TELKOMNIKA, Vol.17, No.4, August 2019, pp.2076~2086
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v17i4.12780

AUTO-CDD: automatic cleaning dirty data using machine learning techniques

Jesmeen M. Z. H.*[1], Abid Hossen[2], J. Hossen[3], J. Emerson Raja[4],
Bhuvaneswari Thangavel[5], S. Sayeed[6], Tawsif K.[7]
[1,3,4,5,7]Faculty of Engineering and Technology, Multimedia University, Melaka, 75450, Malaysia
[2]Department of Computer Science and Engineering, Khulna University Bangladesh, India
[6]Faculty of Information Science and Technology, Multimedia University, Melaka, 75450, Malaysia
*Corresponding author, e-mail: jesmeen.online@gmail.com

Abstract

Cleaning the dirty data has become very critical significance for many years, especially in medical sectors. This is the reason behind widening research in this sector. To initiate the research, a comparison between currently used functions of handling missing values and Auto-CDD is presented. The developed system will guarantee to overcome processing unwanted outcomes in data Analytical process; second, it will improve overall data processing. Our motivation is to create an intelligent tool that will automatically predict the missing data. Starting with feature selection using Random Forest Gini Index values. Then by using three Machine Learning Paradigm trained model was developed and evaluated by two datasets from UCI (i.e. Diabetics and Student Performance). Evaluated outcomes of accuracy proved Random Forest Classifier and Logistic Regression gives constant accuracy at around 90%. Finally, it concludes that this process will help to get clean data for further analytical process.

Keywords*: classification, data cleaning, dirty data, feature selection, gini index, random forest*

1. Introduction

Data quality is generally described as "the capability of data to satisfy stated and implied needs when used under specified conditions" [1]. Data accuracy, completeness, and consistency are most popular initiatives to address Data quality [2, 3], besides other dimensions like Accessibility, Consistent representation, timeliness, understandability, Relevancy, etc. [2]. Moreover, data quality is a combination of data content and form. Where data content must contain accurate information and data form essential be collected and visualized in an approach that creates data functioning. Content and form are the significant consideration to reduce data mistakes, as they illuminate the task of repairing dirty data needs beyond simply providing correct data.

Likewise, while developing a scheme to enhance quality of data it is essential to classify the primary reasons for causing data to be dirty [4, 5]. The causes are categories into organized and unintentional errors. Basic sources of producing systematic errors include while programming, the wrong definition for data types, rules not defined correctly, data collection's rules violation, badly defined rules, and trained poorly. The sources of random errors can be errors due to keying, unreadable script, data transcription complications, hardware failure or corruption, and errors or intentionally misrepresenting declarations on the portion of users specifying major data. Human role on data entry usually result in an error, this error can be typos, missing types, literal values, Heterogeneous ontologies (i.e. Different nature of data), outdated values or Violations of integrity constraints.

The system becomes very complex on implementing data cleaning process while processing data from heterogeneous sources. However, ignoring the process in data analytics may cause economic costs. Results obtained from the survey in 2014, that due to dirty data around 13 million dollars were costs annually in an organization and around 3 trillion per year was calculated in US economy. Another estimation of 6.8 Billion dollars to 1.5 Billion dollars spent on bad data management in US Postal service [6]. In medical case, these dirty data have ability to kill patients or induce damage to health of the patient which may be long-lasting issue. This bad data not only effects economical costs, it also may cost human life, such as in 1999 an

Received October 7, 2018; Revised February 5, 2019; Accepted March 3, 2019

institute of Medicine reported [7] calculations that minimum 44,000 to 98,000 patients had to lose their lives every year for medical data errors.

In the case of Iot Applications, most of the data are electronically collected, which may have serious data quality problems. Classic data quality problems mainly come from software defects, customised errors, or system misconfiguration. Authors in [8] discussed about cleaning data obtained from sensors. Here other method with ARIMA method was compared and they concluded that with a lower noise ratio, better results were obtained compared to higher noise ratio. The main advantage of their method is that it can work with huge data in a streaming scenario. However, if the data set is batch data it will not perform as expected.

In [9], the problem of cleaning is overcame using DC-RM model, where it supports better Pre-processing and Data Cleaning, Data Reduction, and Projection phases. If the data set contains missing values, the format of missing values was prepared and imputed. In cleaning phase performing removal of unwanted and undesired data is required with elimination of the rows which contains null data [10].

Eliminating data redundancy which usually available in different datasets on same datasets. These data redundancy can cause to database system defection and increase the unwanted cost of transmitting data. These defects can be useless occupying storage space, reducing data reliability, leads to higher data inconsistency, and destroying data. Hence, different reducing techniques were proposed for data redundancy, for example data filtration, data redundancy detection, and data compression. These techniques may be applicable to various data sets. However, it may also bring negative issues, such as compressing data and then decompressing those data may lead to additional computational load. Hence, it is important to balance the process and the cost. An author also indicates that after data collection process cleansing data is compulsory according to previous different datasets can be handled [11].

