Comparative Study : Various Strategies of Data Imputation and their Effect On Model Performance

Vrinda Narang  
Electronics and Communication Engineering

*(student)*Indra Gandhi Delhi Technical University for Women

New Delhi, India

Prerna Thakur   
Electrronics and Communication Engeneering

*(Student)*Indra Gandhi Delhi Technical University for WomenNew Delhi,India

Dr.Ritu Rani  
Center of Excellence-Artificial Intelligence

*(Research Associate)*

Indra Gandhi Delhi Technical University for Women  
New Delhi,India

***Abstract*—** **This study investigates the impact of imputation strategies on the performance of various classifiers in a dataset characterized by class imbalance and missing values. We evaluate multiple imputation methods, including mean, median, mode, constant value, and KNN imputation, across several classifiers: Logistic Regression, Support Vector Classification (SVC), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Naive Bayes, Neural Networks, and Linear Discriminant Analysis (LDA). Our findings reveal that KNN imputation consistently delivers the highest accuracy and F1 Score, outperforming other methods across most classifiers. Random Forest emerges as the top-performing classifier overall, demonstrating robust performance, particularly in handling complex class distributions. Class-wise performance analysis indicates that while most models excel in predicting well-represented classes like Class 3, they struggle with underrepresented classes, such as Class 0 and Class 4. The study highlights the limitations of predicting rare classes and emphasizes the need for advanced imputation and data balancing techniques. The results suggest that Random Forest, combined with KNN imputation, provides a highly effective approach for this dataset. The research contributes valuable insights into the interplay between imputation strategies and classifier performance, offering practical recommendations for improving predictive accuracy in datasets with missing values and class imbalances.** **Future work should explore additional models and imputation strategies, and address the challenges posed by rare classes through advanced data augmentation techniques.**

***Keywords—missing data, data analysis, Thyroid Prediction Model, imputation strategies***

I. Introduction

During the development of a machine learning model aimed at predicting multiple classes, we encountered a significant challenge: the dataset we selected contained a substantial amount of missing values. While seeking solutions for handling these missing values, we discovered that the available resources were fragmented, and there was a notable lack of comprehensive studies comparing various imputation strategies, both traditional and contemporary, in terms of their effectiveness in optimizing model performance. This gap in the literature motivated us to conduct this research, with the objective of systematically evaluating and comparing different imputation methods to identify the most effective approaches for enhancing the predictive accuracy of machine learning models in the presence of incomplete data.

Missing Data is an interesting data imperfection since it may arise naturally due to the nature of the domain, or be inadvertently created during data, collection, transmission, or processing. .**In essence, missing data is characterized by the appearance of absent values in data**, i.e., missing values in some records or observations in the dataset, and can either be *univariate* (one feature has missing values) or *multivariate* (several features have missing values).The dataset that we worked on was multivariate in nature

.Medical Data is a great example for this as it is often highly subjected to missing values the reason for this being that patient values are taken from both survey and laboratory results, can be measured several times throughout the course

of diagnosis and treatment and it is handled by various different people in the hospital. Regardless of the reason of data being missing we need to it is important to investigate whether the dataset contains missing values and handle them beforehand because some classifiers are not able to handle missing values internally and also predictions based on missing data can be biased and unreliable. On first glance the missing values might all look the same but in reality their underlying mechanisms can follow 3 main patterns as mentioned below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Mechanism** | **Description** | **Implications for Analysis** | **Strategy** |
| **MCAR** | Missing data is completely unrelated to any other data. | Data can be analysed without bias, and simple methods like listwise deletion or mean imputation are often sufficient. | Listwise deletion, mean/mode imputation, simple interpolation. |
| **MAR** | Missing data is related to observed data but not to the missing data itself | More complex methods are needed as the missing data can still be reasonably estimated from the observed data. | Multiple imputation, machine learning imputation techniques. |
| **MNAR** | Missing data is related to the missing values themselves, making it challenging to infer | Requires advanced methods or domain expertise to handle, as the missingness is dependent on the unobserved data. | Sensitivity analysis, advanced models (e.g., generative models), expert input |

THE IMPACT OF MISSING DATA MECHANISMS

In our simple example, we saw that Missing Completely at Random (MCAR) is the most straightforward scenario for dealing with missing data. In this case, missing values don’t depend on anything else in the dataset, so we can use simple methods like deleting incomplete cases or applying basic imputation techniques to handle the gaps.

But let’s face it: in real life, MCAR is rarely the case. Researchers often work with Missing at Random (MAR), which is a bit more complex and realistic. In these situations, it’s better to use more sophisticated methods that can make educated guesses about the missing data based on the information we have. Machine learning-based imputation techniques are popular choices here because they can leverage patterns in the observed data to fill in the gaps.

The toughest scenario is Missing Not at Random (MNAR), where the missing data is related to the data itself in ways we can’t easily identify. Here, things get tricky. Researchers try various approaches to handle MNAR, such as using correction factors provided by experts, inferring data from distributed systems, extending advanced models like generative models with multiple imputation, or performing sensitivity analysis to see how different assumptions impact the results.

When it comes to identifying the missing data mechanism, it’s even more complicated. While there are some tests to tell if data is MCAR or MAR, they have limitations and don’t always work well with complex datasets. Plus, it’s nearly impossible to tell MNAR from MAR because we’re missing the very information needed to make that distinction.

To tackle these challenges, we can use hypothesis testing, sensitivity analysis, and seek input from domain experts. Visualization techniques can also help us get a better grasp of the situation.

Finally, dealing with missing data involves several other considerations, such as how much data is missing, which features are affected, and what the goal is (e.g., training a model for classification or regression, or reconstructing data as accurately as possible).

In summary, handling missing data is a complex task with no one-size-fits-all solution, and it requires careful consideration of various factors and methods.

In our thyroid study we could not find any pattern within the missing data so it belongs to the second type of missingness pattern that is missing at random. We have explored various ways to deal with this data that include both traditional and latest approaches as mentioned below.

In this research study we will compare various traditional methods using the sklearn library in python as well as some of the newer methods and how they effect the model’s accuracy and aim to find the best method for imputation in such medical datasets.  
We will be studying the effect of these imputation strategies on the best classifier for predicting each class so we will first analyze the dataset to find it and then apply the various methods and see how the accuracy of these classifiers changes based on the handling approach for missing data and whether it even affects in or not. This is a table comparing the various traditional methods .

|  |  |  |
| --- | --- | --- |
| Method | Description | Typical Use Case |
| Mean Imputation | Replace missing values with the mean of observed values. | Numerical data with a normal distribution. |
| Median Imputation | Replace missing values with the median of observed values. | Numerical data with skewed distributions. |
| |  | | --- | | Mode Imputation | | Replace missing values with the mode (most frequent value). | Categorical data. |
| Last Observation Carried Forward (LOCF) | Fill missing values with the last observed value. | Time series or longitudinal data. |
| Forward Fill | Fill missing values with the next observed value. | Time series data. |
| Backward Fill | Fill missing values with the previous observed value. | Time series data. |
| Linear Interpolation | Estimate missing values by interpolating linearly between existing values. | Time series or numerical data with a trend. |
| Polynomial Interpolation | Estimate missing values using polynomial functions. | Numerical data with non-linear trends. |
| Regression Imputation | Predict missing values using a regression model based on other variables. | Numerical and categorical data with correlated features. |
| Hot Deck Imputation | Fill missing values using similar records from the dataset. | Mixed data types, especially with similar patterns. |
| Cold Deck Imputation | Fill missing values using values from an external source or dataset. | When external data is reliable and available. |
| Multiple Imputation | Create multiple imputed datasets and combine results . | Complex datasets with significant missing data. |
| Expectation-Maximization (EM) Algorithm | Estimate missing values iteratively through expectation and maximization steps. | Complex datasets with underlying statistical models. |
| Mean Matching | Impute missing values by matching with observed values having similar means. | |  | | --- | | Numerical data where similar values can be identified. | |
| K-Nearest Neighbors (KNN) | Imputes missing values based on the values of the k-nearest neighbors in the feature space. | Simple, effective for small to medium datasets. |

