics Pearson uCertify Course and Labs Access Code Card: Data Mining, Machine Learning and Data Science for Practitioners*. FT Press, 2020.

[16] Liu, Quansheng, et al. "Prediction model of rock mass class using classification and regression tree integrated AdaBoost algorithm based on TBM driving data." *Tunnelling and Underground Space Technology* 106 (2020): 103595.
[17] Muhammad, L. J., et al. "Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery." *SN computer science* 1.4 (2020): 206.
[18] Block, Sharon, and David Newman. "What, where, when, and sometimes why: Data mining two decades of women's history abstracts." *Journal of Women's History* 23.1 (2011): 81-109.
[19] He, Jing. "Advances in data mining: History and future." *2009 Third International Symposium on Intelligent Information Technology Application*. Vol. 1. IEEE, 2009. 20-Mann, Jatinder. "The introduction of multiculturalism in C anada and A ustralia, 1960s–1970s." *Nations and Nationalism* 18.3 (2012): 483-503.
[20] Fleck, James. "Development and establishment in artificial intelligence." *The Question of Artificial Intelligence*. Routledge, 2018. 106-164.
[21] Newell, Allen. "Intellectual issues in the history of artificial intelligence." *Artificial Intelligence: Critical Concepts* (1982): 25-70.
[22] Fayyad, Usama. "Knowledge discovery in databases: An overview." *International Conference on Inductive Logic Programming*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997.
[23] Piatetsky-Shapiro, Gregory, et al. "Kdd-93: Progress and challenges in knowledge discovery in databases." *AI magazine* 15.3 (1994): 77-77.
[24] Kroese, Dirk P., Zdravko Botev, and Thomas Taimre. *Data science and machine learning: mathematical and statistical methods*. Chapman and Hall/CRC, 2019.
[25] Liao, Shu-Hsien, Pei-Hui Chu, and Pei-Yuan Hsiao. "Data mining techniques and applications–A decade review from 2000 to 2011." *Expert systems with applications* 39.12 (2012): 11303-11311.
[26] Shu, Xiaoling, and Yiwan Ye. "Knowledge Discovery: Methods from data mining and machine learning." *Social Science Research* 110 (2023): 102817.
[27] Ara, Anjuman, et al. "The Impact Of Machine Learning On Prescriptive Analytics For Optimized Business Decision-Making." *International Journal of Management Information Systems and Data Science* 1.1 (2024): 7-18.
[28] Guleria, Pratiyush, and Manu Sood. "Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling." *Education and Information Technologies* 28.1 (2023): 1081-1116.
[29] Reddy, R. Pallavi, Ch Mandakini, and Ch Radhika. "A Review on Data Mining Techniques and Challenges in Medical Field." *International Journal of Engineering Research and Technology* 9 (2020): 329-333.
[30] Palanivinayagam, Ashokkumar, and Robertas Damaševičius. "Effective handling of missing values in datasets for classification using machine learning methods." *Information* 14.2 (2023): 92.

[31] Rashid, Wajeeha, and Manoj Kumar Gupta. "A perspective of missing value imputation approaches." *Advances in Computational Intelligence and Communication Technology: Proceedings of CICT 2019*. Springer Singapore, 2021.

[32] Fuchs, Matthias, and Wolfram Höpken. "Clustering: Hierarchical, k-Means, DBSCAN." *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*. Cham: Springer International Publishing, 2022. 129-149.

[33] Ghosal, Attri, et al. "A short review on different clustering techniques and their applications." *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018* (2020): 69-83.

[34] Maulud, Dastan, and Adnan M. Abdulazeez. "A review on linear regression comprehensive in machine learning." *Journal of Applied Science and Technology Trends* 1.2 (2020): 140-147.

[35] Loh, Wei-Yin. "Logistic regression tree analysis." *Springer handbook of engineering statistics*. London: Springer London, 2023. 593-604.