Research Gap. Usually multiple manual scrubbing process is executed to overcome and solve the poor data issues. This often involves more processing time and human resources. This results in slowing down any company operation performances and leave less time for analysing and optimising program. It increases cost for leads involving revenue reduction and profit margin. The issue will be solved if the cleaning phase is automatic. The tools available in market, are third party application. However, if the DA process is implement by using programming language it is important to make this process as fast and accurate as possible. Here, a predictive model will be useful to impute accurate missing data.

Problem Statement. In Data Analytics (DA) processing, data cleaning is most important and essential step. Inappropriate data may lead to poor analysis and thus yield unacceptable conclusions [12]. Some authors [13-16] ocused on the problem of duplicate identification and elimination. Their research focused on data cleaning partially and hence received only little attention in the research community. Different information system required to repair data using different rules. It is first required to overcome the dirty data dimensions from the structured data for better DA process. Data cleaning is the process of overcoming dirty data dimensions; such as incompleteness (missing values), duplication, inconsistency, and inaccuracy. Under these requirements, researchers developed tools to detect and repair Data Quality issues by specifying different rules between data, and normally different dimension issues requires different techniques, e.g., imputing missing value in the multi-view and panoramic dispatching [17]. There is scope for research in achieving better data cleaning. It can be achieved by introducing automatic data cleaning process with the help of Machine Learning (ML). Sampling technique is also integrated into the process considering the size of data. Because of the ML ability, the Auto-CDD system can learn from the data and predict the missing class in order to perform Automatic Missing Value Imputation. It is also required to select the suitable features for the suitable ML models automatically, depending on the form of the data set obtained from various domain. These abilities of data cleaning process can enhance the performance of DA, by replacing the current manual data cleaning with an intelligent one.

In the report [18], it has analysis of data issues obtained by companies of differing sizes and operational goals according to business-to-business (B2B) industries (i.e. Small and Medium Business (SMB), enterprise businesses and media companies). The final calculation of data issues is almost same for three categories. The percentages are 38%, 29% and 41% for SMB, enterprise and media companies respectively. The results indicated that the causes of

dirty data is always same. It is clear that the three categories which contain highest percentage of dirty data are:
a) Missing values
b) Invalid values
c) Duplicated data

In this research, the main objective is to overcome issues of incomplete data, due to missing data is produced by data sets basically missing values. These type of data considered concealed when the amount of values identified in a set, but the values themselves are unidentified, and it is also known to be condensed when there are values in a set that are predicted. The following research questions were addressed to be more exact:
a) How to train model to predict if the value is missing ?
b) How to repair the dirty data ?
c) What is the best Machine Learning Algorithm for building the model ?

The rest of paper is organized as follows: Section 2 presents the comparison between existing function in Python and developed function (AutoCDD). Section 3 demonstrates and evaluated performance of Auto-CDD system to make sure the prediction value's accuracy is precise. Then, Section 3 explains in details of developed System Design clearly. Lastly, Section 5 concludes the paper and discusses future prospects.

2. Comparison

As stated earlier, to develop the script of cleaning data Python Language a comparison is shown in Table 1 between existing functions in Python library and Auto-CDD. In the table, the column "Function" contains the task title of the method presented in "Call function example" column. Next, column "Description" contains the definition of the function written in python's Pandas official website. Finally, Pros and cons are written to understand the good and bad side of available functions.

Table 1. Comparison of Methods used for Cleaning Missing Data

Function	Call Function example	Description	Pros	Cons
Deleting Rows	data.dropna(inplace = True) [19]	"Return object with labels on given axis omitted where alternately any or all of the data are missing"	Complete removal of data with missing values results in robust and highly accurate model Deleting a particular row or a column with no specific information is better since it does not have a high weightage	Loss of information and data Works poorly if the percentage of missing values is high (say 30%), compared to the whole dataset
Replace With Mean /Median /Mode	data['age'].replace (np.NaN, data['age'].mean()) [20]	"Replace values given in 'to_replace' with 'value'"	This is a better approach when the data size is small It can prevent data loss which results in removal of the rows and columns	Imputing the approximations add variance and bias Works poorly compared to other multiple-imputations methods
Assigns a Distinct Category	data['age'].fillna('U') [21]	"Fill NA/NaN values using the specified method"	Fewer possibilities with one extra category, resulting in low variance after one hot encoding — since it is categorical Negates the loss of data by adding a unique category	Adds less variance Adds another feature to the model while encoding, which may result in poor performance
Predicts missing value	autocdd(data)	Predicts by selecting other features of missing attributes.	Assigning missing values data other than deleting the row/column is more effective for better performance It can help to predict numerical and non-numerical/categorical values. (Classification used for categorical prediction and Regression used for numerical prediction). It's not guessing the missing values, its rather predicting value using other variables.	As prediction depends on other values, unstable outcome may arise if most of the other values are incomplete.

3. System Design

The central goal of this study is to build a system for deriving a quality data set by detecting, analyzing, identifying and predicting the missing values. This task can be implemented using different Machine learning paradigm [4]. This system will able to perform independently without the help of any pre-developed software. As the system is developed using python Language. The system life cycle is divided into two stages, i.e. training/testing and prediction. Details of the phased are described in details in this section.

