# APPLICATION OF DATA PREPROCESSING IN HEALTH CARE

**Rohith S, Dr. S. Anupama Kumar**

VI[th] Semester, Associate Professor

Department of MCA,

RV college of Engineering,

Bangalore, Karnataka.

## Abstract

Internet is being used by many of user everyday as a result of this huge amount of data is being generated constantly. Mining and analyzing such data will help for an organization for research and development. Data preprocessing will play an important role in the data mining and analysis. This paper is intending to bring an insight to the importance of preprocessing and various preprocessing techniques that can be used in real world applications. Health care data is often incomplete, and the data set is not in the proper format to apply machine learning techniques. Data preprocessing eliminates the missing data and the data will be set in the effective integrated for the purpose of implementation. The different data preprocessing functions that brings in meaning to the raw data and prepare the data for further process are discussed in this paper.

**Keywords:** Machine learning, data-preprocessing, missing data, feature, scaling.

## 1. Introduction

Data preprocessing is important procedure for data mining. which includes data cleaning, data transformation and feature selection. Data cleaning and transformation are methods used to remove outliers and standardize the data so that they take a form that can be easily used to create a model. The value of machine learning in healthcare has the ability to process huge datasets beyond the scope of human capability, and then reliably convert analysis of that data into clinical insights that aid physicians in planning and providing care, ultimately leading to better outcomes, lower costs of care, and increased patient satisfaction.

Resource allocation in hospitals are very difficult these days using the machine learning algorithms. The proposed work is a prediction system to analyze the allocation of resources in hospital operations. The data set for this work is collected in the form of different spread sheets and need to be cleansed, transformed and integrated for further process. Therefore, it becomes necessary to understand the data better to bring out useful meaning from it. The data set consists of lot of noise and missing values and need to be handled properly. This paper mainly concentrates on the different ways on elimination of missing values and the various pre processing techniques.

Before starting handling missing values, it is important to identify the missing values and know with which value they are replaced. This can be handled by combining the metadata information with exploratory analysis. This helps to understand the entries and decide whether to keep that entry or to neglect.

Second method is encoding categorical data by encoding the data in order to bring it to such a state that the machine now understands it. Feature encoding is basically performing transformations on the data such that it can be easily accepted as input for machine learning. Splitting data into training and test set.

## 2. Literature survey

In [1], the authors J. Srivastava, P. Desikan and V. Kumar have given brief overview about the data mining and its applications in various domains.

The information about web usage mining is the application of data mining techniques to large web data repositories in order to produce results that can be used in the design tasks such as web site design, web server design is explained by [2].

The paper [3] contains the detailed information about the use of the machine learning algorithms and deep learning neural networks.

Definition, model, development stage, classification and commercial application of machine learning, and emphasizes the role of machine learning in data mining are explained by the authors Teng Xiuyi1, Gong Yuxia1 [4].

The survey research article [5] prefers two famous supervised machine learning algorithms that is decision trees support vector machine and presented the recent works carried out.

The authors S.B. Kotsiantis describes various supervised machine learning classification techniques [6].

Author Rob Law, reports on a study about applying neural networks to the forecast of room occupancy rates. The significance was tested with the real world data set [7].

The authors Zhongsheng Hua, Bin Zhang, are explaining how the support vector machines (SVMs) are adapted to forecast occurrences of nonzero demand of spare part, and a hybrid mechanism for integrating the SVM forecast results and the relationship of occurrence of nonzero demand with explanatory variables is proposed [8].

Authors Real Carbonneau, Kevin Laframboise, Rustam Vavilov, are investigate the applicability of advanced machine learning techniques, including neural networks, recurrent neural networks, and support vector machines, to forecasting distorted demand at the end of a supply chain [9].

Authors Kristin P. Bennett, Emilio Parrado-Hernandez explains the Large Scale Optimization and Machine Learning emphases on the core optimization problems fundamental machine learning algorithms. Seek to inspect the interface of state-of-the-art machine learning and mathematical programming [10].

## 3 Proposed system

The work is conducted using the data, which is collected from a hospital. Resource allocation in the health care sector is difficult since the existing data will be available in various forms and formats. Therefore, it becomes necessary to cleanse, transform and then integrate it. The following figure1 is the proposed system to implement the various pre processing techniques.

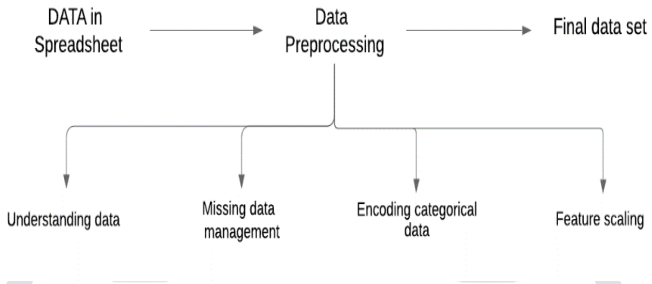



**Figure 1: Proposed System**

The below mentioned methods are used in the proposed system and prepare it for further process.