[36] Li, Jinbo, et al. "Clustering-based anomaly detection in multivariate time series data." *Applied Soft Computing* 100 (2021): 106919.

[37] Thudumu, Srikanth, et al. "A comprehensive survey of anomaly detection techniques for high dimensional big data." *Journal of Big Data* 7 (2020): 1-30.

[38] Albalawi, Rania, Tet Hin Yeap, and Morad Benyoucef. "Using topic modeling methods for short-text data: A comparative analysis." *Frontiers in artificial intelligence* 3 (2020): 42.

[39] Zong, Chengqing, Rui Xia, and Jiajun Zhang. *Text data mining*. Vol. 711. Singapore: Springer, 2021.

[40] Shrivastava, Anshu, J. Jaın, and Dipti Chauhan. "Literature review on tools & applications of data mining." *International Journal of Computer Sciences and Engineering* 11.4 (2023): 46-54.

[41] Costa, Vinícius G., and Carlos E. Pedreira. "Recent advances in decision trees: An updated survey." *Artificial Intelligence Review* 56.5 (2023): 4765-4800.

[42] Blockeel, Hendrik, et al. "Decision trees: from efficient prediction to responsible AI." *Frontiers in Artificial Intelligence* 6 (2023).

[43] Hassanien, Aboul Ella, Ashraf Darwish, and Sara Abdelghafar. "Machine learning in telemetry data mining of space mission: basics, challenging and future directions." *Artificial Intelligence Review* 53.5 (2020): 3201-3230.

[44] Represa, Natacha Soledad, et al. "Data mining paradigm in the study of air quality." *Environmental Processes* 7.1 (2020): 1-21.

[45] Alabadla, Mustafa, et al. "Systematic review of using machine learning in imputing missing values." *IEEE Access* 10 (2022): 44483-44502.

[46] Farhangfar, Alireza, Lukasz A. Kurgan, and Witold Pedrycz. "A novel framework for imputation of missing values in databases." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 37.5 (2007): 692-709.
[47] Gudivada, Venkat, Amy Apon, and Junhua Ding. "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations." *International Journal on Advances in Software* 10.1 (2017): 1-20.
[48] Emmanuel, Tlamelo, et al. "A survey on missing data in machine learning." *Journal of Big data* 8 (2021): 1-37.
[49] Fleming, Thomas R. "Addressing missing data in clinical trials." *Annals of internal medicine* 154.2 (2011): 113-117.
[50] Wunderlich, Gooloo S., Dorothy P. Rice, and Nicole L. Amado. "INSTITUTE OF MEDICINE and Committee on National Statistics Division of Behavioral and Social Sciences and Education."
[51] Lin, Wei-Chao, and Chih-Fong Tsai. "Missing value imputation: a review and analysis of the literature (2006–2017)." *Artificial Intelligence Review* 53 (2020): 1487-1509
[52] Kaiser, Jiří. "Dealing with Missing Values in Data." *Journal of Systems Integration (1804-2724)* 5.1 (2014).
[53] Le Morvan, Marine, et al. "What'sa good imputation to predict with missing values?." *Advances in Neural Information Processing Systems* 34 (2021): 11530-11540.
[54] Moons, Karel GM, et al. "Using the outcome for imputation of missing predictor values was preferred." *Journal of clinical epidemiology* 59.10 (2006): 1092-1101..
[55] Khan, Shahidul Islam, and Abu Sayed Md Latiful Hoque. "SICE: an improved missing data imputation technique." *Journal of big Data* 7.1 (2020): 37.
[56] Johnson, Thomas F., et al. "Handling missing values in trait data." *Global Ecology and Biogeography* 30.1 (2021): 51-62.
[57] Khan, Shahidul Islam, and Abu Sayed Md Latiful Hoque. "SICE: an improved missing data imputation technique." *Journal of big Data* 7.1 (2020): 37.
[58] Woods, Adrienne D., et al. "Best practices for addressing missing data through multiple imputation." *Infant and Child Development* 33.1 (2024): e2407.
[59] Petrazzini, Ben Omega, et al. "Evaluation of different approaches for missing data imputation on features associated to genomic data." *BioData mining* 14 (2021): 1-13.
[60] Austin, Peter C., et al. "Missing data in clinical research: a tutorial on multiple imputation." *Canadian Journal of Cardiology* 37.9 (2021): 1322-1331.
[61] Huque, Md Hamidul, et al. "Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model." *Biometrical Journal* 62.2 (2020): 444-466.
[62] Wutchiett, David, and Claire Durand. "Multilevel and time-series missing value imputation for combined survey and longitudinal context data." *Quality & Quantity* 56.3 (2022): 1799-1828.