3.1. Training Phase

The first stage is Training Phase, as shown in Figure 1, the selected classification or regression machine learning model is trained using selected data sets. Initially, data is retrieved from .csv file and detect the column need to be cleaned. Next step is Feature Selection step, to obtain the important features to train with. After selecting the important features in this training phase, a machine learning model will be produced and will be saved. Finally, an evaluation is held to make sure the stored model produces accurate results.

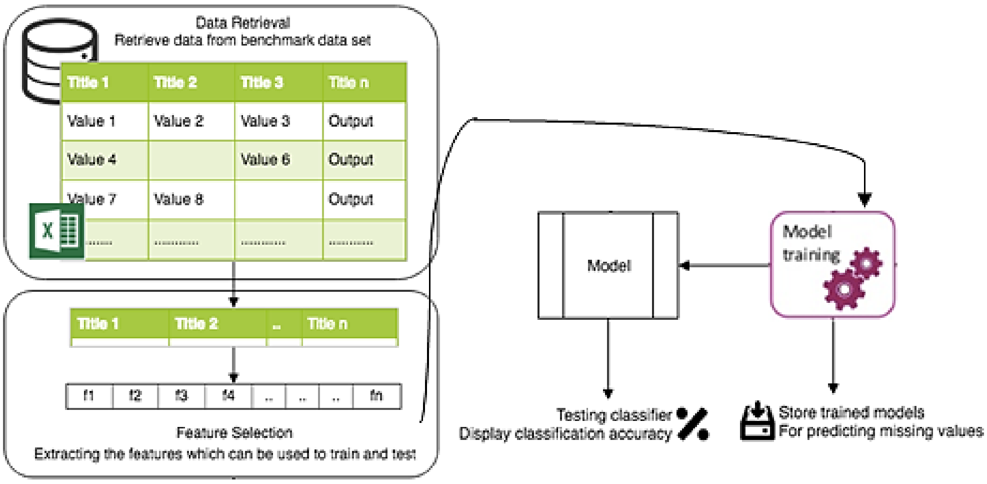


Figure 1. Training phase

3.1.1. Retrieving Data

The cleaning process is mostly processed on the stored dataset; since the system will be responsible for cleaning dirty data (such as missing data) it is important to retrieve data to process. As mentioned earlier, to develop the system python is used, hence 'PANDAS' was imported which is the best tool for data munging. It is a library of high-level data structuring dataset and manipulating tools, which helps to make analyzing data faster and easier. The dataset retrieved data from is stored in comma separated values (.csv) file. For the task reported in this paper, three sets of data selected which have missing values, as it will help to validate the system will work for cleaning data. The data set is selected according to the requirements of the system input. In the developed system three datasets are used. Details of data sets used are presented in Table 2.

Table 2. Data Sets used for Evaluating Developed

#	Data Repository	Data set	Features Characteristics	Number of Attributes
Data set 1	(UCI) [22]	Diabetics	Mixed	55
Data set 2	(UCI) [23]	Student Performance	Mixed	33

3.1.2. Feature Selection Based on Random Forest

In this stage Random Forest feature selection method is used. The steps of Random Forest algorithm includes:

Step 1: Extract feature sets from dataset including personalized and non-personalized features.

Step 2: Take M subset samples at random, without replacement from original feature sets.
Step 3: Build decision tree for each subset samples and calculate Gini index of all features.
Step 4: Rank Gini index in a descending order.
Step 5: Set the thresholds value, and then features with high contribution are selected as the representative features.

The columns selected to train the Machine Learning model by feature importance, the values are plotted in a Cluster Bar chart, as shown in Figures 2 and 3.

Data set 1 (student performance)

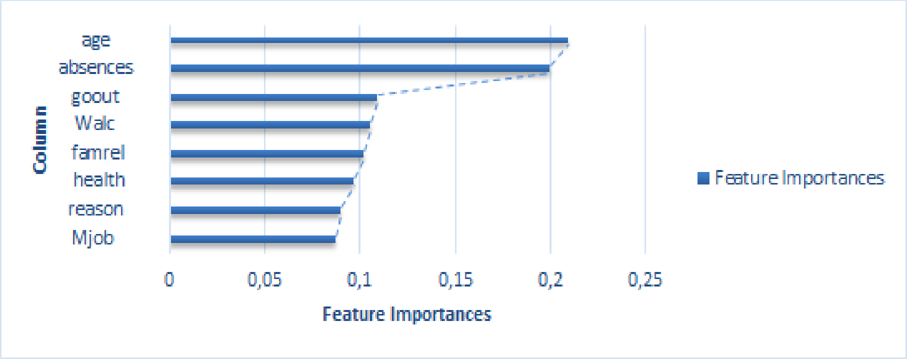


Figure 2. Feature importance (student performance)

Data set 2 (Diabetics Data)

