International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

An User Friendly Interface for Data Preprocessing and Visualization using Machine Learning Models

Mr. S. Yoganand[1], Bharathi Kannan R[2], Daya Meenakshi B[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Agni College of Technology Chennai-130, Tamil Nadu, India.

[2,3]UG Student, Department of Computer Science and Engineering,Agni College of Technology Chennai-130, Tamil Nadu, India.

Abstract – *Machine learning is one of the most efficient techniques for prediction and classification related problems. In this modern era, most of the industries all over the world depend upon the machine learning models which lead into the data analytics century. There is no proper and efficient tool for handling the datasets which use machine learning models for data prediction and Visualization. So, in this paper a novel idea is proposed for making the user-friendly approach to handle the machine learning models for data prediction and visualisation. A tool is developed, such that it performs data cleaning which will be a prerequisite for data analysis and then provides a visible representation of the cleansed data. The developed tool will take the input as structured dataset that contains both textual and numerical data which are then processed using machine learning algorithms to obtain a pre-processed dataset. This process may undergo series of steps to produce visualized and predicted data as per the chosen effective algorithm to obtain efficient result.*

***Key Words*: Machine learning, visualization, pre-processing, Tool, user interface.**

1. INTRODUCTION

An organisation uses the dataset for predictive analysis and an important concern in these cases is data quality. Using noisy data can hamper with the correctness of analysis. The common errors are missing values, duplicates and other errors. These errors need to be corrected for reliable decisions and analytics. The users must know that the effects of using the noisy data before proceeding with the cleaning process. Noise removal will improve the model performance, due to the fact that noises may disturb the discovery of important information.

Machine learning is the appreciated application of Artificial Intelligence. It is used to learn automatically without any human assistance that provides huge dataset for analysing with a large number of data fields. With the data provided by the system after implementing the machine learning algorithms, organizations are able to work more effective and acquire profit over their competitors. The system that uses machine learning technique will be able to predict how the structure looks like and adjust the data according to their structure. The main challenges in machine learning model is to deal with large data sources for data cleaning process. Data cleaning process is carried by taking in huge datasets which are checked for the possible errors by using data pre-processing techniques. The other challenges include avoiding learning process from noisy data, avoiding building a prejudiced model, not giving reasons for compromising with the quality of the data. The best practices for data cleaning using machine learning techniques that are filling missing values, removing unnecessary rows, reducing the size of the data and implementing a good quality plan.

The success of machine learning applications depends on the amount of good quality data that is given to it. But this process of cleaning may not be considered as a main area in data pre-processing. The system that uses powerful algorithms to process the noisy data can yield bad results if irrelevant or wrong training set of data is given. In the proposed model ML algorithms to find out the different patterns in the data and group it by itself into clean and noisy data which will help in reducing execution time.

2. Related Work:

Data Pre-processing is used to convert the raw data into pre-processed data set. [1] In Machine Learning, the data pre-processing is used to transform or encode the data easily by their algorithm. It consists of interactive steps as follows. Data cleaning is used to detect and correct inaccurate records from a record or tables, and then replacing, modifying or deleting this noisy data. Data integration will combines the data residing indifferent sources that provides user with a unified view of these data [2]. The process of selecting suitable data for a research project will impact data integrity where Data transformation converts data from a source data format into resultant data [3].

The tools which are available to process the data in data processing and visualizing are Knime, Shogun, Oryx 2, Tensor flow, Weka, RapidMiner, Trifacta Wrangler, Python [12] [13]. In this paper, we will focus on removing the noisy data that identifies the numerical values, predicting and filling in missing values and detect outliers which hamper with data analysis [11]. We propose a system that simplifies the process for the user and allows for better processing. In summary, Machine learning for data cleaning might be the only way to provide complete and trustworthy data sets for

effective analytics, so we provide an user friendly interface for pre-processing and model analysis with visualization for the ease of user.

3. System Design:

The Data Pre-processing is done with three methods they are Data Cleaning, Data Transformation and Data Reduction. The data cleaning application is to process the raw dataset containing both textual and numerical data that convert it into a cleaned dataset which can be used for data analysis. Initially, users must upload the dataset in which they perform the analysis. They can choose the operations that they want to perform on their dataset from the modules provided. This application performs a series of operations which includes removing columns with less information or no information, removing unnecessary rows, identifying the numerical values, filling in the missing fields and identifying the outliers. Some columns may contain less information or no information that makes it hard to rely on such columns for analysis and so such columns can be removed and they don't cause significant damage to the data.