[63] Odu, Anthomy, and Daniel Adedokun. "Incorporating Active Learning Techniques into Multi-Domain Performance Optimization: A Comprehensive Investigation into the Synergistic Integration of Data-Driven Strategies and Adaptive Learning Models to Enhance Cross-Domain Efficiency and Performance across Diverse Application Areas." (2023).

[64] Thomas, Tressy, and Enayat Rajabi. "A systematic review of machine learning-based missing value imputation techniques." *Data Technologies and Applications* 55.4 (2021): 558-585.

[65] Alabadla, Mustafa, et al. "Systematic review of using machine learning in imputing missing values." *IEEE Access* 10 (2022): 44483-44502

[66] Raja, P. S., and K. J. S. C. Thangavel. "Missing value imputation using unsupervised machine learning techniques." *Soft Computing* 24.6 (2020): 4361-4392.

[67] Piri, Saeed. "Missing care: A framework to address the issue of frequent missing values; The case of a clinical decision support system for Parkinson's disease." *Decision Support Systems* 136 (2020): 113339.

[68] Nijman, Steven Willem Joost, et al. "Real-time imputation of missing predictor values improved the application of prediction models in daily practice." *Journal of clinical epidemiology* 134 (2021): 22-34..

[69] Thomas, Tressy, and Enayat Rajabi. "A systematic review of machine learning-based missing value imputation techniques." *Data Technologies and Applications* 55.4 (2021): 558-585.

[70] Thomas, Tressy, and Enayat Rajabi. "A systematic review of machine learning-based missing value imputation techniques." *Data Technologies and Applications* 55.4 (2021): 558-585.

[71] Maheswari, K., et al. "Missing data handling by mean imputation method and statistical analysis of classification algorithm." *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing: BDCC 2018*. Springer International Publishing, 2020.

[72] K.G. Jöreskog, Structural Equation Modeling with Ordinal Variables Using LISREL, Technical Report, Scientific Software International, Inc., Lincolnwood, IL, 2005

[73] A. Pantanowitz, T. Marwala, Evaluating the impact of missing data imputation, in: Advanced Data Mining and Applications, Springer Berlin Heidelberg, 2009, pp. 577–586

[74] M. Huisman, Imputation of missing network data: Some simple procedures, J. Soc. Struct. 10 (1) (2009) 1–29

[75] S.C.W.C. Albright, W. Winston, C. Zappe, Data Analysis and Decision Making, Cengage Learning, 2010.

[76] S. Tufféry, Data Mining and Statistics for Decision Making, Wiley, John & Sons, 2011.

[77] K.J. Lee, J.C. Galati, J.A. Simpson, J.B. Carlin, Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study, Stat. Med. 31 (30) (2012) 4164–4174.