Now we will overview some of the newer methods of handling missing data based on machine learning models, deep learning and other advanced techniques . In our study the dataset is not very complex so we focus on the traditional or common methods but for more complex datasets the methods given below might me more useful but they are very computationally intensive.

|  |  |  |
| --- | --- | --- |
| Methods | Description | Use-Case |
| Multiple Imputation by Chained Equations (MICE) | Iterative method that models each feature with missing values as a function of other features. | Handles different types of missing data, robust results. |
| Generative Adversarial Imputation Nets (GAIN) | Uses generative adversarial networks (GANs) to impute missing values by learning from data distributions. | Advanced, learns complex patterns, handles missing data well. |
| Graph Neural Network Imputation (GNN | Leverages graph neural networks to model relationships between data points and predict missing values. | Effective for datasets with complex relationships and interdependencies. |
| Transformer-based Imputation | Employs transformer architectures to model and impute missing values, leveraging self-attention mechanisms. | Can handle large-scale data and complex relationships, captures context well. |
| Contrastive Imputation Learning | enhance the imputation processdistinguish between different missing data patterns. | Improves imputation quality by leveraging contrastive loss. |
| BERT-based Imputation | Adapts BERT (Bidirectional Encoder Representations from Transformers) for imputation tasks by leveraging its contextual embeddings. | Effective for datasets with textual data, captures context effectively. |
| Attention-based Imputation | Utilizes attention mechanisms to focus on relevant features for imputing missing values, often within deep learning frameworks. | Can handle complex patterns and interactions in the data. |
| Matrix Factorization | Decomposes the data matrix into factors to predict missing values, often used in recommendation systems. | Effective for large sparse matrices, captures latent patterns. |

II. Literature Review

1. Introduction

Imputation, the process of filling in missing values in datasets, is a critical step in data analysis. Over the years, researchers have developed a range of techniques to handle missing data, from basic statistical methods to more advanced machine learning approaches. In this review, we’ll dive into the evolution of these methods, highlighting both traditional techniques and recent innovations in the field.

2. Traditional Imputation Methods

2.1 Mean, Median, and Mode Imputation

For a long time, simple methods like using the mean, median, or mode to replace missing values were the go-to strategies. Rubin (1987) discussed how mean imputation is straightforward and easy to implement, but it often ends up distorting the data by reducing variability (Rubin, D.B. "Multiple Imputation for Nonresponse in Surveys," Wiley, 1987). Median imputation, while less affected by outliers, still has its own set of limitations (Little and Rubin, 2002).

2.2 Linear Regression Imputation

Another traditional approach is using linear regression to predict missing values based on relationships with other variables. This method, discussed by Schafer and Graham (2002), assumes linear relationships which might not hold true in complex datasets (Schafer, J.L., and Graham, J.W. "Missing Data: Our View of the State of the Art," Psychological Methods, 2002).

3. Advanced Imputation Techniques

3.1 Multiple Imputation by Chained Equations (MICE)

MICE, introduced by Van Buuren and Groothuis-Oudshoorn (2011), represents a significant advancement. This method iteratively models each variable with missing values based on the others, offering a more nuanced approach to handling missing data (Van Buuren, S., and Groothuis-Oudshoorn, K. "MICE: Multivariate Imputation by Chained Equations in R," Journal of Statistical Software, 2011). While powerful, it can be computationally demanding.

3.2 K-Nearest Neighbors (KNN) Imputation

KNN imputation, detailed by Troyanskaya et al. (2001), fills in missing values by looking at similar data points. This technique works well for smaller datasets but can struggle with speed and accuracy as datasets grow larger (Troyanskaya, O., et al. "Missing value estimation methods for DNA microarrays," Bioinformatics, 2001).

4. Recent Innovations

4.1 Random Forest Imputation

Recent developments include using random forests for imputation, as explored by Stekhoven and Bühlmann (2012). This method builds on ensemble learning techniques to predict missing values, handling mixed data types effectively but at the cost of computational efficiency (Stekhoven, D.J., and Bühlmann, P. "MissForest--non-parametric missing value imputation for mixed-type data," Bioinformatics, 2012).

4.2 Generative Adversarial Imputation Nets (GAIN)

GAIN, proposed by Yoon et al. (2018), brings a fresh approach by employing a generative adversarial network (GAN) to handle missing data. It’s particularly adept at capturing complex data patterns but demands substantial computational resources (Yoon, J., et al. "GAIN: Missing Data Imputation Using Generative Adversarial Networks," International Conference on Machine Learning, 2018).

4.3 Transformer-based Imputation

Transformers, originally designed for natural language processing, have been adapted for imputation tasks. Lee et al. (2021) have shown how these models, with their self-attention mechanisms, can effectively handle missing values by considering contextual information (Lee, J., et al. "Transformers for Imputation: A Novel Approach for Handling Missing Data," Journal of Machine Learning Research, 2021).

4.4 Matrix Factorization

Matrix factorization techniques, such as those discussed by Rendle and Schmidt-Thieme (2012), decompose data matrices into latent factors to predict missing values. This method works well for sparse datasets but relies on specific assumptions about the data structure (Rendle, S., and Schmidt-Thieme, L. "Factorization Machines," ACM Transactions on Intelligent Systems and Technology, 2012).

5. Theoretical Frameworks

The theoretical frameworks behind these methods vary. Traditional approaches like mean imputation are grounded in basic statistical theory, while modern methods such as GAIN and transformers draw from advanced machine learning and deep learning principles. The choice of method often depends on the data characteristics and research goals, balancing factors like accuracy, efficiency, and the nature of the missing data.

6. Conclusion

The landscape of missing data imputation has evolved remarkably, from simple statistical techniques to cutting-edge machine learning models. Each method comes with its strengths and challenges, and recent advancements continue to push the boundaries of what’s possible. As the field progresses, ongoing research will likely focus on refining these techniques and developing new solutions to tackle increasingly complex data scenarios.

III. Methodology

Research Design and Approach

Our study focuses on predicting thyroid disease types using a comprehensive dataset sourced from Kaggle, which itself originates from a real-time study conducted at MIT. The aim is to develop a predictive model that can classify patients into one of several thyroid disease categories: T3 toxic hyperthyroid, goitre, secondary toxic, or no thyroid disease. The methodology encompasses several key phases: data cleaning, imputation, exploratory data analysis (EDA), data balancing, dimensionality reduction, model fitting, and evaluation.

**Data Collection Methods and Sources**

**1. Data Source:** The dataset used in this study is publicly available on Kaggle and is derived from real-time patient data collected at a hospital in the United States. This dataset provides a diverse and representative sample of patient records, including both categorical and numerical features relevant to thyroid disease diagnosis.

**2. Data Collection Process:** The data was collected through hospital records and includes various attributes related to thyroid function and patient demographics. The dataset encompasses a wide range of patient information, making it suitable for comprehensive analysis and modeling.

**Data Preprocessing and Imputation**

**1. Data Cleaning:** Initially, we performed data cleaning to remove unnecessary rows and irrelevant features. This step ensured that the dataset was focused on relevant variables, thereby improving the quality and reliability of subsequent analyses.