Some rows may contain empty fields which will again tamper with the proper pre-processing of the dataset. Hence such values are identified and removed. The dataset will contain categorical features ranging from numerical to non-numerical values. This application requires only numerical data which is used for analysis and prediction, such that the fields containing numeric values are identified. If you try to remove them, you might reduce the amount of data that is available. So, these fields need to be filled in appropriate values.

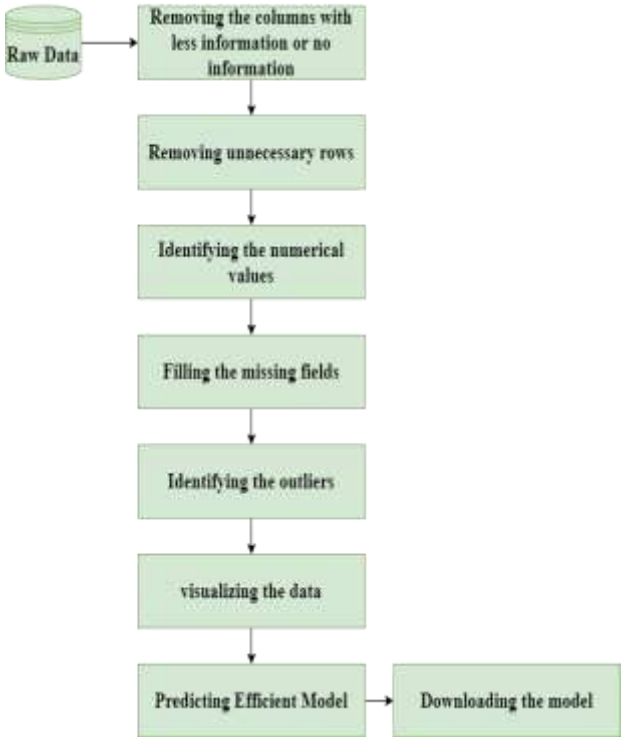
4. Implementation:

The outliers with data points are really far from the rest of your data points. Mathematically, an outlier is usually defined as an observation over three standard deviations from the mean. They can show up due to errors in data entry or measurement, or just because there's a variation in the population. Identifying and handling outliers is an important part of data cleaning.

In Data Analysis we are using the subsequent algorithms to analyse the cleansed data. Linear regression, SVM (Support Vector Machine), KNN (K-Nearest Neighbours), Logistic Regression, Decision Tree, K-Means, Random Forest, Naive Bayes, Dimensional Reduction Algorithms, Gradient Boosting Algorithms.

Linear Regression algorithm will use the info points to seek out the simplest fit line to model the info. A line can be represented by the equation, y = m*x + c where y is the dependent variable and x is the independent variable. Basic calculus theories are applied to seek out the values for m and c using the given data set. The SVM will separate the data points using a line. The KNN will predict unknown data point with its k nearest neighbours. The value of k is a critical factor regarding the accuracy of prediction. It determines the nearest distance using basic distance functions like Euclidean. This algorithm has to be a high computation power and that we have to normalize the information initially to bring every datum within the same range. The Decision Tree algorithm is used to solve classification problems. Some techniques are used to categorize the data they are Gini, Chi-square, entropy etc. K-Mean is an unsupervised algorithm that provides a solution for clustering problem. The algorithm will follow the procedure to form a cluster which contains homogeneous data.

Random forest is identified as a collection of decision trees. Every tree will try to estimate a classification and this is called as a vote. We consider each vote from every tree and chose the maximum voted classification. Naive Bayes can be applied only if the features are independent to each other. Gradient Boosting Algorithm uses multiple weak algorithms to form accurate algorithm. Instead of using the single estimator, will create a more stable and robust algorithm. Based on the data set the algorithm is predicted and provides an efficient result for data analysing process.