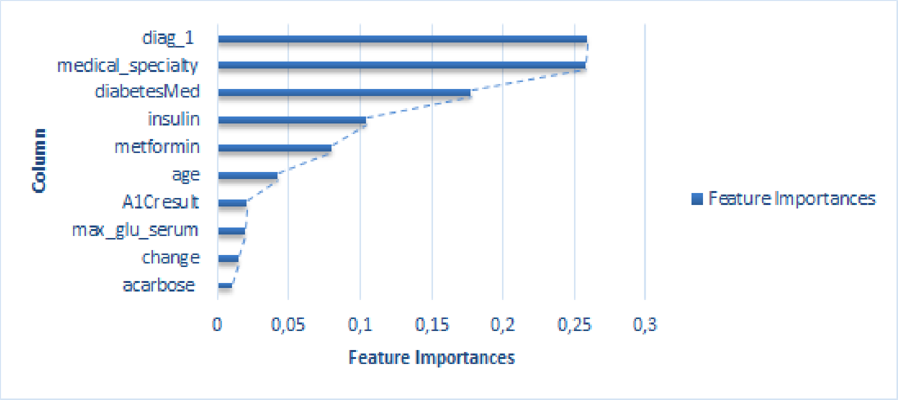


Figure 3. Feature importance (diabetics)

3.1.3. Training a Classifier Model

A set of features for each missing value's attributes are retrieved and then the old model is retrained to get better accuracy for predicting anomalies of data using the trained Machine Learning model. For training the model three common Machine Learning techniques are used, they are Random Forest, Linear SVM, and Linear Regression.

a. Random forest model

According to the system's requirement a supervised learning algorithm can be selected, where Random forest Algorithm is shown to provide a prediction with contains more than one Decision trees, and these trees are independent with each other [24]. It was implemented in different areas and proved to give great prediction accuracy, such as Network Fault Prediction [25]. Suppose there are T classes of samples in set C, then its Gini index is defined in (1):

$$\mathrm{gini}(\mathrm{T}) = \sum_{\mathrm{i}=1}^{\mathrm{nc}} \mathrm{p_i}(1 - \mathrm{p_i}) \quad (1)$$

where nc is the number of classes in set T (the target variable) and pi refers ratio of this class i. If considering dataset C splatted into two class, T1 and T2 with amount of data N1 and N2 respectively, then the Gini index for T is defined in (2).

$$\mathrm{Gini_{split}}(\mathrm{T}) = \frac{\mathrm{N_1}}{\mathrm{N}}\mathrm{Gini}(\mathrm{T_1}) + \frac{\mathrm{N_2}}{\mathrm{N}}\mathrm{Gini}(\mathrm{T_2}) \quad (2)$$

b. Support vector machine (SVM) model

Another supervised learning algorithm is selected, which is known to be strong algorithm used for classification and regression used in different domain, such as Healthcare [26], intrusion detection system [27], lymphoblast classification [28] and driving simulators [29]. It also helps to detect outliers using a built-in function. Implementation of Linear SVM, 'LinearSVC' option was used for able to perform multi-class classification. The (3) used for predicting new input in SVM by means of the dot product of input (x) with every support vector (x_i):

$$\mathrm{f(x)} = B_o + \mathrm{sum}(a_i * (x, x_i)) \quad (3)$$

where x is new input, and B_o and a_i value of each input is obtained from training data through the SVM algorithm. Whereas in Linear SVM the dot product is known as the kernel, the value defines comparison or a gap measure between new data and the support vectors. It can be re-written in form of (4)

$$\mathrm{K}(x, x_i) = \mathrm{sum}(x * x_i) \quad (4)$$

c. Logistic regression

One of the most common ML algorithm is Logistic Regression (LR). LR is not a regression algorithm it is one of the probabilistic classification model. Where, the ML classification techniques works as a learning method, which contains an instance mapped with one of the many labels available. Then machine learns and trains itself from the different patterns of data in such a way that it is able to represent correctly with the mapped original dimension and suggest the label/output without involving a human expert. The sigmoid function graph is plotted using (5):

$$S(x) = \frac{1}{1 + e^{-x}} \quad (5)$$

it makes sure that the produced outcome is always in between 0–1, as the denominator is greater than numerator by 1, as shown in (6).

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}. \quad (6)$$

3.2. Prediction Phase

The prediction phase shown in Figure 4, can be integrated into any pre-processing system, which detects and identifies missing value. Our system first retrieves data contains the missing value. Afterward, our system extracts feature, then predict the missing data by using the stored trained Machine Learning Model and provide predicted missing value. Finally, replace the NAN values with predicted values.

