Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

Impact of Missing Values on Machine Learning Classification Using a Mean Imputation Strategy

Fazal Malik (Corresponding Author)
Department of Computer Science, Iqra National University Peshawar, Khyber Pakhtunkhwa, Pakistan. Email: fazal.malik@inu.edu.pk

Muhammad Suliman
Department of Computer Science, Iqra National University Peshawar, Khyber Pakhtunkhwa, Pakistan. Email: salman.u360@gmail.com

Atiq Ur Rahman
Faculty of Computer Information Science, Higher Colleges of Technology, Ras Al Khaimah Campus, United Arab Emirates. Email: arahman@hct.ac.ae

Rahmat Hussain
Institute of Computer Science and Information Technology, University of Science and Technology Bannu, Khyber Pakhtunkhwa, Pakistan
Email: imrahmatwazir@gmail.com

Muhammad Javed
Institute of Computer Science and Information Technology, University of Science and Technology Bannu, Khyber Pakhtunkhwa, Pakistan
Email: drjaved@ustb.edu.pk

Ashraf Ullah
Institute of Computer Science and Information Technology, University of Science and Technology Bannu, Khyber Pakhtunkhwa, Pakistan
Email: ashrafbth@gmail.com

Afsheen Khalid
Center for Excellence in IT, Institute of Management Sciences, Peshawar, Khyber Pakhtunkhwa, Pakistan. Email: afsheen.khalid@imsciences.edu.pk

Abstract
Having missing data frequently leads to lower classification accuracy and adds bias to the learning process. Although there are several ways to impute missing data, their impact at different amounts of missingness is not explored enough, especially in Random Forest. The study is designed to (1) understand how much missing data affects the Random Forest algorithm's evaluation on the Titanic dataset and (2) check if the algorithm's accuracy can be preserved with mean imputation. We used mean imputation to fill in the missing data in the dataset to represent three different levels of how data could be missing. The Random Forest classifier with 100 decision trees is used for classification. Accuracy, precision, recall, and the F1-score are used to assess all the models, with a train-test split of 80% to 20% (random_state=42) for the same results every time. The research found that the precision of predictions decreases as the amount of missingness increases. 87.00% (20%), 85.33% (40%), and 79.29% (60%). There is little change in the precision

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

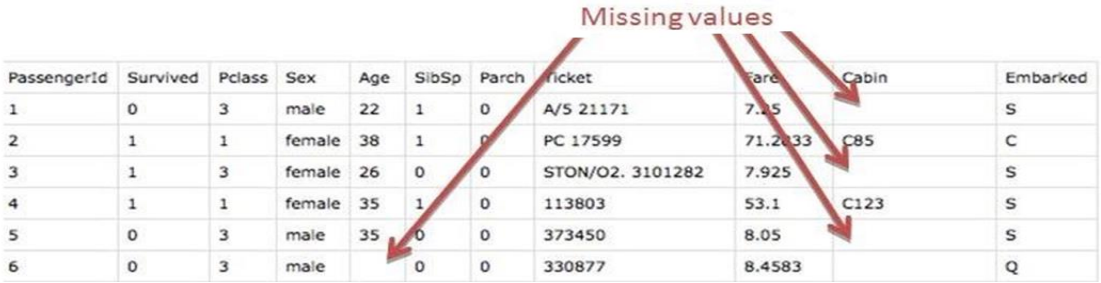
(78%–86%), though the level of recall fell to 67.23% at 60%, reflecting reduced awareness of minority-class outcomes. Even with a basic imputation, Random Forest is able to outperform other methods like K-Nearest Neighbors imputation using the Instance-Based Learning algorithm IB3 (73% at 20%). According to these findings, mean imputation works well with missingness up to 20%, but shows its weakness at higher levels. It shows that Random Forest can handle missing data well and advises using advanced imputation methods, such as MICE or K-Nearest Neighbors, when the data has a lot of missing parts.

Keywords: Missing data, mean imputation, Random Forest, classification accuracy, Titanic dataset.

1. Introduction

1.1. Missing Values

Currently, datasets commonly have missing values, incorrect encoding, and other data issues that prevent them from being used for making decisions or building models. It is possible for these missing values to be NaN (Not a Number), Null, a filler encoding value like -1, or an empty string in CSVs. Since missing value datasets do not have a regular pattern, it is difficult to handle them. Experiments have tested a number of algorithms to take care of missing fields [1]. The Figure 1 introduces missing values in a data set through an example.



PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

Figure 1. Example Illustrating Missing Values in a Dataset

1.2. Effects on Classification Performance

Missing values limits machine learning in classification by reducing learning effectiveness and poorer prediction accuracy. They cause data to deteriorate and add bias, meaning it is important to use reliable imputation. The missing data can be found in three different forms. The three types are Missing Completely at Random (MCAR), missing at Random (MAR), and Missing Not at Random (MNAR). Of these, MNAR is the most significant reason for poor model performance. It is important to use imputation methods such as MICE or k-NN to restore the quality of data and improve the reliability of classifying the data [2].

1.3. Classification Performance

Machine learning allows computers to make predictions about the future based on information. Such models are most useful for predicting new cases, as they combine various predictions to give a better result. Not having enough data often reduces the accuracy of predictions during classification [3].

1.4. Missing Data Mechanism

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

How and why some data is actually missing is an important matter to investigate. Missing data can occur for three main reasons. The missing data cases fall under MCAR, MAR, and MNAR [4].

1.4.1. Missing Completely at Random (MCAR)

MCAR states that the chance of missingness is not linked to any of the observed or unobserved data. No relationship between the missing values and any other variable in the data shows that they happened at random.

1.4.2. Missing at Random (MAR)

MAR means that the missingness is connected to the information that has been recorded, yet not to the missing data. The likelihood of missing values is influenced by other measurements, but not by the actual value that is missing.

1.4.3. Missing Not at Random (MNAR)

MNAR happens when the missing data is connected to the unobserved information. As a result, the probability of encountering missing values relies on their own pattern, so it is the hardest mechanism to address.

1.5. Strategies for Handling Missing Data

Different solutions are available for handling missing data, since each case is unique and depends on data features [5]. Approaches for handling missing information are outlined below:

1.5.1. Complete Case Analysis

Missing records are identified and removed as part of this method. In most cases, it is applied to small datasets where the missing data quantity is not significant. It could cause a lot of data loss, and the results might become biased if the missing part isn't randomly chosen.

1.5.2. Available Case Analysis

Also referred to as pairwise deletion, it allows all data to be used, even if some fields are missing. This is often used to find values such as the mean, median, or mode, and each variable gets analyzed using the given data.

1.5.3. Weighting Procedures

To account for responses that are missing, survey analysts often use weighting. The probabilities of each sample being chosen are inversely linked to their assigned weights. It is necessary to adjust these weights to get rid of bias and better match the sample design.

1.5.4. Imputation Procedures

Imputation fills in the spaces where no data is available with estimates of the missing figures. Popular ways of handling missing values are simply using the mean, median, or mode of the related variable. In some cases, more advanced methods estimate missing values using the information in the other attributes of the data.