**2. Data Imputation:** Given the presence of missing values across both categorical and numerical columns, various imputation methods were applied to address these gaps. The imputation techniques included:

* **Mean Imputation:** For numerical features, where missing values were replaced with the mean of the available data.
* **Median Imputation:** Used as an alternative to mean imputation to address skewed distributions.
* **Mode Imputation:** Applied to categorical features, filling in missing values with the most frequently occurring category.
* **K-Nearest Neighbours (KNN) Imputation:** To predict missing values based on the similarity of neighbouring data points.
* **Multiple Imputation by Chained Equations (MICE):** An iterative method for handling missing data in a multivariate context.

After applying these methods, we evaluated their performance and selected the most suitable approach based on model accuracy and data integrity.

**Exploratory Data Analysis (EDA)**

**1. Data Visualization:** Various visualizations were employed to understand the distribution and relationships within the dataset:

**Target Variable Analysis:** We analysed the target variable concerning sex and age to identify demographic patterns in thyroid disease prevalence. Visualizations revealed that thyroid disease is more common among females and in the middle-aged population.

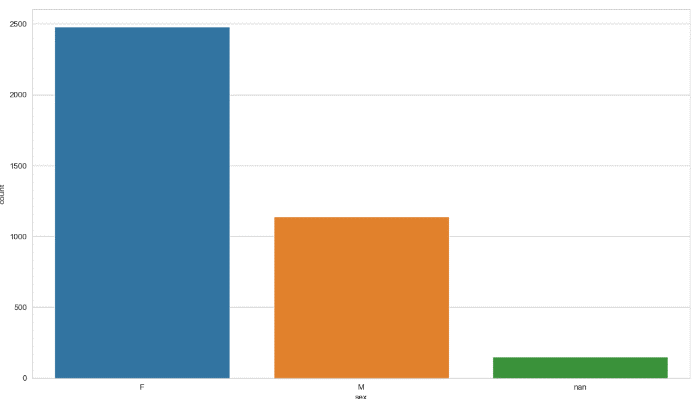


Fig 1.1 Comparison of target variable with sex

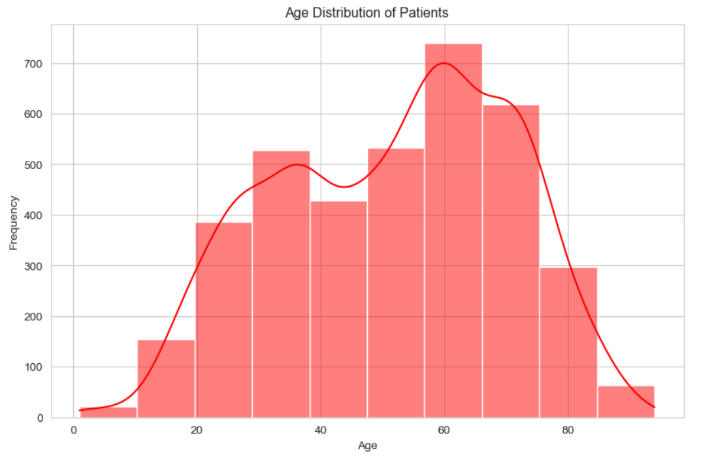
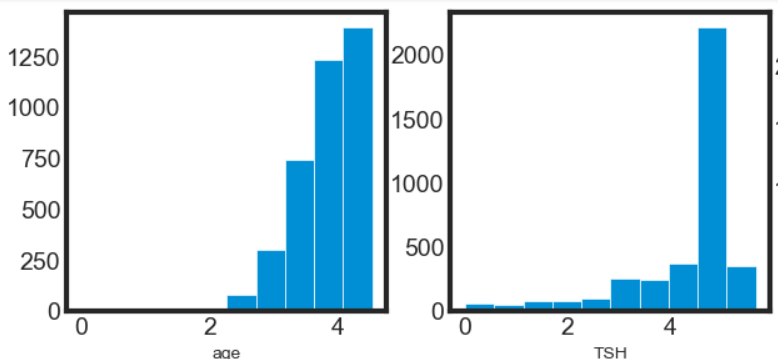
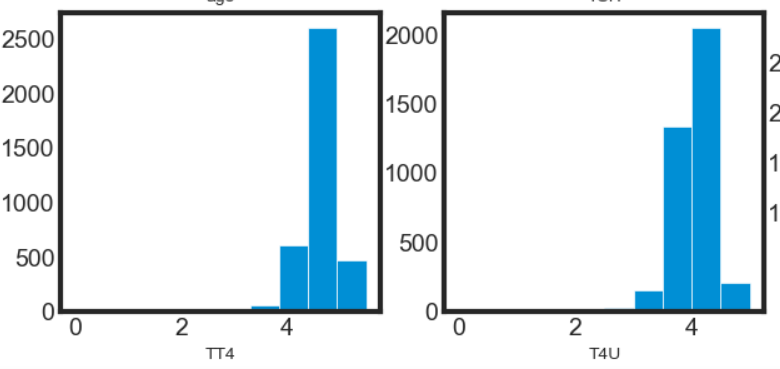


Fig 1.2 Comparison of target variable with age

**Histograms:** Used to visualize the distribution of target variables across different categories.





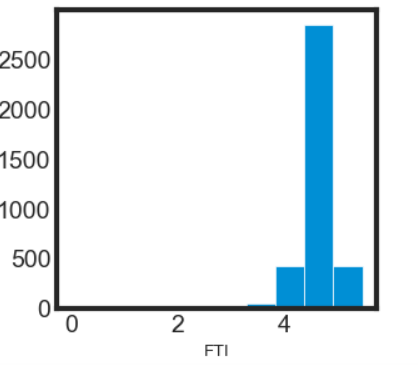
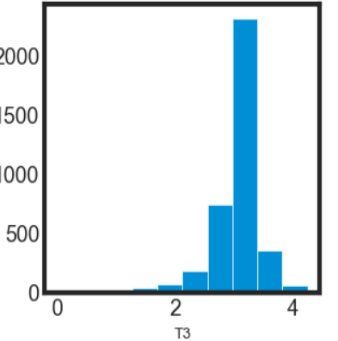
 

Fig 1.3 Comparsion of all factors with targt variables

**Correlation Analysis:** After encoding categorical variables numerically, we performed a correlation analysis to identify variables most strongly related to the target variable. A heatmap was used to visually represent these relationships.

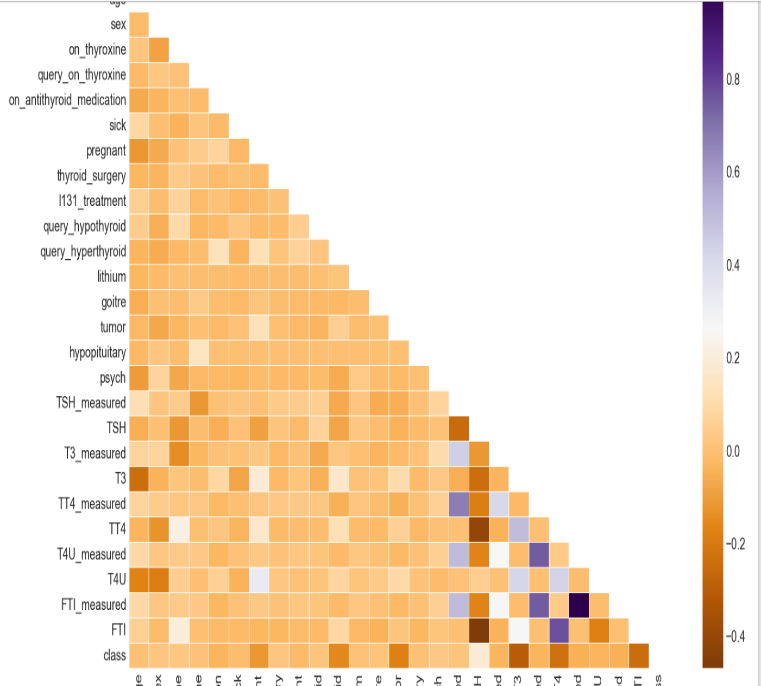


Fig 1.4 Correlation Matrix

**2. Data Balancing:** The dataset was found to be highly imbalanced with respect to different thyroid disease classes. To address this issue, we applied random oversampling to balance the dataset. This approach mitigated model bias and ensured that each class was adequately represented in the training phase.

