An Overview on Data Preprocessing and Data Cleaning Techniques with Their Impact on Model Performance.

TOUHID ALAM, 22-46330-1, 22-46330-1@student.aiub.edu,

American International University- Bangladesh (AIUB)

1. Introduction:

Machine learning is a part of AI where the machine learns from input data to do some tasks or to predict something or to make a logical decision without implementing any code. Nowadays, in every field of our day-to-day life Machine Learning has been used to make our life easier. Data Preprocessing is an important and crucial part of Machine Learning where data is prepared to be fed to a Machine Learning algorithm or model to make the algorithm work on any task or to predict. The following project is designed to discuss the importance of Data Preprocessing for Machine Learning, discuss some common techniques to Data Preprocessing and to discuss their impact on the performance of the machine learning algorithm or model.

2. Importance of Data Preprocessing:

The term Data Preprocessing in Machine Learning means preparing data to be used in Machine Learning Model. In today's world there are numerous numbers of data. Due to many reasons, like poor data quality, they cannot be directly used for Machine Learning or Data Mining [1]. There may be many missing values in the data, or the data of the dataset may not be feasible or consistent. Furthermore, with the introduction of Big Data in recent times, more dimensions have been introduced which tends to result in questioning the quality of collected data furthermore [2]. These problems can majorly impact the performance of the model. Data quality can be measured by investigating several factors such as completeness, timeliness, conformity, uniqueness, consistency or accuracy of the data. Data Preprocessing is not only important for Machine Learning. Ensuring the quality of data is majorly important to be processed furthermore in because the data could be related to so many important sectors from science to e-commerce [3]. Moreover, the machine learning model works based on the data fed to the model.

That is why, Pre-processing data and cleaning data is crucial for Machine Learning. Data Preprocessing is the process to make data usable for Machine Learning. The focus of Data Preprocessing is to make sure the data is ready for use and to ensure the data quality for the best possible result of the model [4]. It is considered as the core stage in Machine Learning due to the sensitivity of the nature it offers to the whole Machine Learning process [5].

3. Common Techniques of Data Preprocessing:

Usually, the Data Preprocessing is done in a structural way step by step. The process used for Data Pre-process may be very based on the data. Two common processes of Data Preprocessing are shown in Fig. 01 and Fig. 02.

Figure 01: An Approach of Pre-processing of a dataset [6]

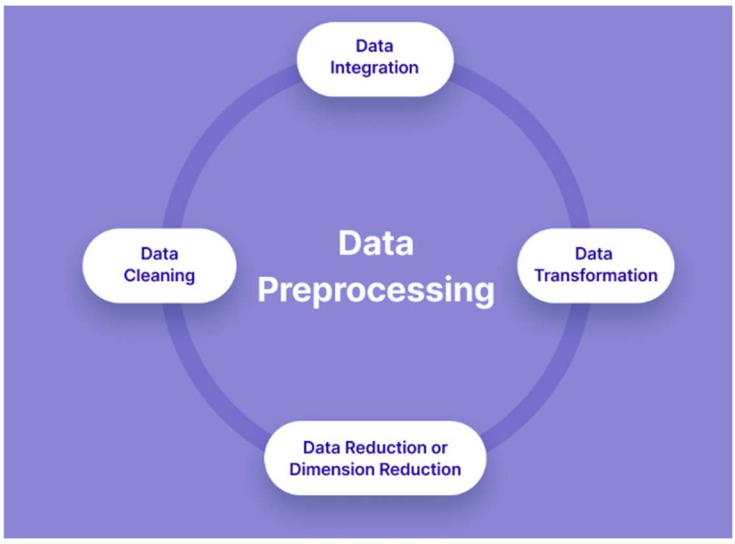

Figure 02: Common Steps of Data Preprocessing [7].

The project considered all the steps until model building in Data Preprocessing. Let's discuss some common techniques used for Data Preprocessing in Machine Learning such as Data Collection, Data Exploration, Data Cleaning, Feature Engineering, Data Balancing and Splitting Data.