1.6. Imputing Methods for Missing Values

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

Imputation methods help fill in gaps with missing data in datasets, making the data better and helping machine learning models work more accurately. Various ways to deal with missing data are available, and some of the most common are using the mean, the most frequently seen value (mode), the middle number (median), or k-Nearest Neighbors (k-NN). These techniques try to find and fix missing values in the data, so that the analysis can work better and give clearer results [6].

1.6.1. Mean and Median Imputation

These are common and easy-to-use ways to fill in missing data in groups. They fill in empty values by using the average or the middle number of the column they belong to. Although these methods are easy to use, they don't look at how different variables are connected and make the assumption that the data is normally distributed.

1.6.2. Hot Deck Imputation

Hot deck imputation steps in when there's a missing value in the data by filling it in with results from others that look similar to the one missing. Common in survey data analysis, it can work for both types of variables, like ones that are divided into categories or that have a number range.

1.6.3. k-Nearest Neighbors (k-NN) Imputation

This method fills in missing values by taking the average, or most common, value from the nearest k data points, which are picked out using a distance measure. It works well with all kinds of data, including numbers and categories, and looks at how the different things in those groups are connected.

1.6.4. Iterative Model-Based Imputation

This technique starts by making a simple predictive model for each missing value in the data, and keeps improving the model little by little until it gets better at filling in the gaps. Initial values may be filled in using the mean or by picking from nearby values, and then the model can be improved step by step, which helps it work well for tricky data.

1.6.5. Factorial Analysis for Mixed Data

Proposed by Josse and Husson [7] , this method uses component analysis to work with both numbers and categorical data, and it is really useful for dealing with datasets that have different kinds of information in them.

1.6.6. Random Forest or Bootstrap Imputation

Introduced by Stekhoven and Bühlmann [8], this method uses random forest models to step-by-step guess and replace missing values until they converge. It starts by trying a rough guess (such as the mean) and then teaches the model to get better at filling in the missing values step by step, so it can handle different kinds of data too.

Growing use of data in decision making by different industries has shown that high-quality, accurate data is essential. Despite that, missing data is a frequent issue and contributes to the reduced trustworthiness of machine learning solutions. One way to reduce this is for researchers to implement various

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

imputation methods from simple to advance. Nevertheless, few studies explore how the absence of data from input samples impacts the results of Random Forest and similar ensemble techniques.

Most research studies so far have looked at approaches to treating missing values in regression or classification problems using Support Vector Machines and k-Nearest Neighbors algorithms. At the same time, there is not much research on the effects of missing values on Random Forest classifiers. Additionally, there is not enough attention given to looking at imputation effectiveness across different numbers of missing data, such as 20%, 40%, and 60%.

This study aims to: Firstly, see if a lack of data affects the performance of a supervised classification algorithm; secondly, test how effective mean imputation is for various levels of missing data; and thirdly, check how the new method compares with other similar strategies.

The dataset used for the research is the Titanic dataset from Kaggle, with 20%, 40%, and 60% of data missing. The data's missing values are first replaced using the mean, and then classification is performed with Random Forest. The performance of a model is measured using accuracy, precision, recall, and F1 score. It provides a step-by-step process to understand how missing data affects the accuracy of classification. It shows that Random Forest can handle some missing data quite successfully but implies that the mean imputation method is not effective at higher levels. The results help to choose the most appropriate imputation methods for the particular data set.

The subsequent Section 2 explained the research methods found in the literature, and Section 3 discussed the methodology followed thereafter. The last section (Section 4) presented the findings alongside a discussion, with recommendations being provided in the Conclusion (Section 5).

2. Literature Review

This section presents research on missing data and the strategies for measuring its effects on different machine learning models.

Using the Statistical Package for the Social Sciences (SPSS), Leeuw et al. [9]carried out a case study on the analysis of missing data. They also made use of SPSS's ability to deal with missing data to help their decisions. Marvin L. et al. [10] studied the role of missing values in data mining by using the standard methods of imputing data using means, modes, medians, regression, and multiple regression. They wanted to see how these approaches affect the model's performance by looking at the results obtained with the same dataset. Missing data and its effects on Learning Classification Systems (LCS) are studied by John H. Holmes and other researchers [11]. They conducted experiments with two strategies and found that "missing values" negatively influenced the speed of training and the accuracy of predictions. Peng Liu and his team [12] encouraged using the Bayesian classifier to manage missing data, since it is equipped to use probabilistic methods and fill in the gaps using several methods of imputation. Huisman and his colleagues [13] pointed out that simple ways of imputing network data are effective. They introduced both direct imputation and imputation by description to help in managing missing values in networks. M. Kashif Gill et al. [14] looked at the consequences of missing data for predictive accuracy in hydrology. Support Vector Machine (SVM) and Artificial Neural Networks (ANNs) are used, and they learned that missing values negatively affected the model's performance.

ISSN Online: 3007-3154
ISSN Print: 3007-3146

The method used by Peter Schmitt and others [15] compared results of six imputation tools. Some unsupervised machine learning methods are KNN, FKM, SVD, Bayesian PCA, and MICE. It was found that Bayesian PCA and FKM gave the best results for both large and small datasets, whereas other methods are not as useful or took more time to run when the data was large. They looked at Linear Discriminant Analysis and KNN classifiers to insert missing data from twelve datasets by applying four methods: mean, median, deletion, and KNN [16]. It was found that using deletion imputation gives acceptable results as long as the missingness does not exceed 60%. Sutthipong Meeyai et al. [17] looked into how well multiple imputation works for problems of missing data with three mechanisms: For imputation to be accurate, designers and researchers must understand the differences between MCAR, MAR, and MNAR because they all depend on data structure. Farhangfar et al. [18] looked at how to replace missing values in discrete data with hot-deck, polytomous regression, and Naïve Bayes methods for various data loss percentages. Julie Josse et al. introduced missMDA, an R package that can perform principal component analysis on missing data and display graphical representations of the imputed values [7]. Using multiple imputation (MI) in a regression analysis, M. Kenward [19] managed to gain useful results when 50% of the data was missing. Audigier et al. found that mixing PCA with factorial analysis helps to deal with missing data when mixed datasets are used [20]. For dealing with huge datasets, Husson et al. [21] used SVD-based single imputation, proving it to work efficiently and in large scale. Kimberly T. To1 et al. [22] studied how different imputation methods affect the mining process.
According to data analysts Lynette A. Hunt et al., [23] some of the top imputation techniques for missing data are Hot Deck Imputation, IRMI, FAMD, and MissForest. They showed that these methods are accurate and thus proven to be reliable. In a similar way, Grigorios Papageorgiou and his colleagues [24] focused on why missing data handling is important in research and described useful methods to choose from. In [25], Rianne Margaretha Schouten et al. suggested a model that handles student answers and the likelihood of missing data while also including teacher and student effects. The method works to rectify bias in teacher value-added scores that result from missing observations. The researchers [26] tackled several connected areas, for example, low-rank matrix completion, multi-task learning, and reduced-order regression. By applying generalized kernel regularization, they dealt with hard non-convex problems and implemented solvable convex optimization approaches for missing information in data. Fithian et al. [27] suggested an algorithm for decision trees that is able to successfully model predictors affected by missing values using subspace clauses. Doing things this way results in more accurate and easy-to-understand results for simulated and real datasets alike. Jehanzeb R. Cheema [28] conducted a review of research on missing data methods in education and pointed out their benefits, drawbacks, and the direction for future studies. Using eight missing data techniques, Liesje Coertjens and her colleagues [29] carried out sensitivity analysis to study students' changing use of learning strategies across multiple rounds of data collection. The findings between methods differed, highlighting the role of different assumptions about missing data. Yufeng Ding et al. used Monte Carlo simulations to evaluate six well-known ways of handling missing data in classification trees [30]. Their work focused on the connections between missing data and the results they are studying, mainly in the case of company bankruptcy prediction.