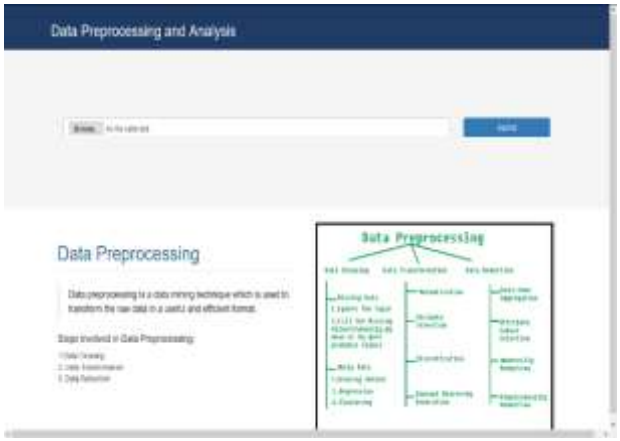


5. Results:

The user can click on the Submit button that is provided and then select the operations they wish to perform on their dataset from the list of operations provided. The user can then upload the dataset into the application by click on the Upload button to start the pre-processing. Initially the original dataset is displayed and then dataset after operation 1 will be displayed as cleansed dataset. The selected operations are performed with the

Cleansed dataset; finally the user will perform the data analysis with the required algorithm to obtain the result in visualization and it can be download by the user.

Upload Noisy Dataset:



Displaying the noisy Dataset:

Preprocessing the Data:

Applying Machine Learning Modal:

Output:

6. Conclusion:

Our developed system performs Data Cleaning, Data Transformation and Data Reduction in data pre-processing. Our system which takes the raw datasets into the application which are then pre-processed to clean up all the noisy data using pre-processing techniques and the cleansed data is visualized to the users after all the pre-processing is done. This system saves a lot of time since manual cleaning can be avoided. After cleansing the user can choose or select the machine learning model which will provide efficient results as plots. This serves as an effective purpose for the users who wants to clean huge datasets and visualizes the analysis of pre-processed data. In future the accuracy and comparison of the machine learning algorithms can be done within the friendly user interface.

REFERENCES

[1] Cristian Felix, Anshul Vikram Pandey, and Enrico Bertini, "TextTile: An Interactive Visualization Tool for Seamless Exploratory, Analysis of Structured Dataand Unstructured Text", IEEE-2018.

[2] Data,Huawen Liu, Xuelong Li, Jiuyong Li, andShichao Zhang, "Efficient Outlier Detection for High-Dimensional", IEEE-2019.

[3] M. Bostock, V. Ogievetsky, and J. Heer, "Datadriven documents," IEEE-2011.

[4] F. Beck, S. Koch, and D. Weiskopf, "Visual Analysis and Dissemination of Scientific Literature Collections with SurVis", IEEE-2016.

[5] Parke Godfrey, JarekGryz and PioterLasek,"Interactive visualisation of large datasets", IEEE-2016.

[6] Dileep kumarkoshley and Raju Hadler, "Data Cleaning: An Abstraction-based approach", IEEE-2015.

[7] Mehmet Adil Yalçın;NiklasElmqvist; Benjamin B. Bederson,"Keshif :Rapid and Expressive Tabular Data Exploration for Novices", IEEE-2018.

[8] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsosk, "LOCI: Fast outlier detection using the local correlation integral," IEEE 19th Int. Conf. Data Eng. (ICDE), Bengaluru, India, 2003, pp. 315–326.

[9] Y. Pang, J. Cao, and X. Li, "Learning sampling distributions for efficient object detection", IEEE Trans. Cybern., vol. 47, no. 1, pp. 117–129, Jan. 2017.

[10] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey", IEEE Trans. Knowl. Data Eng., vol. 26, no. 9, pp. 2250–2267, Sep. 2014.

[11] S. F. Roth and J. Mattis, "Automating the presentation of information," in Artificial Intelligence Applications, 1991. Pro-ceedings. , Seventh IEEE Conference on, vol. 1.IEEE, 1991, pp. 90–97.

[12] M. Bostock and J. Heer, "Protovis: A graphical toolkit for visualization," Visualization and Computer Graphics, IEEE Transactions on, vol. 15, no. 6, pp. 1121–1128, 2009.

[13] A. Dziedzic, J. Duggan, A. J. Elmore, V. Gadepally, and M. Stonebraker, "Bigdawg: a polystore for diverse interactive applications," in IEEE Viz Data Systems for Interactive Analysis, 2015.

[14] P. Bohannon, W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis "Conditional functional dependencies for data cleaning. In Data Engineering", IEEE 23rd International Conference on, pages 746–755. IEEE, 2007.