3.1. Data Collection:

Data Collection can be considered as the most important part for building a Machine Learning model. The whole model is based on what kind of data we are working data. Unfortunately, there may be some issues with the collection of data. Roh et al. presented two major issues of data collection are:

- Insufficient labeled data for Machine Learning in new applications [8].
- Deep learning techniques requiring more labeled data [8].

There are some factors to consider while collecting data. Such as:

- Data much be indexed and published for sharing so that the validation of the data can be verified [8].
- The timeliness of the data.

There may be some problems with data collection such as less data collected and low recall rates, Data Sparseness [4]. Sometimes, more than one dataset may need together to build specific dataset for the purpose. Data Integration is the process of combining multiple datasets together. But several challenges need to be faced while doing the process such as Data Inconsistency, Inadequate Resources, Scope of Data [9]. One technique of Data Collection is generating data via Crowd-based techniques such as Active Learning and Crowdsourcing (survey) [8]. Moreover, there are many popular online database platforms for data science and machines for the sole purpose of finding and publishing datasets such as Kaggle [10].

3.2. Exploratory Data Analysis (EDA):

Data exploration can help to understand some characteristics of a dataset. Such as:

- What kind of data we are working with.
- Dimensions of the dataset or any attribute not related to the target attribute.
- If there are any missing values, outliers in the dataset.
- The value types of the attributes of the dataset.
- Data Visualization for understanding the patterns, trends and correlations within the dataset and can also help with decision making [11].

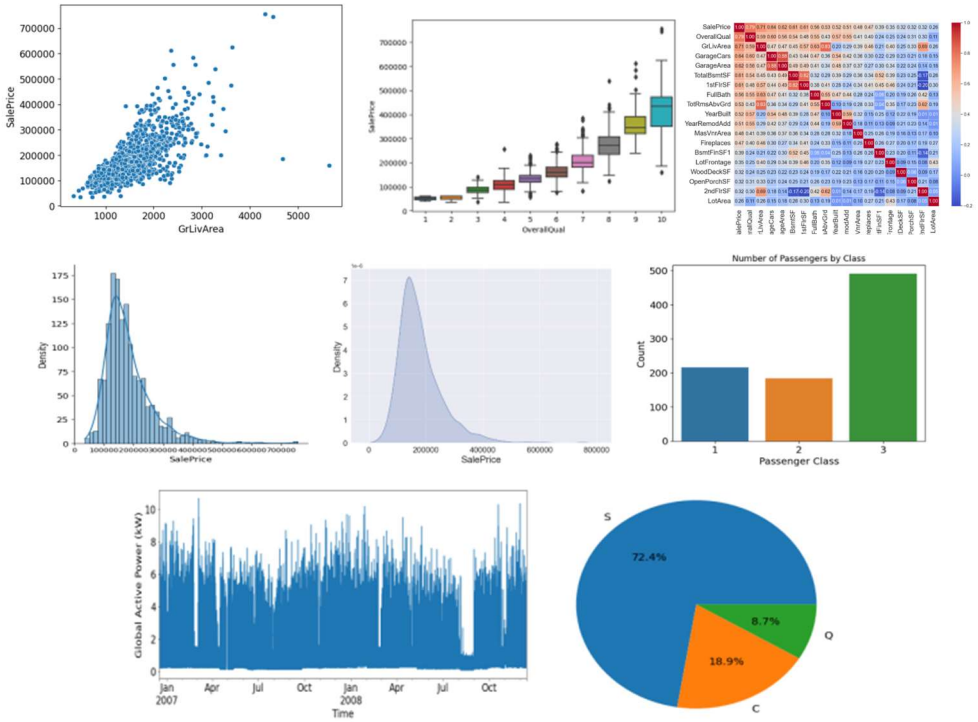

Figure 03: Some Common Data Visualization Techniques [11].