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Hyun Kang argued [31] that missing data can lead to mistakes and reduced results in well-organized studies. Pedro J. García-Laencina et al. [32] looked at classification problems and evaluated the most widely used techniques for imputation. Also, Maytal Saar-Tsechansky et al. [33] suggested several hybrid solutions that allow for both efficiency and predictive power, performing well even in situations with a limited amount of resources. Ghorbani et al. [34] presented 14 sets of experiments on binary data with many different inputs and a variety of missing data rates from 5% to 50%. Based on the 10-fold cross-validation results, performance varied widely among classifiers, and also depended on the rate of missing data in the datasets. The MIEM method was found to be the most successful in all of the different scenarios. While barely making a difference, the NB model did, in comparison, as did the LR and TAN models that followed. However, methods including Mean Field (MF), TAN, and Hot-deck loared Support Vector Machine (SVM)'s performance when there was more than 30% missing data. The focus was on the importance of thoughtfully picking which imputation technique to use. Jonathon J. O'Brien et al. [35] tested a Bayesian selection model using both simulated and real data. It delivered higher precision and included more data than imputing data or discarding when data is missing. According to Ivan Jordanov et al. [36], radar signal data's quality greatly affected the classifier's performance. The authors tested different approaches to interpolate missing data, and they found that the methods worked well up to 60% gaps in the data set.

Daniel J. Mund and his co-authors [37] used information from heart disease patients to try to guess which ones might have newly developed health problems after staying in the hospital, using nine different factors. They looked at mean replacement, regression interpolation, and the hot-deck method to see which one did the best job when filling in missing data. Mean and hot-deck interpolation worked better than regression interpolation no matter if the predictors are categories or numbers. Julie Josse and her colleagues [38] came up with a simpler version of Principal Component Analysis that helps improve how often the method finds the correct results and avoids making detailed models that only work for the specific data set they're looking at. They also used nonparametric multiple imputation to check how variable the parameters are and tried out different ways of picking which features are important, using cross-validation as a check. Their method was put into the R package missMDA so it could be used by other researchers. Cuicui Sun and his team [39] made the KNNI method more accurate by switching out the usual Euclidean distance with the Gray Relational Analysis formula. This modification helped cut down on unnecessary noise and made it easier to work with things like separate parts or features. In tests with real crime data that had many missing values, their GMKNN algorithm was able to accurately identify whether a crime was serious or not around 77% of the time. Taeyoung Kim and his team[40] used four different ways to fill in missing data, and they used Support Vector Regression to help estimate how much power solar panels would generate. The k-Nearest Neighbors (KNN) method stayed consistent in making predictions even when more information was missing. They found that using KNN worked best for filling missing data in these types of applications.

Alireza Farhangfara and his team [18] found that if take the average of given samples during preprocessing, it helps improve how well a classifier does its job. Their results found that there wasn't one best way to judge polls because different people reacted differently to the way the questions are asked. Naive Bayes

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

imputation worked much better for the RIPPER algorithm when up to half the data was missing, while multivariate regression interpolation seemed to help out more with SVM using polynomial kernels. Their framework did really well when used with SVM and KNN for testing. Except for the average interpolation, all methods made fewer mistakes when the amount of missing values was more than 10%. The study also showed that some classifiers can handle missing data without much trouble, while other models do much better when the missing values are filled in. It was found by Min-Wei Huang et al. [41] that selecting the most relevant instances within a dataset improved the outcomes for numeric data. Using both attribute selection and smart imputation led to better results on datasets with different types of data. The average accuracy from using KNN and SVM for classification was 81%. Still, selecting instances did not help improve the accuracy of imputation for categorical data.

3. Methodology

In this research, we intend to find out how missing values influence the outcomes of machine learning classifiers. We obtain the Titanic dataset on Kaggle, as it includes details on passengers and their survival, and also has many missing values.

We create three versions of the dataset where 20%, 40%, and 60% of the data are missing. After that, we replace the values with the mean for the imputation process. Once the data has been preprocessed, the model uses the Random Forest classification algorithm to examine the accuracy of the predictions.

The main aim is to discover how missing data influences how well our model classifies, and to compare our observations with existing literature. All the steps in the methodology are seen in Figure 2, which gives an overview of the proposed research.

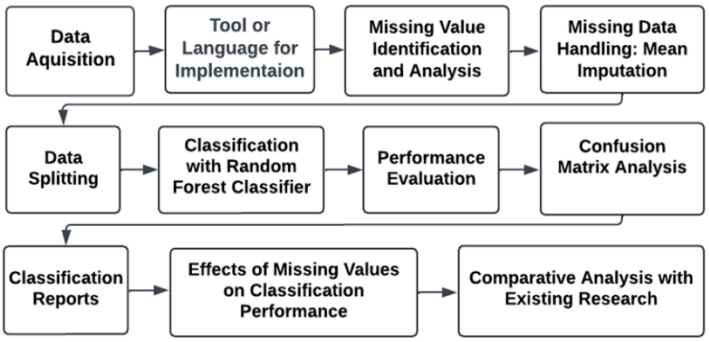


Figure 2. Proposed Workflow for Missing Data Impact on Classification

3.1. Dataset Acquisition

For this study, we use a dataset with many missing values so we can see how the amount of missing data influences the models. As the Titanic dataset from Kaggle [1] is popular in data science and has many missing values, we choose it for our experiments.

The Titanic dataset can be found online and is often used for creating predictions and teaching. It contains a lot of details about Titanic passengers, highlighting features such as Pclass, name, sex, age, ticket fare, SibSp, Parch, cabin, and survival (Survived) as shown in Table 1.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

The data is made up of 891 passenger records and 12 different features. It includes multiple data types: There are 2 columns using float, 5 using integers, and 5 using object data types.