[78] I. Eekhout, R.M. de Boer, J.W. Twisk, H.C. de Vet, M.W. Heymans, Missing data: a

systematic review of how they are reported and handled, Epidemiology 23 (5) (2012) 729–732.
[79] K. Bache, M. Lichman, UCI Machine Learning Repository, UC Irvine, CA, USA, 2013, URL http://archive.ics.uci.edu/ml.
[80] K.J. Lee, J.C. Galati, J.A. Simpson, J.B. Carlin, Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study, Stat. Med. 31 (30) (2012) 4164–4174.
[81] R. Core, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2014.
[82] S.J. Choudhury, N.R. Pal, Imputation of missing data with neural networks for classification, Knowl.-Based Syst. 182 (2019) 104838.
[83] U. Pujianto, A.P. Wibawa, M.I. Akbar, et al., K-nearest neighbor (k-NN) based missing data imputation, in: International Conference on Science in Information Technology, ICSITech, IEEE, 2019, pp. 83–88.
[84] S. Mercaldo, J.D. Blume, Missing data and prediction: the pattern submodel, Biostatistics 21 (2) (2020) 236–252
[85] C.-Y. Hung, B.C. Jiang, C.-C. Wang, Evaluating machine learning classification using sorted missing percentage technique based on missing data, Appl. Sci. 10 (14) (2020) 4920.
[86] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, O. Tabona, A survey on missing data in machine learning, J. Big Data 8 (1) (2021) 1–37.
[87] P.C. Austin, I.R. White, D.S. Lee, S. van Buuren, Missing data in clinical research: a tutorial on multiple imputation, Can. J. Cardiol. 37 (9) (2021) 1322–1331.
[88] A.R. Ismail, N.Z. Abidin, M.K. Maen, Systematic review on missing data impu[1]tation techniques with machine learning algorithms for healthcare, J. Robotics Control (JRC) 3 (2) (2022) 143–152.
[89] P.C. Chiu, A. Selamat, O. Krejcar, K.K. Kuok, S.D.A. Bujang, H. Fujita, Miss[1]ing value imputation designs and methods of nature-inspired metaheuristic techniques: A systematic review, IEEE Access (2022).
[90] U. Shahzad, N.H. Al-Noor, M. Hanif, I. Sajjad, M. Muhammad Anas, Imputation based mean estimators in case of missing data utilizing robust regression and variance–covariance matrices, Comm. Statist. Simulation Comput. 51 (8) (2022) 4276–4295.
[91] W.-C. Lin, C.-F. Tsai, J.R. Zhong, Deep learning for missing value imputation of continuous data and the effect of data discretization, Knowl.-Based Syst. 239 (2022) 108079.
[92] H. Ahn, K. Sun, K. Kim, Comparison of missing data imputation methods in time series forecasting, Comput. Mater. Continua 70 (1) (2022) 767–779.
[93] W.H. Finch, Imputation methods for missing categorical questionnaire data: A comparison of approaches, J. Data Sci. 8 (3) (2010) 361–378
[94] E. Acuna, C. Rodriguez, The treatment of missing values and its effect on classifier accuracy, in: Classification, Clustering, and Data Mining Applications, Springer, 2004, pp. 639–647.
[95] H.W.H. Hui, W. Kong, H. Peng, W.W.B. Goh, The importance of batch sensitization in missing value imputation, Sci. Rep. 13 (1) (2023) 3003.
[96] S. Tufféry, Data Mining and Statistics for Decision Making, Wiley, John & Sons, 2011.