3.3. Data Cleaning

When we collect data, there is a possibility that there will be many problems with the dataset. For example, there may be missing values in the dataset. Missing value in dataset can occur for several reasons. Moreover, there may be some noisy values and outliers in the dataset. Noisy value is the random error input in the dataset and the outliers are the datapoints that are severely different from the rest of the values of datasets.

3.3.1. Missing Values:

There are different types of missing values. Figure 03 shows the three different types of missing values MCAR, MAR and NMAR.

Missing Completely at Random (MCAR): if the missing variable's probability does not relate to other variables [12].

Missing at Radom (MAR): if the missing variable's probability depends on the information available [12].

Not Missing at Random (NMAR): if the missing variable's probability depends on the variable itself [12].

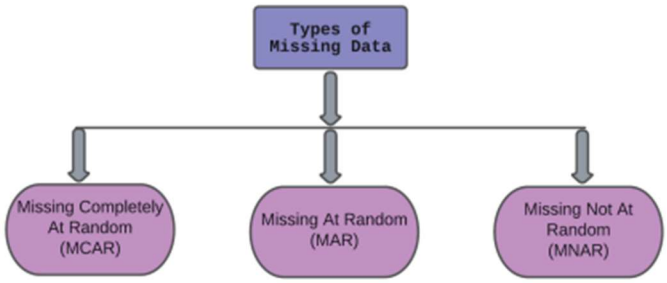

Figure 04: Types of missing values [6].

The conventional method of dealing with missing value is ignoring or deletion but it can severely impact the performance of the method used [12]. That's why recovering these missing values is important for better machine performance. Pratama et al. discussed several missing value recovery systems such as Mean/Mode Imputation, Hot and Cold Deck Imputation, Autoregressive, Genetic Algorithm Optimization Based Algorithm, Support Vector Machine, Interpolation, Fuzzy-rough set and Maximum Likehood [12]. The method required to recover these missing values depends on the nature of the missing value of the dataset.

3.3.2. Noise Values:

Noise values are the anomalies of the dataset. Abnormalities and inaccuracies in performance are common problems that occurred because of noise values [12]. Pratama et al. suggested that the

data should be filtered as null check and type anomalies for discovering noise values. [12]. Using Regression and clustering these values can be recovered. Clustering helps to identify similar data in groups.

3.3.3. Outliers:

Outliers are the values that are severely different from the rest of the dataset in ranging. One common way of detecting outliers is the Boxplot Technique. The outlier values can be recovered with the Median Imputation technique because doing mean imputation may cause biasness in recovering the value.

3.4. Data Labeling:

Data Labeling means to determine which attributes of the dataset will be the feature and which attribute of the dataset will be the output or the target. The whole model working method is based on determining the target based on the features. As a result, it is important to determine the features and the target or output properly.

3.5. Feature Engineering:

Feature Engineering refers to the process of creating new features or deleting existing features or transforming existing features based on the nature of the dataset. Some common step of feature engineering is Feature Reduction using PCA or LDA, Creating New Feature based on the nature of the data and Transformation the data of the features. The project discusses some simple approaches of Feature Engineering but in broad, feature engineering is very big.

3.5.1. Data Transformation:

Data Transformation refers to the process of manipulating the data for the better output of the performance. It is a common step of Feature Engineering. There can be some mandatory data transformation such as transforming the non-numeric attributes in numeric attributes and resizing the inputs to fixed sizes [10]. Data Transformation is commonly used to reduce the skewness of the data.

3.5.1.1. Normalization:

Normalization is the process of converting the features in a particular range for better performance. Normally in range of [0,1]. Some common approaches of normalization are min max method, log scaling. Normalization techniques use Quantile to normalize the data. Obiad et al. used some normalization method and a model based on J48 CLASSIFIER to show how much time some normalization methods take to pre-process the data showing their effectiveness and they used the NSL-KDD benchmark dataset. Table -1 shows the result of their research.

Table 1. The accuracy and consumed time by the pre-processing normalization methods [5].