Table 1. Structure and Description of the Titanic Dataset

Passenger Features	Non-Null Entries	Data Type	Description
PassengerId	891	int64	Unique passenger identifier
Survived	891	int64	Survival status (0=No, 1=Yes)
Pclass	891	int64	Ticket class (1=1st, 2=2nd, 3=3rd)
Name	891	object	Passenger name
Sex	891	object	Gender (Male/Female)
Age	714	float64	Age in years (contains nulls)
SibSp	891	int64	Number of siblings/spouses aboard
Parch	891	int64	Number of parents/children aboard
Ticket	891	object	Ticket number
Fare	891	float64	Passenger fare
Cabin	204	object	Cabin number (mostly null)
Embarked	889	object	Port of embarkation (C/Q/S)

3.2. Tool or Application for Implementation

For the proposed work, we use Python as the language in Jupyter Notebook. Python is famous for being easy to pick up and use, making code applications quick and understandable. Many people in science and data use Jupyter Notebook, a free, web-based application, since it allows for interactive programming with Python.

Various tasks like data mining, machine learning, deep learning, artificial intelligence, and software development are performed using the Python language. Thanks to its wide use in the scientific field, research tasks are completed promptly with high-quality results. Python works quite well with Jupyter Notebook since Jupyter offers the chance to code, show results, and document everything in an integrated way.

It's simple and efficient nature makes Python perfect for data analysis work. It is easy for Python to help developers work with data science models via specialized libraries. As a consequence, many data scientists and researchers rely on it as their main tool, mainly in the areas of machine learning and deep learning. With Jupyter Notebook, users get a user-friendly interface that works with Python, giving an overall effective platform for working on scientific problems.

3.2.1. Import Data

Python scripts use Jupyter Notebook along with different libraries to load the data. Table 2 walks through how to read and import data in Python using Jupyter Notebook. For this process, we make use of Pandas, NumPy, Matplotlib, Seaborn, and Math, all of which are necessary for working with data. With Pandas mainly can handle data reading and writing, and with the other Python libraries like NumPy, to work with multiple forms of data types.

Table 2. Sample Data Imported Using Python Libraries

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	3	Braun, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25	NaN	S
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0	PC 17599	71.2833	C85	C
2	3	1	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925	NaN	S
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
4	5	0	Allen, Mr. William Henry	male	35	0	0	373450	8.05	NaN	S

3.3. Missing Value Identification and Analysis

3.3.1. Visual Missing Value Detection

The study uses Seaborn, a Python library, to create easy-to-understand heatmaps (Figure 3) to help figure out which values in the data are missing. It displays missing values by painting the relevant region in the matrix white. The visualization interface helps quickly see where missing values are and how they are spread out in the data.

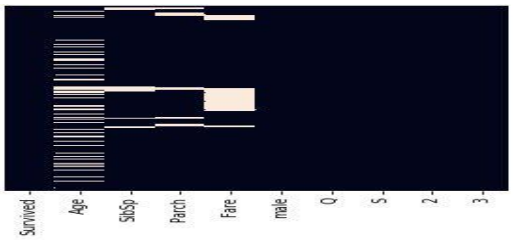


Figure 3. Heatmap Visualization of Missing Values in the Dataset

3.3.2. Missing Value Quantification

Quantitative analysis reveals that some of the most important features in the dataset have missing values. To be precise, the following numbers of missing entries are found: Age: 190 missing entries, SibSp (Siblings/Spouses aboard): 48 missing entries, Parch (Parents/Children aboard): 51 missing entries, and Fare: 144 missing entries. These represent about: Age: 21.3%, SibSp: 5.4%, Parch: 5.7%, and Fare: 16.2%.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

Table 3 represents the number of missing values in every feature. Their detection helps in determining the features that need additional caution while performing imputation, so data quality would be ensured prior to model training.

Table 3. Quantitative Distribution of Missing Values Across Key Features

Features	Missing Count	Description
Survived	0	No missing values.
Age	190	~21.3% of total entries.
SibSp	48	~5.4% of total entries.
Parch	51	~5.7% of total entries.
Fare	144	~16.2% of total entries.
male	0	No missing values.
Q	0	No missing values.
S	0	No missing values.
2	0	No missing values.
3	0	No missing values.

3.3.3. Feature Statistical Analysis

Prior to imputation, the study examines the original statistics for missing-value features. Table 4 presents mean values for these features and are taken as baselines to compare imputation performance and model accuracy. The following means are noteworthy: Age: Mean = 29.74 years, SibSp (Siblings/Spouses aboard): Mean = 0.52, Parch (Parents/Children aboard): Mean = 0.39 and Fare: Mean = 33.08.

Table 4. Mean Values of Features with Missing Data

Features	Missing Values	Mean Value
Survived	0	0.383838
Age	190	29.736334
SibSp	48	0.521945
Parch	51	0.395238
Fare	144	33.079963
male	0	0.647587
Q	0	0.086420
S	0	0.722783
2	0	0.206510
3	0	0.551066

These statistical values are a benchmark to measure how good the model learns and predicts the missing values.

3.4. Missing Data Handling: Mean Imputation

Mean imputation is usually the go-to method in statistical analysis because it helps deal with missing values in data in an easy way. This method means that when there are no values in a column, take the average, middle value, or most commonly seen value from the same kind of data and fill in the missing spots with that number. However, it does not take into account how different variables connect to each other and can make the differences in the data smaller than they really are. The effectiveness of mean imputation mostly depends on how the data is spread out and connected, which influences how well it can fill in missing values.

In this study, we used mean imputation to deal with the missing values. Specifically, we take out some details from the data on purpose, choosing to leave out 20%, 40%, and 60% of the data to see how that affected how well the models

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

worked. For each level, the average value of the feature is used to put in the spot where there is missing data.
After filling in the missing values, we ran the Random Forest classifier on each of the new datasets to see how much the missing data affected how well the model did. The performance is measured by looking at how often the model got the answer correct. Finally, we looked at how different levels of missingness affected the accuracy of the models by comparing the scores and seeing how missing data changed how well they worked when we filled in the gaps with the mean

3.5. Data Splitting

Before applying the classifier, first split the data into rows that show all the information about the case (independent) and another column that has the answer to predict (dependent). The data is then split up into two parts, a set for training and a set for testing, so that the model could be made and checked. This splitting lets the classifier look for patterns in the training data and check if it's working well using other data it hasn't seen before.

3.6. Classification Algorithms

3.6.1. Overview of Classification Algorithms

Supervised classification allows machines to learn from labeled data and later use what they have learned to make predictions about new data. They do this by finding patterns in the training data, creating a way to tell apart the target classes, and using what they learned to classify more data points. Usually, we use splits where 80% of the dataset is used for training and 20% is used for testing to judge if a model is ready to be used.

3.6.2. Random Forest Classifier Implementation

Random Forest is chosen because its decision-making involves multiple decision trees, which gives greater accuracy to the model. Because it reduces the risk of overfitting, deals easily with imbalanced data, and can rate how significant each feature is, it is invaluable for this study.

3.6.3. Experimental Design

There are three main stages in the experimental workflow, each one standing for a unique percentage of missing values.
For the first time through, 20% of the data is randomly taken out of the set. First, mean imputation is used to treat missing data, and the model is then trained using 80% of the data and evaluated with the rest.
In the second run, I increased the amount of missing data to 40%. The train and test sets are still split just as they are before, and the mean is recalculated for the new level of missing values before the model is tested.
The third scenario makes it the most difficult, because 60% of the information is missing. Every program is tested the same way, so all data can be used efficiently to note their relative efficiency.