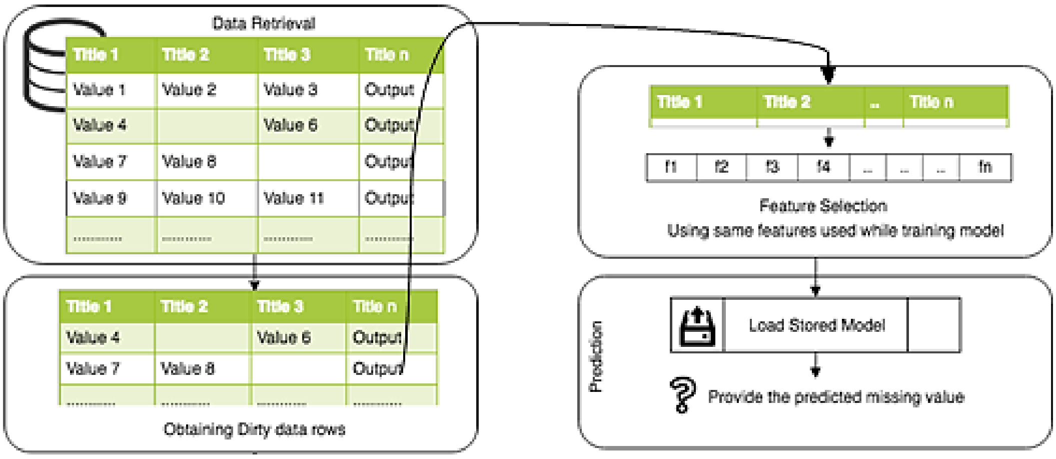


Figure 4. Prediction phase

4. Performance Evaluation

The importance of the performance evaluation is to investigate that how accurate and effective is the developed system, which is able to detect missing values, based on several metrics. Different type of data may give unlike level of prediction accuracy in a classification model. So different models are used and passed selected features from three data sets. Then cross-validation is implemented for further proof of the effectiveness of developed classifiers. More specifically, a selected dataset is divided into test and training sets (Diabetics Dataset obtained from ‘uci').

4.1. Classification Accuracy

The method used for evaluation is by retrieving TP (True Positive), TN (True Negative), FP (False Negative) and FN (False Negative) values. Where, TP is total amount of predicted correct/true value as expected; TN as total amount of predicted correct/true value as not expected; FP is total amount of predicted incorrect/false value as expected; FN as total amount of predicted incorrect/false value as not expected. Finally, accuracy is calculated by using following in (7).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad (7)$$

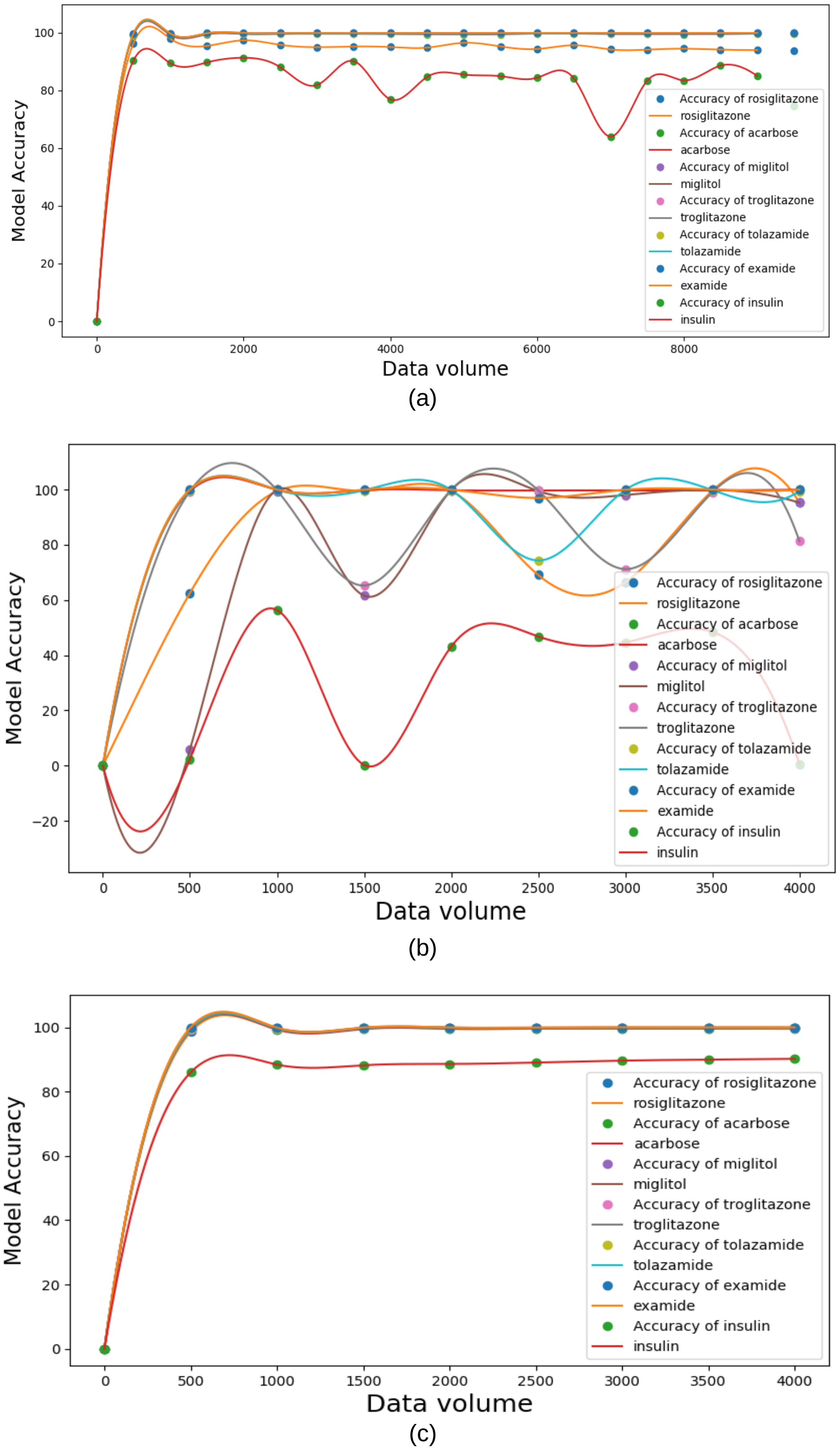
This accuracy of Machine learning Models depends on the data set selected to train. As different type of data sets will predict differently and different Learning models are used to get the best model according to the data set. Data sets were selected and the predicted outcome accuracies on different machine learning where presented in Figures 5-6 in form of graphs. This accuracy is the percentage of predicted missing values for each attribute, for example, in graph predicting values in ‘rosiglitazone’ column obtained from a CSV file. Three well-kwon supervised learning algorithms are used as mentioned earlier and in evaluation process from the three trained model, Random Forest Algorithm and Logistic Regression gave stable accuracy output throughout inputting data. Whereas, LinearSVM shows unstable and comparatively lower accuracy than other selected algorithm.