**(i)Understanding of data:** Choosing right data for machine learning is important. data set should be understood before proceeding further. To apply any of the machine learning algorithm the data set should have dependent and independent variable. Independent variables are the input for a process that is being analyzes. Dependent variables are the output of the process. So, for that dataset understanding is much needed based on the variable can choose which data value should be preprocessed. The following table 1 is the list of attributes identified as a data set.

Table 1: Data set

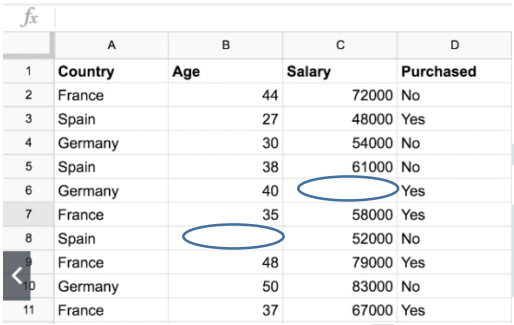
| Sl No | Attribute | Data Type |
|---|---|---|
| 1 | Country | String |
| 2 | Age | Number |
| 3 | Salary | Number |
| 4 | Purchased | String |

Consider the attributes for the following example here there are 4 attribute Country, Age, Salary, Purchased (whether the patient has purchased an insurance or not). In the above table 1 it is necessary to choose a dependent and independent variable. Independent variables are the input for a process that is being analyzed. Dependent variables are necessary for the working towards the output of the process. Age and Salary can be taken as independent variable based on those two variables input dependent variable can be predicted that is based on Age and Salary of the person whether he/she will purchase the product or not can be predicted. Prediction can be made with machine learning regression algorithms. Choosing independent and dependent variable is very much important for that data set should be understood properly. Following methods will specify how the raw data should be converted to clinical data.

**(ii) Missing data:** Real is data is often incomplete. Data missing will happen usual, which may be happened during data collection. Missing data can be handled by predicting the mean of the data. So, the data will be efficient to apply algorithm. Consider an example.

```
dataset$Column = ifelse(is.na(dataset$Column),
avg (dataset$Column, FUN = function(x) mean(x, na.rm = TRUE)),
dataset$Column)
```

These functions are used in the dataset where the column contains missing data and is checked with is.na() function if there are any missing data that can be handled with by considering average of that row by the mean. The following figures 2 and 3 shows the original and preprocessed data set that handles the missing values



| | A | B | C | D |
|---|---|---|---|---|
| 1 | Country | Age | Salary | Purchased |
| 2 | France | 44 | 72000 | No |
| 3 | Spain | 27 | 48000 | Yes |
| 4 | Germany | 30 | 54000 | No |
| 5 | Spain | 38 | 61000 | No |
| 6 | Germany | 40 | | Yes |
| 7 | France | 35 | 58000 | Yes |
| 8 | Spain | | 52000 | No |
| 9 | France | 48 | 79000 | Yes |
| 10 | Germany | 50 | 83000 | No |
| 11 | France | 37 | 67000 | Yes |

Figure 2: Original Data set

| | Country | Age | Salary | Purchased |
|---|---|---|---|---|
| 1 | France | 44.00000 | 72000.00 | No |
| 2 | Spain | 27.00000 | 48000.00 | Yes |
| 3 | Germany | 30.00000 | 54000.00 | No |
| 4 | Spain | 38.00000 | 61000.00 | No |
| 5 | Germany | 40.00000 | 63777.78 | Yes |
| 6 | France | 35.00000 | 58000.00 | Yes |
| 7 | Spain | 38.77778 | 52000.00 | No |
| 8 | France | 48.00000 | 79000.00 | Yes |
| 9 | Germany | 50.00000 | 83000.00 | No |
| 10 | France | 37.00000 | 67000.00 | Yes |

Figure 3: Handled Missing values

In the above figure the mean data column as mention in the above code handles the missing data.

**(iii) Encoding categorical data:** If the data contains categories since machine learning models based on mathematical equation it can understand that if we keep text in categories which will cause problems that why we need to encode text into number. With the example in R.

```
dataset$Column = factors(dataset$Column,
        levels = c('text1','text2','text3')
        labels = c(1, 2, 3)
```

In this function the column of a dataset which contains the characters in the columns will be converted as a number so that the applying of machine learning algorithm will be easier which will not face problems. So, the text1 will be converted into 1 similarly text2 to 2 and text3 to 3.

| Country | Age | Salary | Purchased |
|---|---|---|---|
| France | 44 | 72000 | No |
| Spain | 27 | 48000 | Yes |
| Germany | 30 | 54000 | No |
| Spain | 38 | 61000 | No |
| Germany | 40 | 63777.77778 | Yes |
| France | 35 | 58000 | Yes |
| Spain | 38.77777778 | 52000 | No |
| France | 48 | 79000 | Yes |
| Germany | 50 | 83000 | No |
| France | 37 | 67000 | Yes |

Figure 3: Original Dataset

| Country | Age | Salary | Purchased |
|---|---|---|---|
| 1 | 44.00000 | 72000.00 | 0 |
| 2 | 27.00000 | 48000.00 | 1 |
| 3 | 30.00000 | 54000.00 | 0 |
| 2 | 38.00000 | 61000.00 | 0 |
| 3 | 40.00000 | 63777.78 | 1 |
| 1 | 35.00000 | 58000.00 | 1 |
| 2 | 38.77778 | 52000.00 | 0 |
| 1 | 48.00000 | 79000.00 | 1 |
| 3 | 50.00000 | 83000.00 | 0 |
| 1 | 37.00000 | 67000.00 | 1 |

Figure 4: Categorical Value

In above figure figures country and purchased are categorized into characters to numeric.