3.6.4. Model Evaluation Framework

Model performance is measured with the help of confusion matrix and percent accuracy calculation, so comparisons can be made among models with different levels of missingness. To get a basic performance for the model, Random Forest Classifier from scikit-learn is used with the parameters n_estimators=100 and max_depth=None.

3.6.5. Data Partitioning Strategy

The data is arranged with input features (x) like age, fare, and passenger class on one end, and an outcome variable (y) that shows if someone survived or not on the

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

other end (0 or 1). To keep class distribution the same for both training and testing subsets, stratified sampling is used. It is also possible to reproduce the results by setting a fixed random state number (random_state=42). By doing this, we make sure all assessments are based on the same approach.

3.7. Performance Evaluation

The Random Forest classifier is tested on the Titanic dataset from Kaggle and is measured by accuracy (AC), precision (PR), recall (RE), and F1 score.

Accuracy (AC) score presents at what fraction of the total examples are predicted correctly

$$AC = \frac{TP + TN}{TP + FN + TN + FN} \quad (1)$$

Precision (PR) precision measures how many of the predicted instances are truly positive:

$$PR = \frac{TP}{TP + FP} \quad (2)$$

Recall (RE) (sensitivity) shows the ratio of true positives that are caught correctly:

$$RE = \frac{TP}{TP + FN} \quad (3)$$

F1 Score works by combining precision and recall in order to manage false positives and negatives:

$$F1 - Score = 2 \times \frac{PR \times RE}{PR + RE} \quad (4)$$

The metrics gleaned from the confusion matrix demonstrate how reliable the Random Forest model is in completing bimodal tasks with missing data. The experiments analyze how well a classifier works with different levels of missing data (20%, 40%, 60%) after filling in the values. The findings prove that missing data can influence how reliable predictions are, helping in choosing better preprocessing options for supervised learning jobs.

4. Results And Discussion

The implementation demonstrates the performance of Random Forest classifier on the Titanic data for 20%, 40%, and 60% missing values. The effect of missing data is tested by looking at confusion matrices, using various performance metrics, and comparing to results from previous studies.

4.1. Confusion Matrix Analysis under Mean Imputation Strategy

In the context of determining the effect of missing values on the performance of classification, confusion matrix takes the center stage. It gives a proper and precise insight about how well the machine learning model is performing after implementing the mean imputation strategy for handling missing values.

A confusion matrix provides a summary of correct and incorrect predictions through four significant metrics:

True Positives (TP): Classified positive correctly
True Negatives (TN): Classified negative correctly
False Positives (FP): Classified as positive but are negative
False Negatives (FN): Classified as negative but are positive

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

These measures provide a better understanding of the result of classification and the type of mistakes being performed by the model, particularly for datasets with missing values.

As shown in Figure 4, the confusion matrix serves as the basis for computation of important performance measures, including: Accuracy: General accuracy of the model, Precision: Number of positives predicted that are actually positive, Recall: Number of positives correctly classified and F1-Score: Harmonic mean of recall and precision.

This measure determines the success of mean imputation in maintaining classification performance, and indicates where change is beneficial.

		Condition Phase (Worst Case)		
		Condition Positive/ Shaded	Condition Negative/ Unshaded	
Testing Phase (Best Case)	Test Positive/ Shaded	True positive shaded T_p (*Correct*)	False positive shaded F_p (*Incorrect*)	Precision/Positive Predictive Value (PPV) $\frac{T_p}{T_p+F_p}$ x 100%
	Test Negative/ Unshaded	False negative unshaded F_n (*Incorrect*)	True negative unshaded T_n (*Correct*)	Negative Predictive Value (NPV) $\frac{T_n}{T_n+F_n}$ x 100
		Sensitivity/Recall Rate (RR) $\frac{T_p}{T_p+F_n}$ x 100%	Specificity Rate (SR) $\frac{T_n}{T_n+F_p}$ x 100%	

Figure 4. Confusion Matrix Architecture for Evaluating Classification Model Performance

4.1.1. Confusion Matrix Analysis with 20% Missing Data

For the initial experiment, 20% of the data is deleted and substituted through mean imputation. The Random Forest classifier is used, and the generated confusion matrix is presented in Figure 5. The matrix contains the values: TP = 36, FN = 18, FP = 6 and TN = 119.

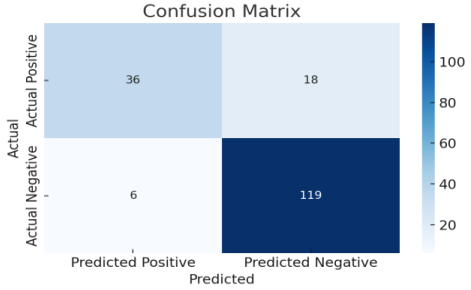


Figure 5. Confusion Matrix with 20% Missing Values

4.1.2. Confusion Matrix Analysis with 40% Missing Data

The data of 40% is replaced by mean values in the second test case and the classifier is implemented. The resulting matrix is depicted in Figure 6. The matrix contains the values: TN = 223, TP = 81, FP = 18 and FN = 35.

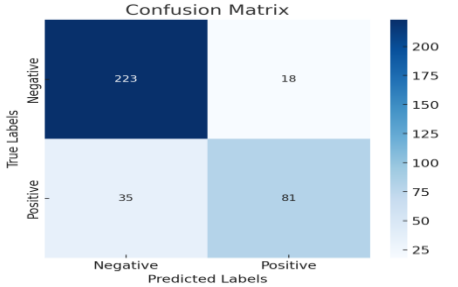


Figure 6. Confusion Matrix with 40% Missing Values

4.1.3. Confusion Matrix Analysis with 60% Missing Data

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

In the last experiment, 60% of the data is imputed and then classified. The confusion matrix in this case is illustrated in Figure 7. The matrix is assigned the values: TN = 281, TP = 143, FP = 40 and FN = 71.

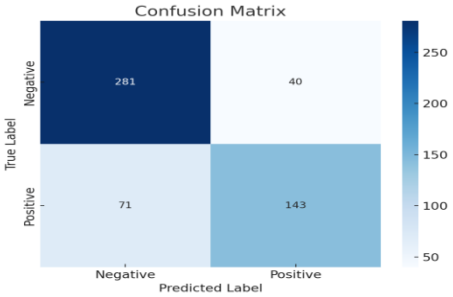


Figure 7. Confusion Matrix with 60% Missing Values.

4.2. Classification Performance with Varying Missing Data Levels

Classification is performed by using machine learning classifier on titanic data to classify survived in titanic data set. The main goal of the experimental analysis to use classifier to apply on a data set that contain different percentage of missing value and identify missing value its effects on classifier. Three-time classifier is applied on data set with different percentage of missing value. The following are the classification results and reports of these three-confusion matrixes.

This section presents the classification performance of the Random Forest classifier applied to the Titanic dataset with varying levels of missing values (20%, 40%, and 60%), each imputed using mean imputation. The classifier is evaluated using key performance metrics: accuracy (AC), precision (PR), recall (RE), and F1 score (F1).