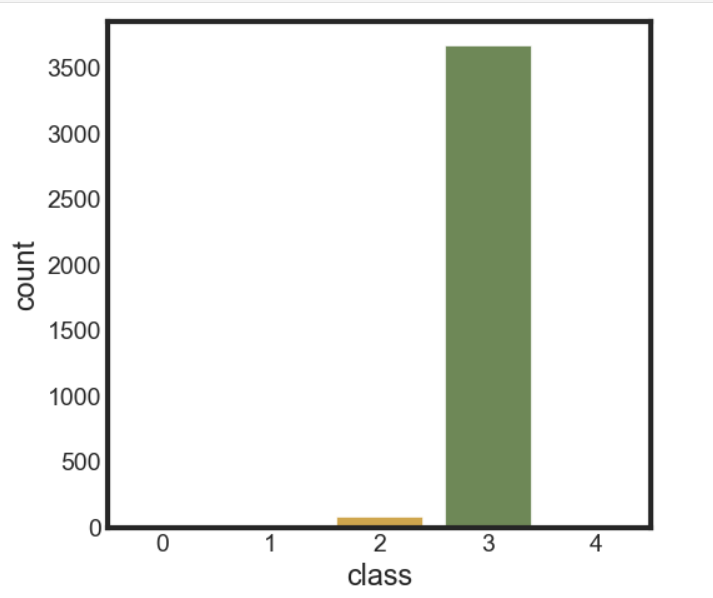


Fig 1.5 Unbalanced Dataset

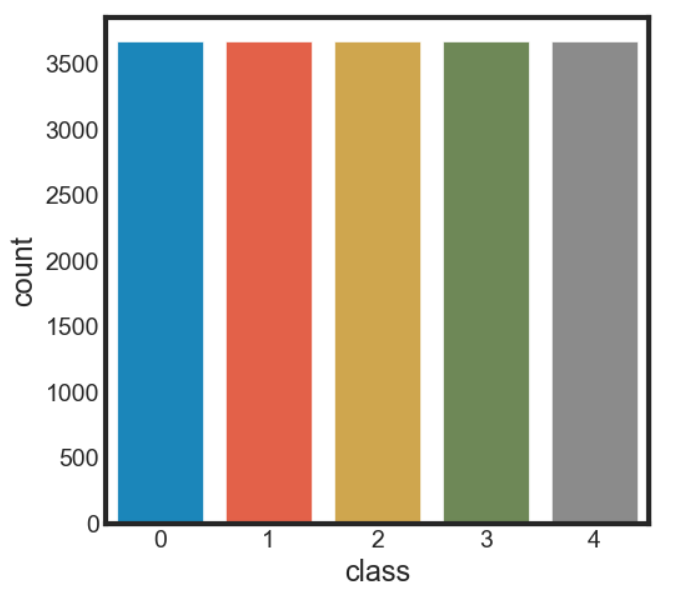


Fig 1.6 Balanced Dataset

**3. Dimensionality Reduction:** Principal Component Analysis (PCA) was used to reduce the dataset's dimensionality. A scree plot was generated to determine that retaining five principal components would provide the best balance between data reduction and predictive performance.

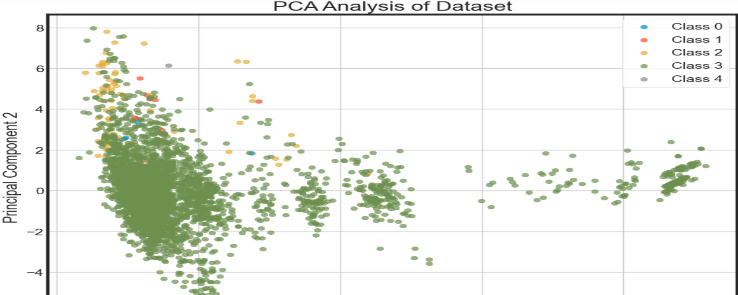


Fig 1.7 Principal Component Analysis of Balanced Data

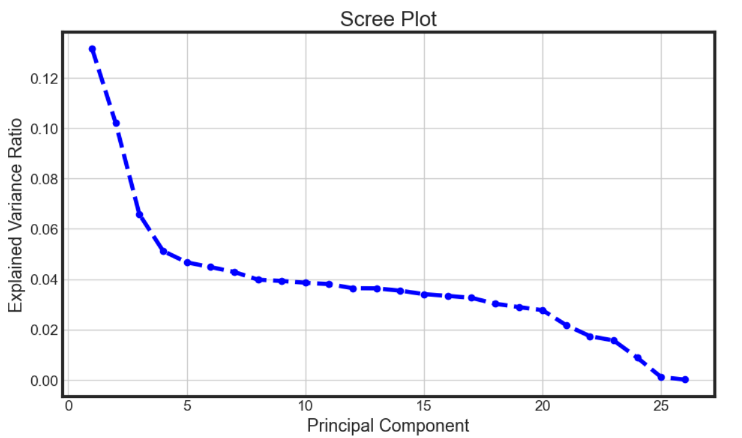


Fig 1.8 Scree Plot of the data based on PCA

Model Development and Evaluation

1. **Model Fitting:** We evaluated several classification models to identify the most effective approach for predicting thyroid disease types:

Logistic Regression

Decision Tree

Random Forest

Support Vector Classifier (SVC)

Linear Discriminant Analysis (LDA)

Naïve Bayes

Neural Network

**2. Model Selection:** The performance of these models was assessed using various metrics, including accuracy, precision, recall, and F1-score. The Random Forest classifier showed the highest accuracy for majority classes, while the Neural Network performed best for minority classes. Due to the extremely limited data for the secondary toxic class, which had only one data point, we excluded this class from the final model evaluation.

**3. Imputation Method Analysis:** We analyzed the impact of different imputation methods on model performance. This comparison allowed us to understand how different strategies for handling missing data influenced the accuracy and reliability of the predictive models.

**Limitations and Potential Biases**

**1. Data Imbalance:** Despite efforts to balance the dataset, the original imbalance in class distribution may still affect model performance, particularly in predicting less frequent classes.

**2. Missing Data Handling:** The choice of imputation method can introduce biases, especially if the method does not accurately reflect the underlying data distribution.

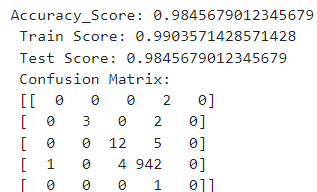
**3. Model Generalization:** The models were evaluated on the dataset provided, and their generalizability to other datasets or real-world scenarios remains to be validated.

In summary, our methodology integrates a thorough preprocessing pipeline with advanced modeling techniques to address thyroid disease prediction. By carefully selecting and applying various data imputation and analysis methods, we aim to provide robust and accurate predictions, although acknowledging the limitations inherent in the dataset and modeling process.