**Feature Scaling:** Data set have entries which have different scale level that will cause some issue in applying algorithm because most of the machine learning algorithm follows the Euclidean Distance. If scales are different then the Euclidean distance may be dominated by one other and the least dominant will be removed to avoid that issue scaling is needed.

dataset [, 2:3] = scale (dataset [, 2:3])

In this function dataset which contains multiple data columns which may have data of different scale feature scaling will help in scaling them up. In the above function the dataset which include multiple columns of data but for the purpose of applying algorithm column 2 and 3 need to be scaled up. Which is possible by the above function using R.

| Country | Age | Salary | Purchased |
|---|---|---|---|
| France | 44 | 72000 | No |
| Spain | 27 | 48000 | Yes |
| Germany | 30 | 54000 | No |
| Spain | 38 | 61000 | No |
| Germany | 40 | 63777.77778 | Yes |
| France | 35 | 58000 | Yes |
| Spain | 38.77777778 | 52000 | No |
| France | 48 | 79000 | Yes |
| Germany | 50 | 83000 | No |
| France | 37 | 67000 | Yes |

Figure 5: Original Data set

| | | | |
|---|---|---|---|
| 1 | 0.90101716 | 0.9392746 | 0 |
| 2 | -1.58847494 | -1.3371160 | 1 |
| 3 | -1.14915281 | -0.7680183 | 0 |
| 2 | 0.02237289 | -0.1040711 | 0 |
| 3 | 0.31525431 | 0.1594000 | 1 |
| 2 | 0.13627122 | -0.9577176 | 0 |
| 1 | 1.48678000 | 1.6032218 | 1 |
| 1 | -0.12406783 | 0.4650265 | 1 |

Figure 6: Scaled dataset

Age and the salary are in different scale now in feature scaling both of them are scaled in same scale.

## 4 Result

In this in this article, different data preprocessing methods were proposed for converting raw clinical data to a format that is acceptable to model learning algorithms. The methods were implemented using R. The raw data can be better understood by applying the various statistical techniques. The application of mean method can be used to replace the missing values which can be handled efficiently. The coversion of numerical values to categorical values will help the application of machine learning algorithms in a larger scale. Feature scaling is important to normalize the data and bring out better understanding of the data set. By using these methods and policies, the manual work required for this process can be reduced and information from various sources can be used in a systematic way. Machine learning algorithms can be applied over this data set in future for optimizing the operational resources in health care sector.

## References

[1] J. Srivastava, P. Desikan and V. Kumar, "Web Mining – Accomplishments and Future Directions", National Science Foundation Workshop on Next Generation Data Mining, pp. 51-56, 2002.

[2] R. Colley, B. Mobasher and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", Journal of Knowledge and Information Systems Springer, vol. 1, no. 1, pp. 5-32, 1999.

[3] Mariette Awad, Rahul Khanna. "Efficient Learning machines: Concepts and Applications". Aspress Publishers, 2015.

[4] Teng Xiuyi1, Gong Yuxia1. "Research on Application of Machine Learning in Data Mining". IOP Conf. Series: Materials Science and Engineering, 2018.

[5] M. Praveena, V. Jaiganesh, "Literature Review on Supervised Machine Learning Algorithms and Boosting Process". International Journal of Computer Applications, ISSN No. 0975 – 8887, vol. 169, 2017.

[6] S.B. Kotsiantis. "Supervised Machine Learning: A Review of Classification Techniques", Informatica. pp 249-268, 2007.

[7] Rob Law, "Room occupancy rate forecasting: a neural network approach", International Journal of Contemporary Hospitality Management, vol. 10 Issue 6, pp 234 – 239, 1998.

[8] Zhongsheng Hua, Bin Zhang, "A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts".

[9] Real Carbonneau, Kevin Laframboise, Rustam Vahidov, "Application of machine learning techniques for supply chain demand forecasting " , European Journal of Operational Research 184, pp 1140 1154, 2008

[10] Kristin P. Bennett, Emilio Parrado-Hernandez "The Interplay of Optimization and Machine Learning Research" Journal of Machine Learning Research 7 (2006) 1265–1281 Submitted 7/06; Published 7/06.

[11] "Machine Learning A-Z™: Hands-On Python & R In Data Science" Udemy course.

[12] F. R. Bach, G. R. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In Proceedings of the Twenty-first International Conference on Machine Learning, 2004.

[13] M. Bazaraa, H. Sherali, and C. Shetty. Nonlinear Programming Theory and Algorithms. Wiley, 2006.

[14] K. P. Bennett and O. L. Mangasarian. Bilinear separation of two sets in n-space. Computational Optimization & Applications, 2:207–227, 1993.