International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05872
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Role of Data Quality in Machine Learning Performance

Nikita Tanwar
Class: M.Sc. Data Science SY

Introduction

Machine Learning (ML) has become an essential technology across diverse sectors, including healthcare, finance, and e-commerce. The effectiveness of ML models is influenced not only by the choice of algorithms but also significantly by the quality of the data on which they are trained. High-quality data enhances model accuracy and reliability, whereas poor-quality data can lead to misleading and inconsistent results. This study aims to examine the impact of data quality on machine learning performance and highlights its critical role in achieving robust and dependable model outcomes.

ProblemStatement

Many ML models fail to achieve expected performance due to poor data quality, including missing values, noise, and inconsistencies. Despite using advanced algorithms, the results remain inaccurate. This study aims to analyse how data quality issues impact ML models and how they can be improved.

Objective

- To analyse the impact of data quality on the performance of Machine Learning models.
- To identify common data quality issues such as missing, noisy, and inconsistent data.
- To evaluate how data preprocessing techniques improve model accuracy and reliability.
- To compare model performance using raw data versus cleaned data.
- To emphasise the importance of high-quality data in building efficient and robust ML systems.

Literature Review

Previous studies highlight that data quality plays a crucial role in Machine Learning performance. Researchers like Han et al. emphasise that real-world data is often incomplete and noisy, making data preprocessing essential. Goodfellow et al. state that high-quality data is necessary for model predictions, especially in deep learning. Accurate studies in areas such as fraud detection show that issues like imbalanced and inconsistent data negatively impact model accuracy. Research also suggests that techniques like data cleaning, normalisation, and feature selection significantly improve results. Overall, existing literature confirms that improving data quality is often more effective than using complex algorithms for achieving better ML performance

Data Collection

Data for this study were collected from reliable sources such as Kaggle and the UCI Machine Learning Repository. The dataset includes structured data relevant to machine learning tasks. Both raw and cleaned data were used to analyse the impact of data quality. The raw data contained issues like missing values and noise, while the cleaned data was prepared using preprocessing techniques for better analysis.

Research Methodology

This study follows an experimental approach to analyse the impact of data quality on machine learning performance. A dataset was collected from reliable sources and divided into two versions: raw data and preprocessed (cleaned) data. Data preprocessing techniques such as handling missing values, removing duplicates, normalisation, and encoding were applied to improve data quality. Multiple machine learning models, including Logistic Regression, Random Forest, XGBoost, and Neural Networks, were trained on both datasets. The performance of these models was evaluated using metrics such as accuracy, precision, recall, and F1-score. A comparative analysis was conducted to assess differences in model performance between raw and cleaned data. This methodology helps clearly identify how data quality affects the efficiency, accuracy, and reliability of machine learning models.

Results

The experimental results clearly demonstrate that data quality has a significant impact on the performance of machine learning models. Models trained on the preprocessed (cleaned) dataset consistently outperformed those trained on raw data. There was a noticeable improvement in accuracy, precision, recall, and F1-score after applying data cleaning and

preprocessing techniques. In contrast, models trained on raw data showed lower performance due to the presence of missing values, noise, and inconsistencies. Additionally, issues such as data imbalance negatively affected prediction accuracy, especially in classification tasks like fraud detection. The comparative analysis confirms that improving data quality leads to more reliable, accurate, and efficient machine learning models. These findings highlight the importance of proper data preprocessing in achieving optimal model performance.

Future Scope of Research

The findings of this study highlight several potential directions for future research in enhancing data quality for machine learning applications. Future work may focus on developing intelligent and automated data cleaning and preprocessing frameworks using artificial intelligence. Additionally, scalable approaches for handling large-scale and real-time data should be explored to ensure consistent and reliable data quality. Further research can examine the impact of data quality across diverse domains such as healthcare, finance, and cybersecurity. Integrating data quality assessment and monitoring mechanisms within machine learning pipelines can significantly enhance model robustness, reliability, and generalisation. Moreover, advanced techniques for handling imbalanced and noisy datasets should be investigated to improve the efficiency and performance of machine learning systems.

Limitation of Research

This study has certain limitations that may affect the generalizability of its findings. The analysis is conducted on a limited dataset, which may not fully capture real-world variability. Additionally, the study focuses on a selected set of machine learning models, and the results may differ with the application of alternative or more advanced algorithms. Furthermore, constraints related to time and computational resources restricted the scale of experimentation. The research primarily considers structured data, while the impact of data quality on unstructured data, such as text and images, remains unexplored. Therefore, the findings, although significant, may have limited applicability across diverse domains and data types.

References

- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics*. MIT Press.
- Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
- García, S., Luengo, J., & Herrera, F. (2016). Data preprocessing in data mining. *Springer*.
- IEEE Research Papers on Data Quality and Machine Learning.
- Kaggle (2023). Machine Learning Datasets. Available at: https://www.kaggle.com
- UCI Machine Learning Repository (2023). Available at: https://archive.ics.uci.edu