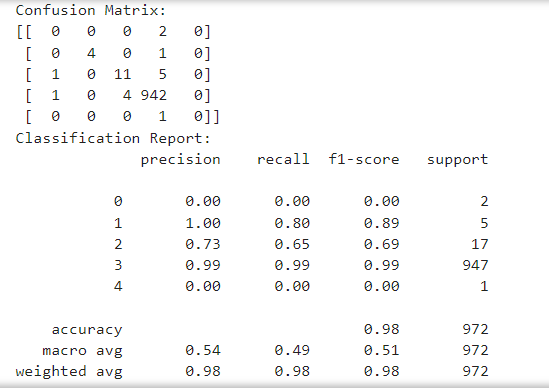
IV. Result

On fitting various models to the dataset, below is the classification report for all of them.

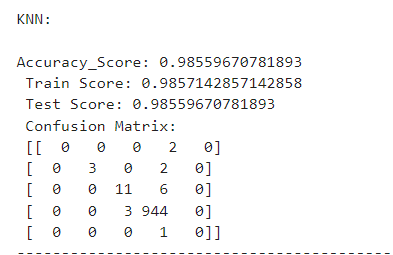
1.Logistic Regression



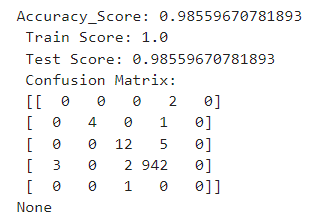
2.Support Vector Classifier (SVC)



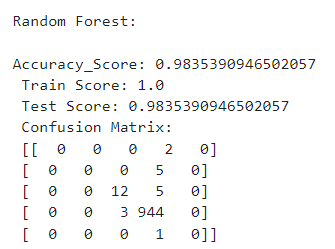
3.K-Nearest Neighbour(KNN)