[97] Jöreskog, Karl G. "Structural equation modeling with ordinal variables using LISREL." (2005): 9.
[98] A.C. Acock, Working with missing values, J. Marriage Fam. 67 (4) (2005) 1012–1028.
[99] L. Rodwell, K.J. Lee, H. Romaniuk, J.B. Carlin, Comparison of methods for imputing limited-range variables: a simulation study, BMC Med. Res. Methodol. 14 (1) (2014) 57.
[100] K. Psychogyios, L. Ilias, C. Ntanos, D. Askounis, Missing value imputation methods for electronic health records, IEEE Access 11 (2023) 21562–21574
[101] S.C.W.C. Albright, W. Winston, C. Zappe, Data Analysis and Decision Making, Cengage Learning, 2010.
[102] Sim, Jaemun, Jonathan Sangyun Lee, and Ohbyung Kwon. "Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications." *Mathematical problems in engineering* 2015 (2015).
[103] E.L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, M.D. Cubiles-de-la Vega, Missing value imputation on missing completely at random data using multilayer perceptrons, Neural Netw. 24 (1) (2011) 121–129.
[104] X. Su, R. Greiner, T.M. Khoshgoftaar, A. Napolitano, Using classifier-based nominal imputation to improve machine learning, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD, Springer, 2011, pp. 124–135
[105] C. Wongkamthong, O. Akande, A comparative study of imputation methods for multivariate ordinal data, J. Surv. Stat. Methodol. 11 (1) (2023) 189–212
[106] N. Sengupta, M. Udell, N. Srebro, J. Evans, Sparse data reconstruction, missing value and multiple imputation through matrix factorization, Sociol. Methodol. 53 (1) (2023) 72–114.
[107] C. Jacobsen, U. Zscherpel, P. Perner, A comparison between neural networks and decision trees, in: Machine Learning and Data Mining in Pattern Recognition, Springer, 1999, pp. 144–158.
[108] F. Honghai, C. Guoshun, Y. Cheng, Y. Bingru, C. Yumei, A SVM regression based approach to filling in missing values, in: Knowledge-Based Intelligent Information and Engineering Systems, Springer, 2005, pp. 581–587.
[109] D. He, Active learning for ordinal classification on incomplete data, Intell. Data Anal. 27 (3) (2023) 613–634.
[110] A.A. Ahmed, New technique for imputing missing item responses for an ordinal variable, 2007, Using Tennessee Youth Risk Behavior Survey as an Example.
[111] A. Palanivinayagam, R. Damaševičius, Effective handling of missing values in datasets for classification using machine learning methods, Information 14 (2) (2023) 92.
[112] S. Pan, S. Chen, Empirical comparison of imputation methods for multivariate missing data in public health, Int. J. Environ. Res. Public Health 20 (2) (2023) 1524.
[113] Hasan, Md Kamrul, et al. "Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021)." *Informatics in Medicine Unlocked* 27 (2021): 100799.

[114] Lin W-C, Tsai C-F. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev 2020;53:1487–509

[115] Alamoodi, A. H., et al. "Machine learning-based imputation soft computing approach for large missing scale and non-reference data imputation." *Chaos, Solitons & Fractals* 151 (2021): 111236.

[116] Blazek, Katrina, et al. "A practical guide to multiple imputation of missing data in nephrology." *Kidney International* 99.1 (2021): 68-74.

[117] Seu, Kimseth, Mi-Sun Kang, and HwaMin Lee. "An intelligent missing data imputation techniques: A review." *JOIV: International Journal on Informatics Visualization* 6.1-2 (2022): 278-283.

[118] Mital, Utkarsh, et al. "Sequential imputation of missing spatio-temporal precipitation data using random forests." *Frontiers in Water* 2 (2020): 20.

[119] Dagdoug, Mehdi, Camelia Goga, and David Haziza. "Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison." *Journal of Survey Statistics and Methodology* 11.1 (2023): 141-188.

[120] Wang, Chunzhi, et al. "A new approach for missing data imputation in big data interface." *Information Technology and Control* 49.4 (2020): 541-555.

[121] Lin, Wei-Chao, Chih-Fong Tsai, and Jia Rong Zhong. "Deep learning for missing value imputation of continuous data and the effect of data discretization." *Knowledge-Based Systems* 239 (2022): 108079.

[122] Lin, Wei-Chao, and Chih-Fong Tsai. "Missing value imputation: a review and analysis of the literature (2006–2017)." *Artificial Intelligence Review* 53 (2020): 1487-1509.

[123] Feng, Runhai, Dario Grana, and Niels Balling. "Imputation of missing well log data by random forest and its uncertainty analysis." *Computers & Geosciences* 152 (2021): 104763.

[124] Bilal, Mehwish, et al. "Auto-prep: efficient and automated data preprocessing pipeline." *IEEE Access* 10 (2022): 107764-107784.