4.2.1. Classification Performance with 20% Missing Data Imputation

In the first experiment, 20% of the data in the Titanic dataset is artificially removed to simulate missingness and subsequently addressed using mean imputation. This method replaced missing numerical values with the mean of each feature. The imputed dataset is then used to train a Random Forest classifier, selected for its robustness and capability to handle high-dimensional, imbalanced data.

Figure 8 shows the distribution of predicted outcomes, indicating a higher number of correctly classified non-survivors than survivors. This suggests potential class imbalance or a model bias toward the majority class—a common challenge in binary classification tasks.

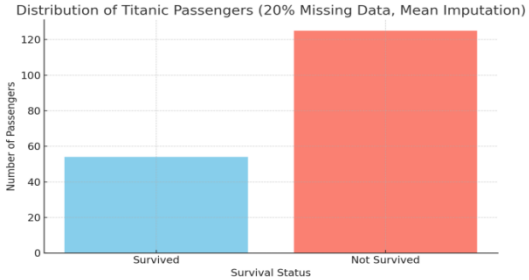


Figure 8. Predicted Outcome Distribution Showing Class Imbalance

Table 5 shows the comparison of the performance of the model. The model achieved 87% accuracy, 86% precision, 67% recall, and 73% F1-score. Although the high precision indicates correct positive predictions, the low recall shows that

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

the majority of actual survivors are mispredicted, illustrating the low minority class sensitivity of the model.

Table 5. Performance Metrics of Random Forest at 20% Missingness

AC (%)	PR (%)	RE (%)	F1-Score (%)
87	86	67	73

Random Forest performed well at this level of missingness. The precision-recall gap, however, indicates that mean imputation may not be maintaining insightful feature patterns well enough, particularly for minority class detection. Future runs would benefit from more sophisticated imputation or class-balancing methods.

4.2.2. Classification Performance with 40% Missing Data Imputation

For the second experiment, 40% of the data are randomly made missing to simulate real-world large-scale data incompleteness. Again, mean imputation is applied to impute and fill the missing values, on which a Random Forest classifier is trained from the imputed dataset.

Figure 9 shows the distribution of predicted survivors and non-survivors. While the model still discriminates between the two classes, there is a noticeable imbalance present, which may reflect the underlying class distribution of the dataset or the restriction of adding further data loss.

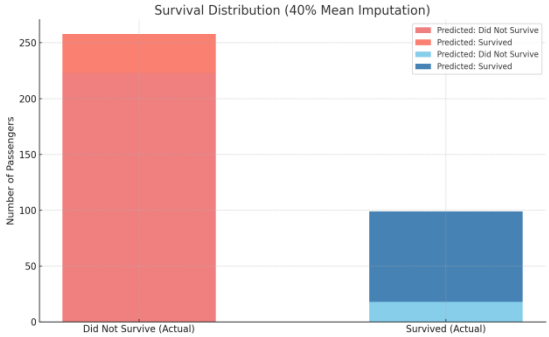


Figure 9. Predicted Class Distribution with Data Loss

Table 6 provides an overview of classifier performance. The accuracy of the model is reported at 85.33%, showing very high overall prediction ability. Precision reported at 82% shows very high positive correct prediction ratios. The recall dropped to 70%, representing low sensitivity towards the actual survivors. The computed F1-score of 72.86% shows this precision-recall disparity.

Table 6. Performance Metrics of Random Forest at 40% Missingness

AC (%)	PR (%)	RE (%)	F1 (%)
85.33	82	70	72.86

These results indicate that the Random Forest model is quite stable with moderate missingness. However, the resulting loss of recall indicates the limit of mean imputation in maintaining key data properties. For better sensitivity under such a condition, more advanced imputation methods or hybrid modeling methods might be required.

4.2.3. Classification Performance with 60% Missing Data Imputation

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

During this phase, the Titanic dataset is exposed to a critical missing data scenario where 60% of values are intentionally removed. Mean imputation is employed to rectify missing data and later a Random Forest classifier is employed. The performance of the model is tested using Accuracy (AC), Precision (PR), Recall (RE), and F1-Score.

Figure 10 presents the distribution of survivors and non-survivors by classification. While the model classifies correctly, there exists some misclassification—predominantly among the survivors—most likely resulting from colossal information loss due to missing values.

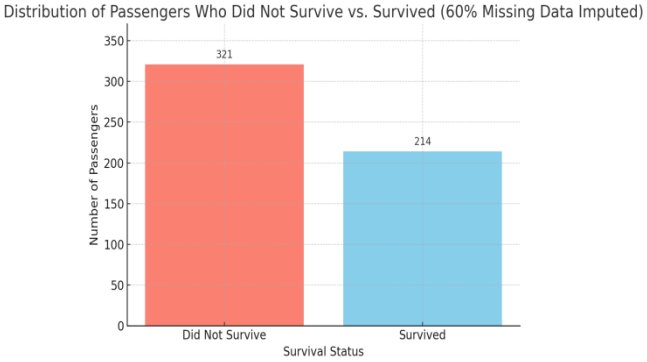


Figure 10. Survivors vs. non-survivors classification distribution

The accuracy in Table 7 demonstrates that nearly 80% of the prediction is correct. Precision is very strong, showing that the model can classify survivors with some confidence. However, the drop in recall shows reduced sensitivity in detecting all actual survivors. The fall reflects the limitation of mean imputation to acquire detailed patterns of data, especially under conditions of heavy missingness.

Table 7. Performance Metrics of Random Forest at 60% Missingness

AC (%)	PR (%)	RE (%)	F1-Score (%)
79.29	78.15	67.23	72

The result is a clear reduction in model precision compared to lower missingness. This confirms that an increase in the rate of missing data lowers classification precision even with using standard imputation techniques. The findings highlight that complex imputation or robust modeling techniques need to be employed in handling high rates of missing data.

4.3. Effects of Missing Values on Classification Performance

This section investigates the effects of higher percentages of missing data—20%, 40%, and 60%—on the classification accuracy of a Random Forest model from the Titanic data set. Mean imputation is employed to handle missing values before model training and testing.

Figure 11 shows a clear negative relationship between missing data percentage and model accuracy. The model is accurate at 87% at 20% missingness, then decreased to 85.33% at 40%, and again decreased to 79.29% at 60%. These results indicate that classification performance decreases as missing data increases.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

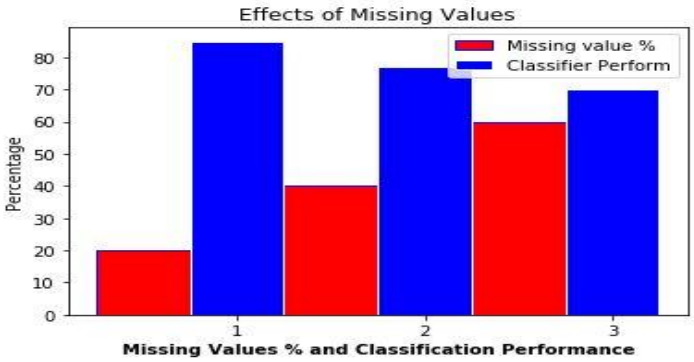


Figure 11. Values Accuracy Declines with Increased Missing Data

The degradation mirrors the shortcoming of mean imputation, especially with high levels of missingness. While mean imputation works best under low levels of missing data by maintaining data integrity, its simplicity distorts data distributions with higher levels of missingness. The distortion weakens the ability of the model to identify compound relationships among features and hence leads to poorer predictive performance.