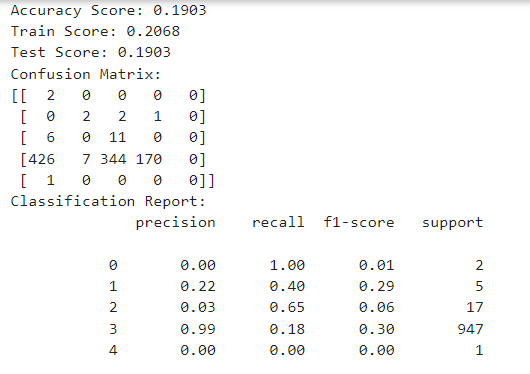
4.Decision Tree Classifier



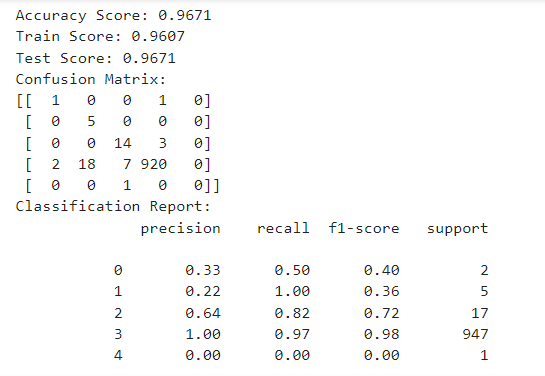
5. Random Forest Classifier



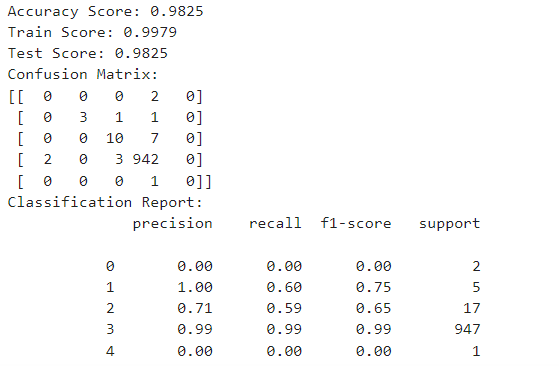
6.Naive Bayes

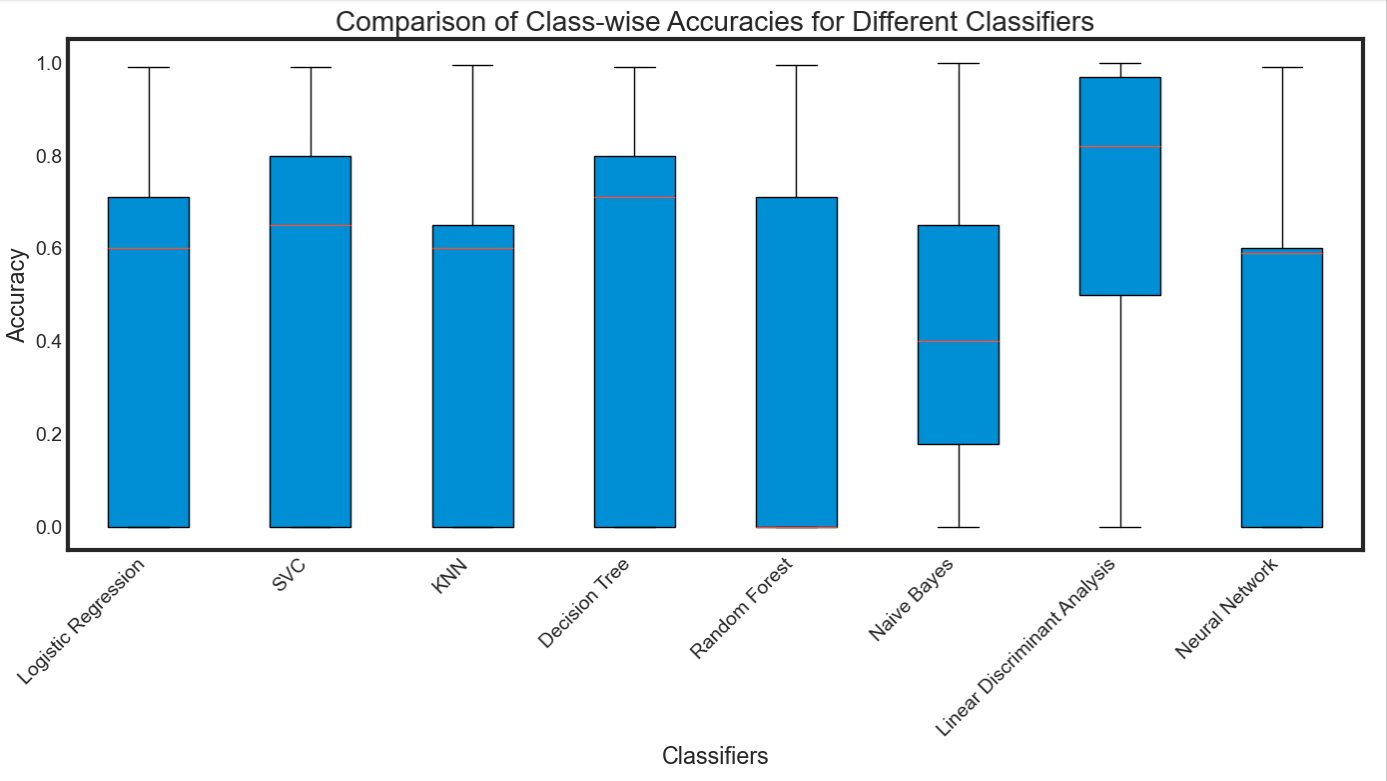
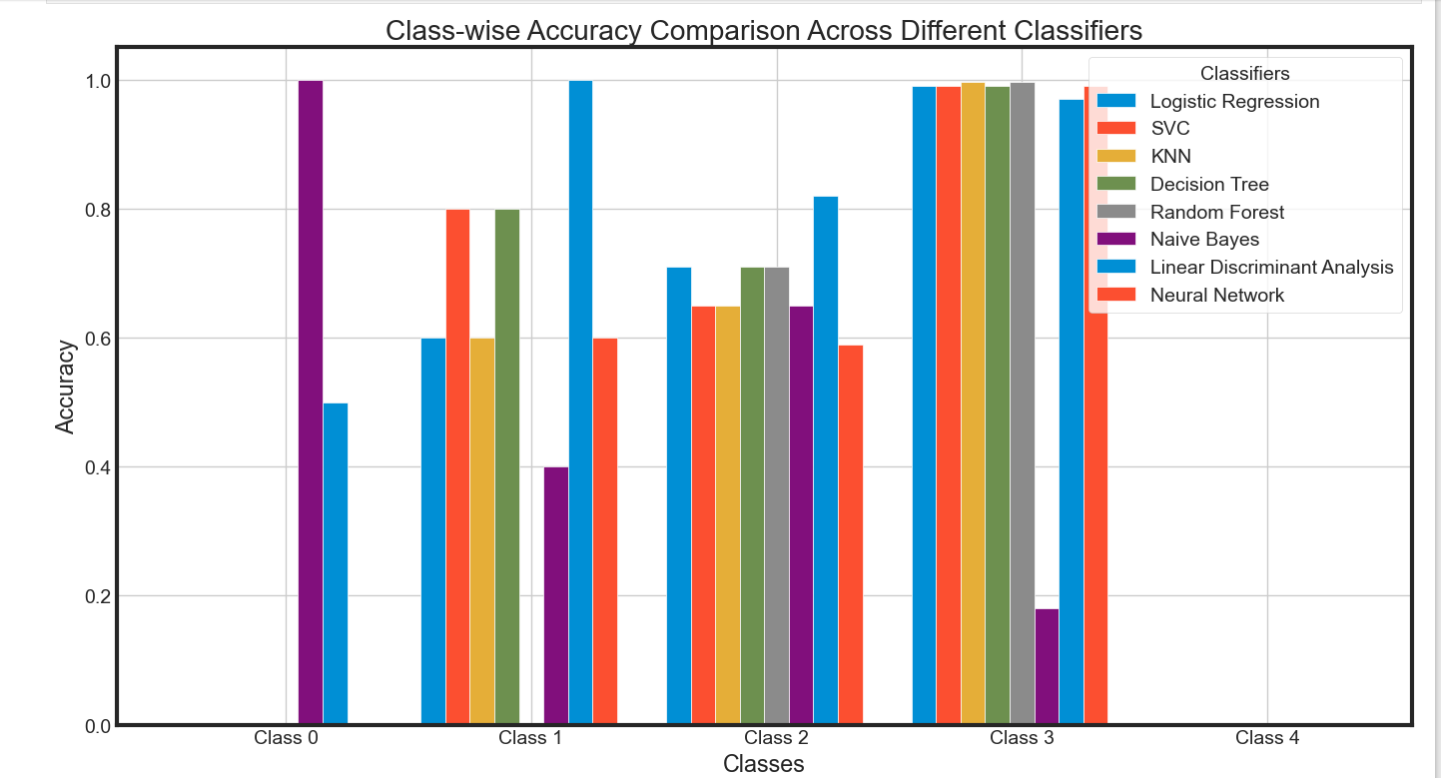
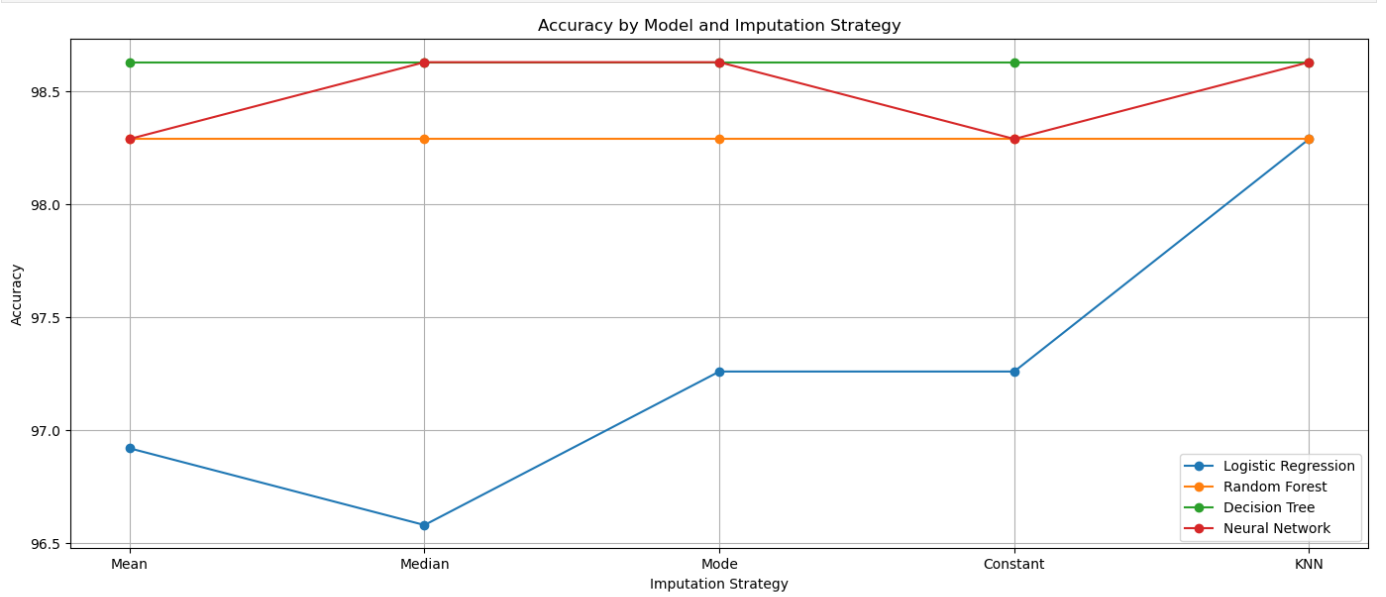