Case 1: Cleaning Dataset1-Diabetics Data:

Trained Random Forest Algorithm gave more than 90% accuracy, as shown in Figure 5 (a). Trained LinearSVM model shows to be an unstable model with lower accuracy of predicting missing values as shown in Figure 5 (b) and Logistic Regression trained algorithm proved to be more than 85% accuracy as shown in Figure 5 (c).

Case 2: Cleaning Data set 2 (Student Performance Data set):

Cleaning this data set, Logistic Regression performs in accuracy of greater than 90% as shown in Figure 6 (c) and Random Forest Algorithm is a close competitor in terms of accuracy 90% as shown in Figure 6 (a). Whereas. Linear Support Vector Machine again gives the bad performance of around 80% accuracy as shown in Figure 6 (b).

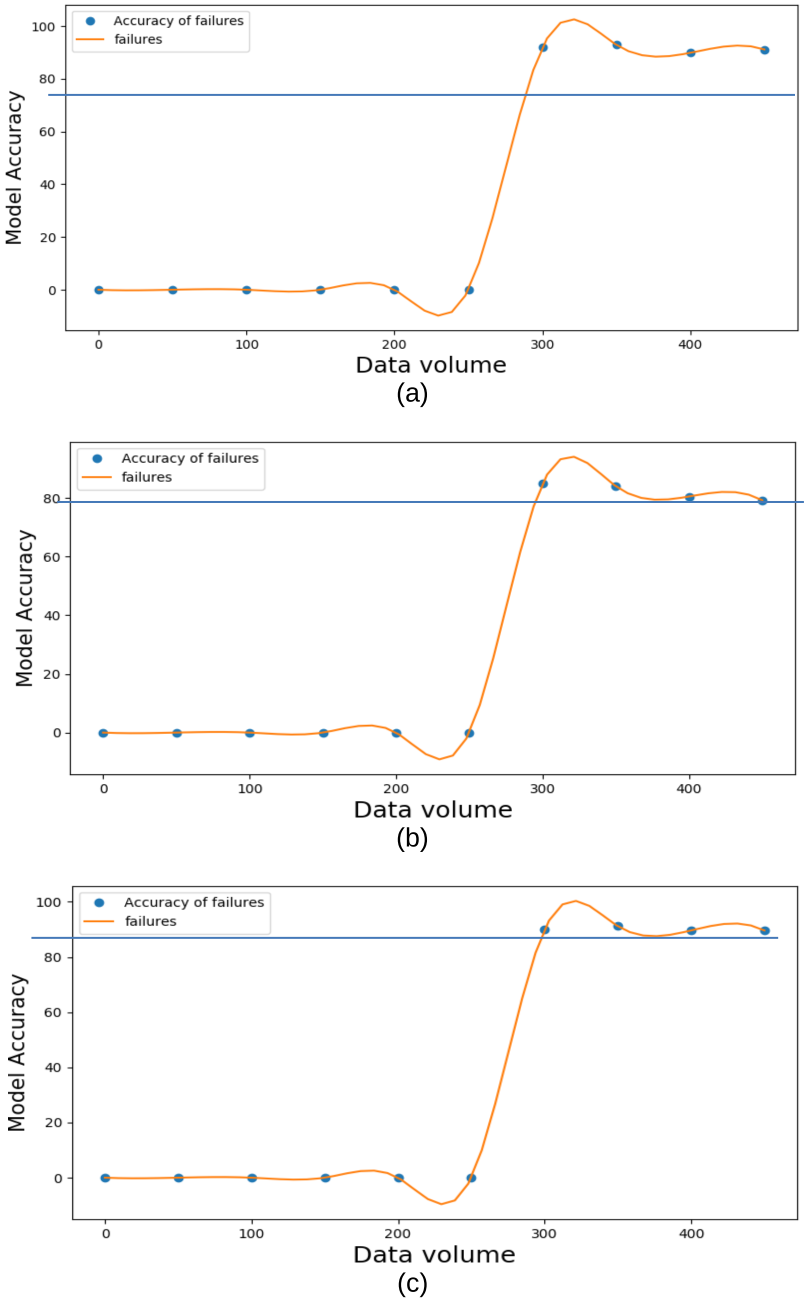


(a)