Pre-Processing Method	Classification Accuracy (%)	Time for Model Building (Sec)	Time for Model Testing (Sec)
Log2	99.66	16.4	0.12
Min-max	99.66	17.7	0.06
Z-score	99.66	21.05	0.04
Decimal Scaling	99.66	18.43	0.13

3.5.1.2. Standardization or Scaling:

The term Scaling means adjusting based on having a mean of 0 and a standard deviation of 1. It is mainly achieved using the method z-score. The method is mainly used when the data has a Gaussian distribution. Standardization can be considered as a part of the Normalization Technique.

3.5.1.3. Encoding Categorical Data:

The categorical Variable sometime needs to be converted into unique numeric value. This is very much applicable for the output variable. One example of Encoding Categorical Data is One Hot Encoding where only one value of the categorical attribute gets converted in 1 and rest are 0.

3.5.2. Dimensionality Reduction:

Dimensionality Reduction is the process of reducing the number of features in a dataset. Having large number of features can take a lot of memory in the system or there maybe some features which do not have any correlation with the target attribute or the output. That's why dimensionality reduction is important.

3.5.2.1. Principal Component Analysis (PCA):

PCA is a common method to reduce the number of features based on the variance in the data using principal component (k). The principal component is the new set of dimensions.

3.5.2.2. Linear Discriminant Analysis (LDA):

LDA is another common method to reduce the number of features used to reduce the number of features in a dataset when there is a need to reduce the time complexity of the model and for a better classification [5].

In their research, Obaid et al. also showed the impact on the accuracy of the model when different number of features are used for the model using different dimensionality reduction methods [5]. Table 2 and Table 3 show the result of their study.

Table 2: Dimensional Reduction Results using LDA [5].

No. of features	Classification accuracy (%)	Time for model building (sec)	Time for model testing (sec)
5	53.36	0.02	0.12
10	99.08	0.04	4.6
20	99.11	0.04	6.72
30	98.99	0.04	10.57
40	99.02	0.04	15.21

Table 3: Dimensional Reduction Results using PCA [5].

No. of features	Classification accuracy (%)	Time for model building (sec)	Time for model testing (sec)
5	99.28	0.03	2.78
10	99.50	0.04	5.12
20	99.52	0.04	12.69
30	98.53	0.04	17.46
40	99.53	0.04	22.47

3.6. Data Balancing:

Another important step of Data Preprocessing is Data Balancing. Data Balancing refers to the process of readjusting the distribution of the data. It is important for specially dataset with categorical target where one category maybe much more than the other category, thus creating biasness in the performance of the model. Some common techniques of data balancing are resampling data and synthetic samples [13].

3.6.1. Resampling Data:

Resampling means either oversampling the target attribute with lower number of instances or under sampling the target attribute with higher number of instances. The purpose of resampling data is to avoid the biasness of model with reaching equal number of instances for the target attribute [13].

3.6.2. Synthetic Samples:

Synthetic Samples are the sample data that are created artificially for balancing the dataset. One of the popular algorithms of creating synthetic instances is Synthetic Minority Over-Sampling Techniques (SMOTE). Shaer et al. approaches to handle imbalance data with SMOTE and other techniques where they found machine gives 18% relative error rate for the imbalance data and only 5% relative error rate in predicting the yield estimation using SMOTE [14]. Table 4 shows their result.

Table 04: Yield Prediction Error [14].

Relative Yield Error (for appreciable yield > 10%) Abs((Yield_modelBased-Yield_CktSim)/ Yield_CktSim)*100			
	Imbalanced	**SMOTE**	**SMOTE + borderline**
Average	18.20	5.70	20.28
Maximum	50.29	20.09	80.04

3.7. Splitting Data:

Splitting Data refers to the process of separating the data into three parts: for training the model, for testing the model and to validate the model. Typically, we split the data in 80-20 for training and testing. But in Big data the ration for training data needs to be much higher. Splitting is important to test if there is overfitting or underfitting in the model via the testing the validation and test data. Also, validation data is important for tuning hyperparameters.