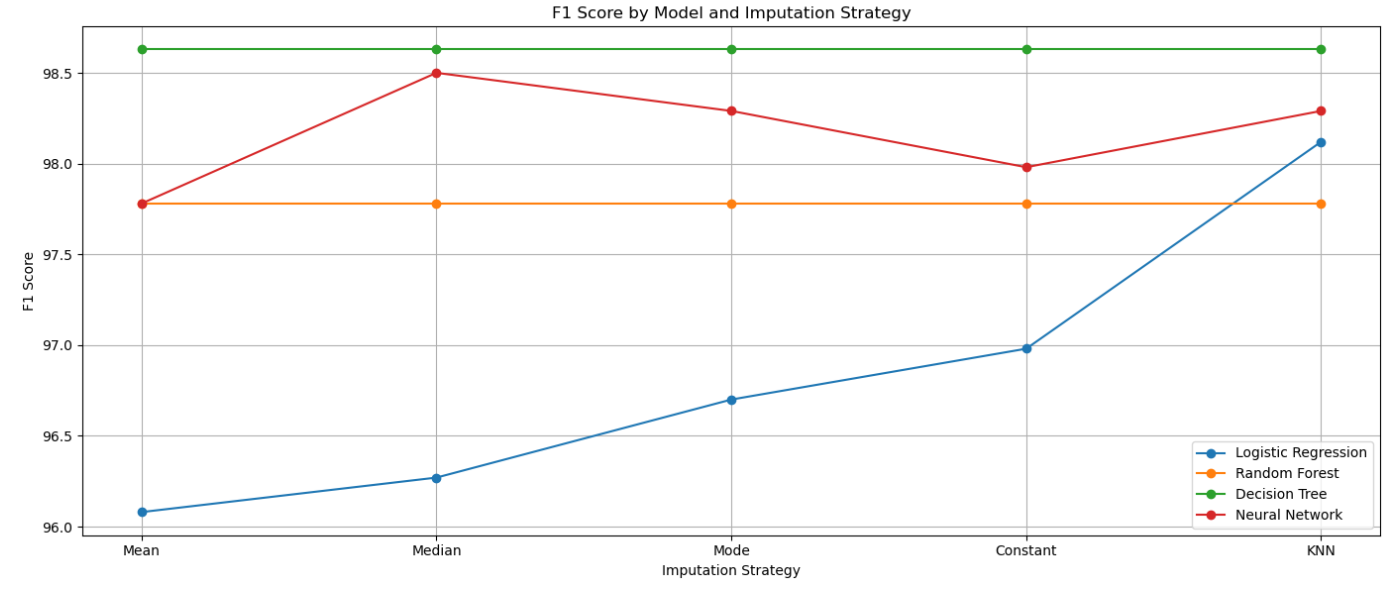
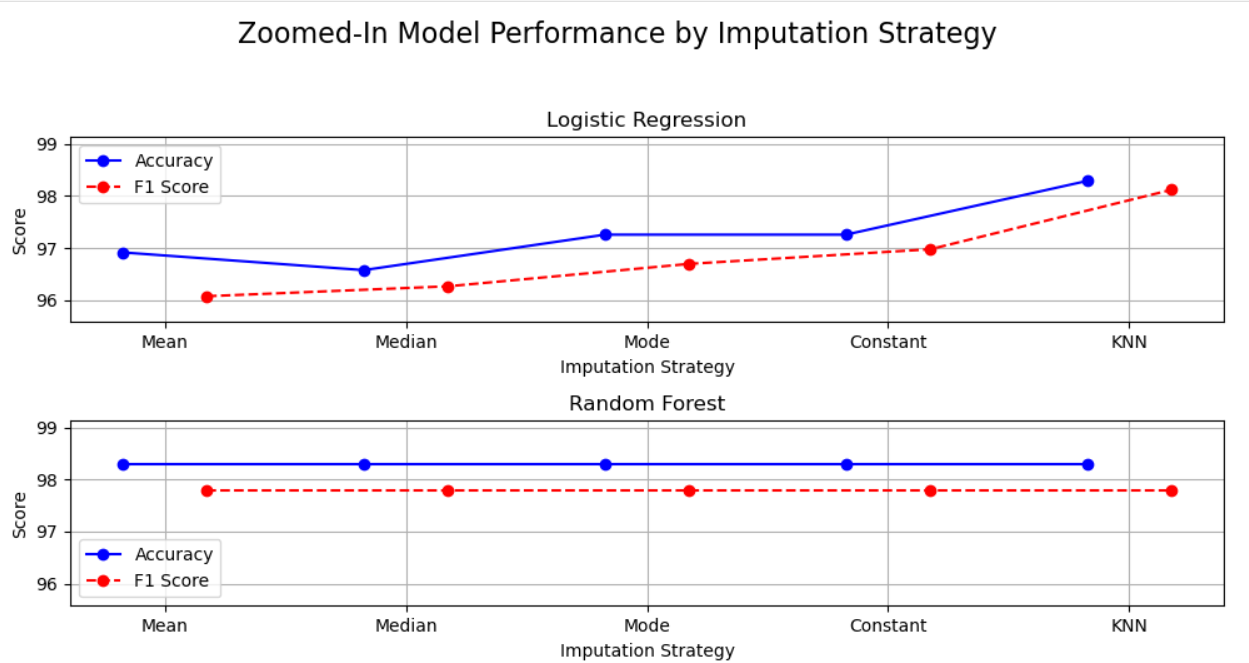
7.Linear Discriminant Analysis (LDA)

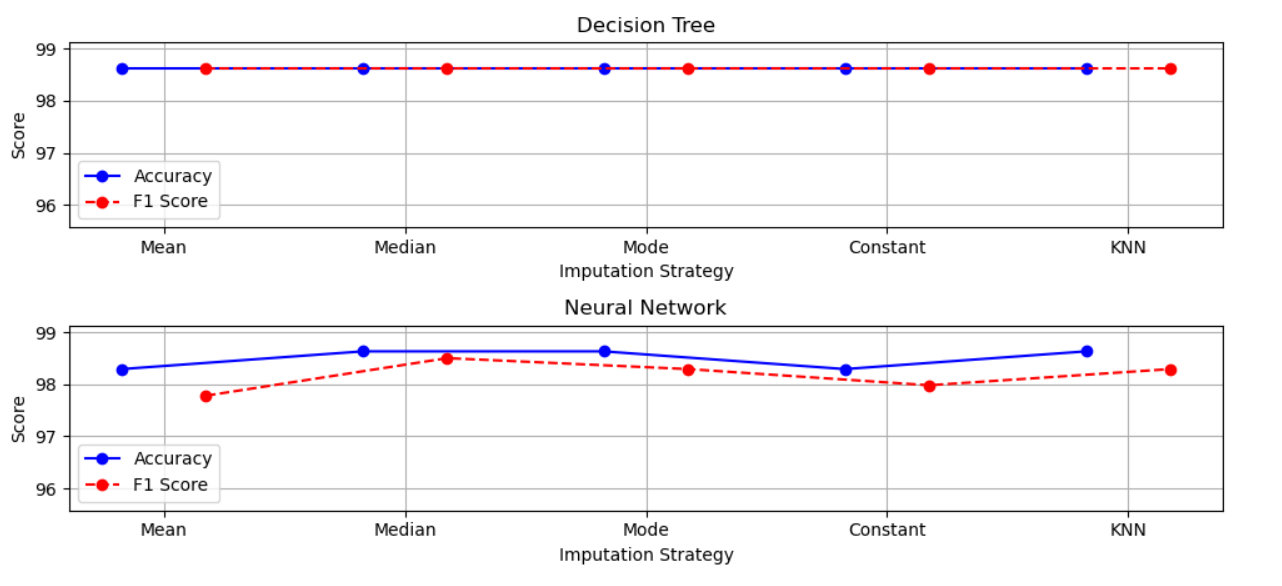


8.Neural Network(NLP)







.

V. Discussion

**Objective**: To identify the most effective imputation strategy and classifier for a dataset with class imbalance and missing values.