(b)

(c)

Figure 5. The accuracy obtained for Dataset 1 (a) accuracy percentage vs data volume for trained random forest (b) accuracy percentage vs data volume for trained linear svm (c) accuracy percentage vs data volume for trained logistic regression



(a)

(b)

(c)

Figure 6. The accuracy of prediction for dataset 2 (student performance) (a) accuracy percentage vs data volume for trained random forest (b) accuracy percentage vs data volume for trained linear svm (c) accuracy percentage vs data volume for trained logistic regression

For cleaning purpose and predicting missing data for each attribute, it’s proved that a trained Random Forest Model and Logistic Regression Model acts a better predictive model. Whereas, a trained LinearSVM shows to be unreliable for this type of prediction cause as it gives lower and unstable accuracy throughout training model by inputting new data into the model. This accuracy is further verified by using cross-validation technique.

4.2. Cross-Validation

Cross-validation technique is important to implement to confirm and examine the trained model can be reliable without issues (such as overfitting). Here, the data set is divided into k parts as shown in Figure 7 (where, k=5). This type of validation is known as k-fold cross-validation used to validate and determine the trained classifiers.

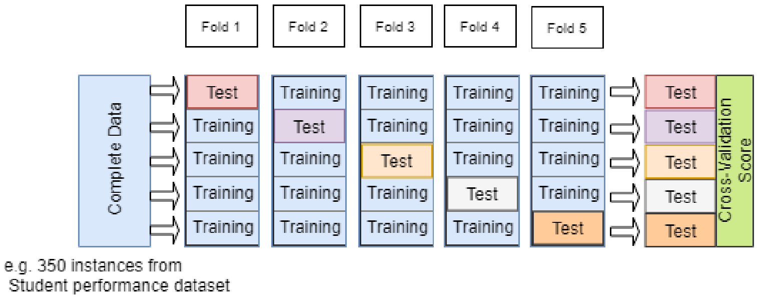


Figure 7. Data splitting in 5-fold cross validation

As the data set is divided into 5-folds, total of 1/5 of complete data used for testing and test data used for training. This training and testing are repeated 5 times, and total of each test accuracy is calculated to get Cross-validation score. The retrieved outcomes are entered into a table (presented in Table 3) with the classification accuracy obtained in previous stage for one column containing missing value(s). The outcomes proved that the model accuracy and cross-validation accuracy is almost close to each other. The trained model is not over-fitted and can be reliable.

Table 3. Cross-Validation Outcomes for Data Set 2 (Student Performance) for Failure

# of Instance	Model Accuracy	Cross-Validation Score
275	88.00%	86.182%
300	92.66%	88.333%
325	90.15	87.385%
350	90.85	87.143%

5. Conclusion

Almost all dataset available in repositories may contain attributes with missing data and it is very important to handle these type of data to overcome any performance issues. As different data set have different formats of data it is quite challenging task to deal with, and it is important to deal intelligently by using robust models. In this paper, a comparison is stated with pros and cons to will help the developer while selecting the best method for cleaning missing values. However, it's not essential to use one method for repairing data. Next, a system is designed and presented by using well-known Machine Learning algorithms for predicting missing data automatically. Three classification algorithms (i.e. SVM, Random Forest, and Logistic Regression) are used to test the process. The evaluation methods proved that two trained models are reliable on the data set selected. The k-fold cross-validation method confirms that the trained model is not over-fitted and can perform well with new dataset. For future work, combination of more than one method needs to be implemented with additional rules for data repair. It is also important to indicate and repair inappropriate or wrong data. Integrity constraints (such as Functional dependencies) can combine with Machine Learning Algorithms to classify the type of error to capture.

References

[1] F Sidi *et al.*, Data Quality : A Survey of Data Quality Dimensions, in 2012 International Conference on Information Retrieval & Knowledge Management (CAMP), 2012; 300–304.

[2] S Juddoo. *Overview of data quality challenges in the context of Big Data*, in 2015 International Conference on Computing, Communication and Security (ICCCS), 2015;

[3] I Taleb, HT El Kassabi, MA Serhani, R Dssouli, C Bouhaddioui. *Big Data Quality : A Quality Dimensions Evaluation*. in 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, 2016; 759–765.

[4] MZH Jesmeen, J Hossen, S Sayeed, CK Ho, K Tawsif, A Rahman. A Survey on Cleaning Dirty Data Using Machine Learning Paradigm for Big Data Analytics. *Indonesian Journal of Electrical Engineering and Computer Science*. 2018; 10(3): 1234–1243.

[5] J Hossen, MZH Jesmeen, S Sayeed. *Modifying Cleaning Method in Big Data Analytics Process using Random Forest Classifier*. 2018 7th International Conference on Computer and Communication Engineering (ICCCE), 2018: 208–213.