4. Discussion:

Data Preprocessing is the most crucial part of Machine Learning. Data Preprocessing determines whether the processed data with will work synchronously with the machine model. Data Preprocessing is processed with some structured steps. Based on the data, it is important to choose which Pre-processing technique to apply. In this project, some common and basic way to Pre-process data has been discussed. The impact of these Pre-processing techniques was also discussed based on some previous studies in Data Preprocessing. Overall, this project gives a basic idea of some common techniques of Data Preprocessing and their impact on the data we are working with.

5. References:

[1] S. Roy, P. Sharma, K. Nath, D. K. Bhattacharyya, and J. K. Kalita, "Pre-Processing: A Data Preparation Step," in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 463–471. doi: 10.1016/B978-0-12-809633-8.20457-3.

[2] D. Rao, V. N. Gudivada, and V. V. Raghavan, "Data quality issues in big data," in *2015 IEEE International Conference on Big Data (Big Data)*, Oct. 2015, pp. 2654–2660. doi: 10.1109/BigData.2015.7364065.

[3] A. Juneja and N. N. Das, "Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Feb. 2019, pp. 559–563. doi: 10.1109/COMITCon.2019.8862267.

[4] P. Zhang, F. Xiong, J. Gao, and J. Wang, "Data quality in big data processing: Issues, solutions and open problems," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2017, pp. 1–7. doi: 10.1109/UIC-ATC.2017.8397554.

[5] H. S. Obaid, S. A. Dheyab, and S. S. Sabry, “The Impact of Data Preprocessing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning,” in *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, Mar. 2019, pp. 279–283. doi: 10.1109/IEMECONX.2019.8877011.

[6] M. Rahman, M. Hasan, M. M. Billah, and R. J. Sajuti, “Grading System Prediction of Educational Performance Analysis Using Data Mining Approach,” *Malaysian Journal of Science and Advanced Technology*, pp. 204–211, Nov. 2022, doi: 10.56532/mjsat.v2i4.96.

[7] “Data Preprocessing in Machine Learning [Steps & Techniques].” Accessed: Jun. 13, 2024. [Online]. Available: https://www.v7labs.com/blog/data-preprocessing-guide, https://www.v7labs.com/blog/data-preprocessing-guide

[8] Y. Roh, G. Heo, and S. E. Whang, “A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, Apr. 2021, doi: 10.1109/TKDE.2019.2946162.

[9] A. Kadadi, R. Agrawal, C. Nyamful, and R. Atiq, “Challenges of data integration and interoperability in big data,” in *2014 IEEE International Conference on Big Data (Big Data)*, Oct. 2014, pp. 38–40. doi: 10.1109/BigData.2014.7004486.

[10] “Kaggle: Your Machine Learning and Data Science Community.” Accessed: Jun. 13, 2024. [Online]. Available: https://www.kaggle.com/

[11] C. Yalim and H. Handley, “The Effectiveness of Visualization Techniques for Supporting Decision-Making,” *Modeling, Simulation and Visualization Student Capstone Conference*, Apr. 2023, [Online]. Available: https://digitalcommons.odu.edu/msvcapstone/2023/datascience/1

[12] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani, “A review of missing values handling methods on time-series data,” in *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*, Oct. 2016, pp. 1–6. doi: 10.1109/ICITSI.2016.7858189.

[13] “Handling Imbalanced Datasets in Machine Learning Projects.” Accessed: Jun. 13, 2024. [Online]. Available: https://www.linkedin.com/pulse/handling-imbalanced-datasets-machine-learning-projects-sandesh-patil

[14] L. Shaer, R. Kanj, and R. Joshi, “Data Imbalance Handling Approaches for Accurate Statistical Modeling and Yield Analysis of Memory Designs,” in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2019, pp. 1–5. doi: 10.1109/ISCAS.2019.8702731.