These results confirm the need for minimizing missing data in supervised learning for model precision. Where strong high missingness cannot be avoided, more sophisticated imputation techniques need to be employed. Techniques like K-Nearest Neighbors (KNN) and Multiple Imputation by Chained Equations (MICE) retain data properties more effectively by accounting for local and global data structures.. Additionally, algorithms such as XGBoost that accommodate missing values by design can have the potential for higher robustness.

While mean imputation suffices for low missingness, it diminishes significantly in effectiveness as missing data increases. Therefore, adapting to preprocessing and imputation strategies as well as robust classifiers is necessary in order to maintain classification accuracy when data is missing.

4.4. Comparative Analysis with Existing Research

Table 8 shows a comparative analysis of the proposed work here with best-class research on the effect of missing data on model performance. The comparison is centered on major features such as datasets, methods, missingness levels, and classification accuracy and provides an integrated view of methodological performance under incompleteness in the data.

Table 8. Comparison of Proposed method with existing approaches in missing data accuracy

Authors	Dataset	Techniques	Missing Data %	Accuracy (%)
Huang et al. [41]	Healthcare	KNNI + IB3	20%	73
			40%	69
			60%	67 (Not evaluated beyond 50%)
Khalil et al. [42]	Insurance Fraud Dataset	Scenario 2, Ridge + SMOTE; 1% overfitting tolerance	20%	84

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 6 (June) (2025)

		Scenario 4, Boosting + ADASYN + KNN Imputation	40%	86
			60%	Not tested at 60% specifically; Column removal used at >15% missing
Kenward, [19]	ELSA	Regression	Up to 46.9%	Not explicitly accuracy, but MI improved coefficient stability and reduced bias compared to complete-case analysis
Proposed Work	Titanic Dataset	Mean Imputation + Random Forest	20%	87
			40%	85.33
			60%	79.29

Huang et al. [41] performed their experiments on a healthcare data set and through the K-Nearest Neighbors Imputation (KNNI) combined with Instance-Based Learning (IB3). They attained a steadily falling accuracy rising with rising missingness proportion—73% at 20%, 69% at 40%, and 67% at 60%, although they reported limited assessment beyond 50% missing data. While their method includes imputation and classification phases, the lower accuracy suggests comparative missing data sensitivity and points towards potential instance-based learner constraints in high-missingness settings.

Khalil et al. [42] tried out a range of scenarios on an insurance fraud detection data set with some of the hybrid ones like Ridge regression using SMOTE (Scenario 2) and Boosting with ADASYN and KNN Imputation (Scenario 4). They have obtained 84% accuracy at 20% missingness and 86% at 40% missingness. They did not try especially at 60% but instead chose to remove columns with over 15% missing values, which represents an inadequacy of coping with extreme missingness.

Kenward [19], based on the ELSA dataset, employed regression-based estimation with Multiple Imputation (MI).Even though accuracy criteria are not distinctly provided, the study emphasized higher coefficient stability and lower bias compared to complete-case analysis, affirming the statistical efficacy of MI in longitudinal data settings. Still, it is applied less for predictive classification than inferential modeling.

Contrarily, the proposed work uses a Random Forest classifier with mean imputation for the Titanic dataset. Even though it uses a quite straightforward imputation method, the accuracy of the model is impressive: 87% at 20%, 85.33% at 40%, and 79.29% at 60% missingness. Not only does this performance beat Huang et al. and closely follows Khalil et al.'s, but also it performs the analysis up to increased missingness, providing more information about the deteriorating patterns under increased data loss. The findings indicate Random Forests, even when combined with simple imputation techniques, to provide high robustness in handling missing data for classification problems.

ISSN Online: 3007-3154
ISSN Print: 3007-3146

The proposed approach exhibits competitive accuracy and scalability on various levels of missing data with varying ease of simplicity, hardness, and completeness of assessment. These results highlight the promise of future integration of ensemble learning approaches with effective preprocessing in scenarios with limited data as well as avenues for future improvement through superior imputation processes.

5. Conclusion

This research investigates the impact of missing data on classification accuracy using the Titanic dataset. It is curious to know how mean imputation performs with Random Forest classification algorithm with various rates of missingness: 20%, 40%, and 60%. Findings reveal that as the proportion of missing data increases, accuracy in classification decreases. Accuracy drops from 87% at 20% missingness to 79.29% at 60%. Accuracy and recall also decrease, with recall (sensitivity) taking the biggest hit, i.e., decreased ability to detect true positives in extreme missingness. Mean imputation performs well at low missingness (≤20%) by preserving model performance. At higher missingness levels (≥40%), however, it distorts feature distributions and reduces predictability. Although this approach outperforms concurrent work and its performance in heavy missingness warrants state-of-the-art approaches like Multiple Imputation by Chained Equations (MICE) and K-Nearest Neighbors (KNN). The article points out that the nature of the data is an important factor in applying machine learning. While naive imputation may compete with itself when there is mild missingness, stronger methods must be used for heavy loss of data. Random Forest offers strength due to its ensemble but loses efficacy if proper imputation is not available. Mean imputation, when used as a single imputation strategy, doesn't consider inter-variable relationships and introduces bias.

Future work should explore hybrid imputation methods, cross-validate scenarios above 60% missingness, and test models like Extreme Gradient Boosting (XGBoost) or deep learning for efficient missing data handling. This study offers an applicable method of assessing missing data's impact and recommends adopting graded, context-aware imputation methods at both data preprocessing and data collection levels.

References

1. Cunha, Maria Nascimento. "Exploring the Titanic Accident: Investigating its Impact on Marine Tourism." AVAHAN: A Journal on Hospitality &Tourism 11, no. 1 (2023).
2. Zhou, Zhihong, Jiao Mo, and Yijie Shi. "Data imputation and dimensionality reduction using deep learning in industrial data." In 2017 3rd IEEE International Conference on Computer and Communications (ICCC), pp. 2329-2333. IEEE, 2017.
3. Bishop, Christopher M., and Nasser M. Nasrabadi. Pattern recognition and machine learning. Vol. 4, no. 4. New York: springer, 2006.
4. Little, Roderick JA, and Donald B. Rubin. Statistical analysis with missing data. John Wiley & Sons, 2019.
5. Little, Roderick J. "Missing data analysis." Annual Review of Clinical Psychology 20 (2024).
6. Das, Dipalika, Maya Nayak, and Subhendu Kumar Pani. "Missing Data Handling." Cuestiones de Fisioterapia 54, no. 3 (2025): 2913-2925
7. Josse, Julie, and François Husson. "missMDA: a package for handling missing