[6] J Bernardino, N Laranjeiro, SN Soydemir, J Bernardino. *A Survey on Data Quality : Classifying Poor Data A Survey on Data Quality : Classifying Poor Data*. in The 21st IEEE Pacific Rim International Symposium on Dependable Computing (PRDC 2015), 2015.

[7] I O MEDICINE. To Err Is Human: Building A Safer Health System. Washington (DC): National Academies Press (US). 2000.

[8] K Kenda, D Mladenić. Autonomous Sensor Data Cleaning in Stream Mining Setting, *Business Systems Research*. 2018; 9(2): 69–79.

[9] DC Corrales. How to Address the Data Quality Issues in Regression Models : A Guided Process for Data Cleaning. *SS symmetry*, 2018; 10(4): 1–20.

[10] M Gupta, S Sebastian. Framework to Analyze Customer's Feedback in Smartphone Industry Using Opinion Mining. *International Journal of Electrical and Computer Engineering*. 2018; 8(5): 3317–3324.

[11] Chanintorn. Granularity analysis of classification and estimation for complex datasets with MOA. *International Journal of Electrical and Computer Engineering*. 2019; 9(1): 409–416.

[12] H Liu, AK Tk, JP Thomas. *Cleaning Framework for Big Data-Object Identification and Linkage*. 2015 IEEE International Congress on Big Data. 2015;

[13] Y Chen, WHe, Y Hua, W Wang. CompoundEyes : Near-duplicate Detection in Large Scale Online Video Systems in the Cloud. in *I*EEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications CompoundEyes. 2016; 1–9.

[14] JM Dupare, NU Sambhe, A Novel Data Cleaning Algorithm Using RFID and WSN Integration 1, 2015;

[15] W Ku, S Member, H Chen. A Bayesian Inference-Based Framework for RFID Data Cleansing. *IEEE Transactions on Knowledge and Data Engineering*. 2013; 25(10): 2177–2191.

[16] BA Høverstad, A Tidemann, H Langseth. Effects of Data Cleansing on Load Prediction Algorithms, in *2013 IEEE Computational Intelligence Applications in Smart Grid (CIASG)*. 2013: 93–100.

[17] L Zhang, Y Zhao, S Member, Z Zhu. Multi-View Missing Data Completion, *IEEE Transactions on Knowledge And Data Engineering*. 2018; 30(7): 1–14.

[18] L. Barber, Data Decay & B2B Database Marketing [Infographic], *zoominfo*. [Online]. Available: http://blog.zoominfo.com/b2b-database-infographic/. [Accessed: 03-Nov-2018].

[19] Pandas, pandas.DataFrame.dropna. [Online]. Available: https://pandas.pydata.org/pandas-docs/version/0.21/generated/pandas.DataFrame.dropna.html. [Accessed: 08-Jun-2018].

[20] Pandas, pandas.DataFrame.replace. [Online]. Available: https://pandas.pydata.org/pandas-docs/version/0.22.0/generated/pandas.DataFrame.replace.html.

[21] Pandas, pandas.DataFrame.fillna. [Online]. Available: http://pandas.pydata.org/pandas-docs/version/0.22/generated/pandas.DataFrame.fillna.html. [Accessed: 03-Jun-2018].

[22] UCI, Diabetes 130-US hospitals for years 1999-2008 Data Set, *Clinical and Translational Research, Virginia Commonwealth University*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008. [Accessed: 03-Jun-2018].

[23] UCI. Student Performance Data Set, *Paulo Cortez, University of Minho, GuimarÃ£es, Portugal*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/student+performance. [Accessed: 25-May-2018].

[24] B Leo, Breiman Adele, Cutler Andy, Liaw Matthew. RandomForest. 2018.

[25] K Tawsif, J Hossen, JE Raja, A Rahman, MZH Jesmeen. *Network Fault Prediction using mRMR and Random Forest Classifier*. in Proceedings of Symposium on Electrical, Mechatronics and Applied Science 2018 (SEMA'18). 2018: 1–2.

[26] CS Sindhu, NP Hegde. A Novel Integrated Framework to Ensure Better Data Quality in Big Data Analytics over Cloud Environment. *International Journal of Electrical and Computer Engineering*. 2017; 7(5): 2798–2805.

[27] A Pradesh, A Info. A novel approach for selective feature mechanism for two-phase intrusion detection system. *Indonesian Journal of Electrical Engineering and Computer Science*. 2019; 14(1): 105–116.

[28] S Nabilah, M Safuan, M R Tomari, W Nurshazwani, W Zakaria. Computer aided system for lymphoblast classification to detectacute lymphoblastic leukemia. *Indonesian Journal of Electrical Engineering and Computer Science*. 2019; 14(2): 597–607.

[29] AM Rumagit, IA Akbar, M Utsunomiya, T Morie. Gazing as actual parameter for drowsiness assessment in driving simulators. *Indonesian Journal of Electrical Engineering and Computer Science*. 2019; 13(1): 170–178.