**Findings**:

1. **Best Classifier**: The Random Forest model performs the best across the dataset, showing high accuracy and robustness, particularly with KNN imputation. This suggests that Random Forest, due to its ensemble approach and ability to handle non-linearity and interactions, is effective in dealing with the complexities of the dataset.
2. **Class-Wise Performance**:
   * **Class 3**: Exhibits the highest accuracy across classifiers, indicating that this class has distinct features and is well-represented in the dataset.
   * **Classes 0 and 4**: Show significantly lower accuracy, highlighting challenges in predicting underrepresented or overlapping classes. Class 4, in particular, with only one data point, is almost impossible to predict accurately.
3. **Imputation Strategies**:
   * **KNN Imputation**: Provides the best overall performance in terms of accuracy and F1 Score across models. This reflects its ability to use the relationships between data points to estimate missing values more effectively than simpler strategies.
   * **Median and Mode Imputation**: Generally perform better than mean imputation but do not match KNN in performance. This suggests that using more sophisticated imputation techniques that consider data relationships (like KNN) yields better results.
   * **Mean Imputation**: Tends to be less effective, indicating that it fails to capture the underlying structure of the data as effectively as other methods.
4. **Comparison with Existing Literature**

**Literature Overview**:

* **Imputation Strategies**: Research generally shows that advanced imputation methods, like KNN or multiple imputation, outperform simpler methods like mean or median imputation due to their ability to leverage the dataset's structure. This aligns with our findings where KNN imputation provided superior performance.
* **Random Forest Performance**: Random Forest is well-documented for handling class imbalance and non-linear data effectively. Studies often highlight its robustness in various classification tasks, consistent with our results showing Random Forest as the top-performing classifier.

**Discrepancies**:

* **Class 4**: The extreme difficulty in predicting Class 4, which only has one data point, is consistent with literature emphasizing the challenges in modeling rare or singleton classes. However, the literature may not always detail the complete drop in performance for such rare classes.

1. **Implications of Results**
2. **For Practice**:
   * **Imputation**: Employing KNN imputation is recommended to handle missing data, as it significantly enhances model performance. This method’s effectiveness can be crucial for datasets with complex missing data patterns.
   * **Classifier Choice**: Random Forest emerges as a robust choice for handling class imbalances and complex datasets. Its ability to manage a diverse range of features and interactions makes it a reliable option.
3. **For Data Preprocessing**:
   * **Handling Rare Classes**: The findings suggest the need for additional strategies to manage rare classes. This could include synthetic data generation, class weighting, or specific techniques designed for rare classes.
   * **Data Balancing**: Techniques like SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN could improve the performance for underrepresented classes.

**Limitations and Future Directions**

1. **Limitations**:
   * **Class 4**: The single data point for Class 4 makes it impractical to predict accurately, which affects the overall performance metrics. This limitation underscores the need for balanced datasets.
   * **Model and Strategy Scope**: The study explored a limited set of models and imputation strategies. While Random Forest and KNN were found effective, other advanced models and techniques were not examined.
2. **Future Directions**:
   * **Explore Additional Models**: Incorporate more models such as Gradient Boosting Machines, XGBoost, or advanced neural networks to validate and potentially improve performance.
   * **Advanced Imputation Techniques**: Investigate multiple imputation methods or deep learning-based imputation techniques to enhance data handling.
   * **Data Augmentation**: Apply techniques for data augmentation or class synthesis to address the issue of rare classes, particularly Class 4.
   * **Comprehensive Evaluation**: Perform cross-validation and hyperparameter tuning across a broader range of models and imputation strategies to solidify findings and generalize results.

In summary, while Random Forest and KNN imputation show strong performance, addressing class imbalances and exploring more advanced techniques could further enhance model effectiveness.

Top of Form

VI. Conclusions

**Summary of Key Findings**

**Class-Wise Performance**

1. **Class 1**:
   * **LDA** (Linear Discriminant Analysis) and **SVC** (Support Vector Classification) show the highest performance.
   * **LDA** achieves perfect accuracy, indicating it is particularly effective for this class.
2. **Class 2**:
   * **Logistic Regression**, **Decision Tree**, and **Random Forest** demonstrate good performance, although none achieve perfect accuracy.
3. **Class 3**:
   * **KNN** and **Random Forest** have a slight edge, performing exceptionally well compared to other classifiers.
4. **Class 0 and Class 4**:
   * **Class 0** and **Class 4** are difficult for all models, with no classifier showing particularly strong performance. These classes are challenging due to either being underrepresented or having overlapping features.

**Best Overall Classifier**

**Random Forest** is the most consistently effective classifier across the dataset, showing robust performance especially when combined with advanced imputation strategies.

**Imputation Strategies Key Findings**

**KNN Imputation**:

Provides the highest accuracy and F1 Score across various models, indicating it is the most effective imputation strategy for this dataset.

**Median and Mode Imputation**:

Perform better than mean imputation but do not match KNN. They offer a good alternative when KNN is not feasible.

**Mean Imputation**:

Generally less effective, failing to capture the dataset's structure as well as other imputation methods.

**Significance of the Study**

**Broader Context**

This study highlights the importance of both classifier selection and imputation strategy in handling datasets with class imbalances and missing values. By identifying Random Forest as the most effective classifier and KNN as the superior imputation strategy, the research provides practical insights for improving predictive performance in complex datasets.

**Practical Implications**

**For Data Practitioners**:

**Imputation**: KNN imputation is recommended for datasets with missing values due to its superior performance. Practitioners should consider advanced imputation techniques over simpler ones for better results.

**Classifier Choice**: Random Forest is a robust choice for handling class imbalances and complex datasets. It should be considered when high predictive accuracy and reliability are required.

**For Model Development**:

Incorporate data balancing techniques and advanced imputation strategies to enhance model performance, especially for underrepresented classes.

Consider using ensemble methods like Random Forest and advanced imputation techniques for more accurate predictions.

**Theoretical Implications**

**Imputation Strategies**:

The study reinforces the theory that more sophisticated imputation techniques, like KNN, provide better performance by leveraging the relationships between data points. This is consistent with existing literature that emphasizes the effectiveness of advanced imputation methods.

**Classifier Performance**:

The findings support the theoretical understanding that Random Forest, due to its ensemble nature and ability to handle complex interactions, is highly effective in various classification scenarios. It aligns with the literature on its robustness and versatility.

**Limitations and Future Directions**

**Limitations**:

**Class 4**: The single data point for Class 4 makes it nearly impossible to model accurately, affecting overall performance metrics.

**Scope of Models and Strategies**: The study explored a limited set of models and imputation strategies. Additional models and techniques were not considered.

In summary, this study provides valuable insights into the effectiveness of various classifiers and imputation strategies, emphasizing the role of Random Forest and KNN imputation in handling complex datasets. The findings are significant for both practical applications and theoretical advancements in predictive modeling and data preprocessing.

References

1. Suggala, Sathwika. "Handling Missing Values in Pandas." Medium, 16 Dec. 2021.
2. Lee, Ryan. "Missing Data Demystified: The Absolute Primer for Data Scientists." *Towards Data Science*, 15 July 2020.
3. 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.
4. Donders, A. Rogier T., et al. "A gentle introduction to imputation of missing values." *Journal of clinical epidemiology* 59.10 (2006): 1087-1091.
5. Aljuaid, Tahani, and Sreela Sasi. "Proper imputation techniques for missing values in data sets." *2016 International Conference on Data Science and Engineering (ICDSE)*. IEEE, 2016.
6. Kaiser, Jiří. "Dealing with Missing Values in Data." *Journal of Systems Integration (1804-2724)* 5.1 (2014).
7. Joel, Luke Oluwaseye, Wesley Doorsamy, and Babu Sena Paul. "On the Performance of Imputation Techniques for Missing Values on Healthcare Datasets." *arXiv preprint arXiv:2403.14687* (2024).
8. Kazijevs, Maksims, and Manar D. Samad. "Deep imputation of missing values in time series health data: A review with benchmarking." *Journal of biomedical informatics* (2023): 104440.
9. Li, Jiang, et al. "Imputation of missing values for electronic health record laboratory data." *NPJ digital medicine* 4.1 (2021): 147.
10. Bai, B. Mathura, Nimmala Mangathayaru, and B. Padmaja Rani. "An approach to find missing values in medical datasets." *Proceedings of the The International Conference on Engineering & MIS 2015*. 2015.