ISSN Online: 3007-3154
ISSN Print: 3007-3146

values in multivariate data analysis." Journal of statistical software 70 (2016): 1-31.
8. Stekhoven, Daniel J., and Peter Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data." Bioinformatics 28, no. 1 (2012): 112-118.
9. Leeuw, Edith D. de, and Joop Hox “Draw more valid conclusions with SPSS Missing Data analysis” SPSS white paper 2016.
10. Brown, Marvin L., and John F. Kros. "Data mining and the impact of missing data." Industrial management & data systems 103, no. 8 (2003): 611-621.
11. John H. Holmes and Warren B. Bilker “The Effect of Missing Data on Learning Classifier System Learning Rateand Classification Performance”P.L. Lanzi et al. (Eds.): IWLCS 2002, LNAI 2661, pp. 46–60, 2003.
12. Liu, Peng, Lei Lei, and Naijun Wu. "A quantitative study of the effect of missing data in classifiers." In The Fifth International Conference on Computer and Information Technology (CIT'05), pp. 28-33. IEEE, 2005.
13. Huisman, Mark. "Imputation of missing network data: Some simple procedures." Journal of Social Structure 10, no. 1 (2009): 1-29.
14. Gill, M. Kashif, Tirusew Asefa, Yasir Kaheil, and Mac McKee. "Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique." Water resources research 43, no. 7 (2007).
15. Peter Schmitt, Jonas Mandel and MickaelGuedj“A Comparison of Six Methods for Missing Data Imputation” ISSN: 2155-6180 JBMBS, an open access journalVolume 6 • Issue 1 • 1000224. 2015.
16. Edgar Acu˜na1 and Caroline Rodriguez2"The treatment of missing values and its effectin the classifier accuracy" 2004.
17. Meeyai, Sutthipong. "Logistic regression with missing data: a comparisson of handling methods, and effects of percent missing values." Journal of Traffic and Logistics Engineering 4, no. 2 (2016).
18. Farhangfar, Alireza, Lukasz Kurgan, and Jennifer Dy. "Impact of imputation of missing values on classification error for discrete data." Pattern Recognition 41, no. 12 (2008): 3692-3705.
19. Kenward, Mike. "Exploring the impact of missing data in multiple regression." (2015).
20. Audigier, Vincent, François Husson, and Julie Josse. "A principal component method to impute missing values for mixed data." Advances in Data Analysis and Classification 10 (2016): 5-26.
21. Husson, F., J. Josse, B. Narasimhan, and G. Robin. "Imputation of mixed data with multilevel singular value decomposition, arXiv e-prints." arXiv preprint arXiv:1804.11087 (2018).
22. To, Kimberly T., Rebecca C. Fry, and David M. Reif. "Characterizing the effects of missing data and evaluating imputation methods for chemical prioritization applications using ToxPi." BioData Mining 11 (2018): 1-12.
23. Hunt, Lynette A. "Missing data imputation and its effect on the accuracy of classification." In Data Science: Innovative Developments in Data Analysis and Clustering, pp. 3-14. Springer International Publishing, 2017.
24. Papageorgiou, Grigorios, Stuart W. Grant, Johanna JM Takkenberg, and Mostafa M. Mokhles. "Statistical primer: how to deal with missing data in scientific research?." Interactive cardiovascular and thoracic surgery 27, no. 2

ISSN Online: 3007-3154
ISSN Print: 3007-3146

(2018): 153-158.
25. Schouten, Rianne Margaretha, Peter Lugtig, and Gerko Vink. "Generating missing values for simulation purposes: a multivariate amputation procedure." Journal of Statistical Computation and Simulation 88, no. 15 (2018): 2909-2930.
26. McCandless, Tyler, Sue Ellen Haupt, and George Young. "REPLACING MISSING DATA FOR ENSEMBLE SYSTEMS." (2011).
27. Fithian, William, and Rahul Mazumder. "Flexible low-rank statistical modeling with missing data and side information." Statistical Science 33, no. 2 (2018): 238-260.
28. Cheema, Jehanzeb R. "A review of missing data handling methods in education research." Review of Educational Research 84, no. 4 (2014): 487-508.
29. Coertjens, Liesje, Vincent Donche, Sven De Maeyer, Gert Vanthournout, and Peter Van Petegem. "To what degree does the missing-data technique influence the estimated growth in learning strategies over time? A tutorial example of sensitivity analysis for longitudinal data." PloS one 12, no. 9 (2017): e0182615.
30. Ding, Yufeng, and Jeffrey S. Simonoff. "An investigation of missing data methods for classification trees applied to binary response data." Journal of Machine Learning Research 11, no. 1 (2010).
31. Kang, Hyun. "The prevention and handling of the missing data." Korean journal of anesthesiology 64, no. 5 (2013): 402-406.
32. García-Laencina, Pedro J., José-Luis Sancho-Gómez, and Aníbal R. Figueiras-Vidal. "Pattern classification with missing data: a review." Neural Computing and Applications 19 (2010): 263-282.
33. Saar-Tsechansky, Maytal, and Foster Provost. "Handling missing values when applying classification models." (2007).
34. Ghorbani, Soroosh, and Michel C. Desmarais. "Performance comparison of recent imputation methods for classification tasks over binary data." Applied Artificial Intelligence 31, no. 1 (2017): 1-22.
35. O'Brien, Jonathon J., Harsha P. Gunawardena, Joao A. Paulo, Xian Chen, Joseph G. Ibrahim, Steven P. Gygi, and Bahjat F. Qaqish. "The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments." The annals of applied statistics 12, no. 4 (2018): 2075.
36. Jordanov, Ivan, Nedyalko Petrov, and Alessio Petrozziello. "Classifiers accuracy improvement based on missing data imputation." Journal of Artificial Intelligence and Soft Computing Research 8, no. 1 (2018): 31-48.
37. Mundfrom, Daniel J., and Alan Whitcomb. "Imputing Missing Values: The Effect on the Accuracy of Classification." (1998).
38. Josse, Julie, and François Husson. "Handling missing values in exploratory multivariate data analysis methods." Journal de la société française de statistique 153, no. 2 (2012): 79-99.
39. Sun, Cuicui, Chunlong Yao, Lan Shen, and Xiaoqiang Yu. "Improving classification accuracy using missing data filling algorithms for the criminal dataset." International Journal of Hybrid Information Technology 9, no. 4 (2016): 367-374.
40. Kim, Taeyoung, Woong Ko, and Jinho Kim. "Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting." Applied Sciences 9, no. 1 (2019): 204.
41. Huang, Min-Wei, Wei-Chao Lin, and Chih-Fong Tsai. "Outlier Removal in

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Model-Based Missing Value Imputation for Medical Datasets." Journal of healthcare engineering 2018, no. 1 (2018): 1817479.

42. Khalil, Ahmed A., Zaiming Liu, Ahmed Fathalla, Ahmed Ali, and Ahmad Salah. "Machine Learning based Method for Insurance Fraud Detection on Class Imbalance Datasets with Missing Values." IEEE Access (2024).