IJARIIT

INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1922)
Available online at: https://www.ijariit.com

An efficient data pre-processing model for machine learning

Piyush Lawatre
piyushlawatre@gmail.com
Anjuman College of Engineering and Technology, Nagpur, Maharashtra

Mohammed Muzzammil
muzzammilsilat56@gmail.com
Anjuman College of Engineering and Technology, Nagpur, Maharashtra

Rishabh Hingal
rishihingal@gmail.com
Anjuman College of Engineering and Technology, Nagpur, Maharashtra

Fatir Khan
khan.fatir99@gmail.com
Anjuman College of Engineering and Technology, Nagpur, Maharashtra

Rizwan Khan
rizwan4242khan@gmail.com
Anjuman College of Engineering and Technology, Nagpur, Maharashtra

ABSTRACT

Currently, data pre-processing is one in all the areas of nice interest as a result of it permits discover hidden andinfrequently attention-grabbing patterns in massive volumes of information. information scientists pay most of their time on information preparation tasks that has investigation regarding the info, loading information and cleanup information, in line with an exploration conducted by Anaconda. The real-world massive informationsets square measure obtained from several sources and contain data that tend to be incomplete, creaky and inconsistent thence required correct investigation. during this context, it's vital to arrange information to satisfy the necessities of information mining algorithms. this can be the role of information pre-processing stage, within which information cleanup, transformation and integration, or information spatiality reduction square measure performed. just about any sort of information analytics, information science or AI development needs some sort of information pre-processing to supply reliable, precise and strong results for enterprise applications..

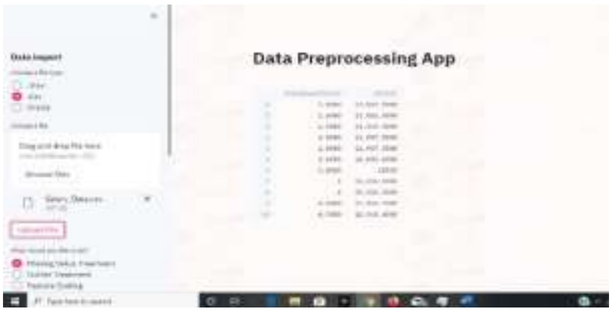
Keywords: *Mode, Median, Mean, Pre-processing, Outlier, Feature Scaling.*

1. INTRODUCTION

Data pre-processing describes any sort of process per- formed on information to organize it for an additional pro-cess procedure. it had been historically used as a prelim- inary step for an information mining method. additional recently, these techniques have evolved for coaching ma- chine learning and AI models and for running inferences against them. Also, these techniques are often employed incombination with a range of knowledge sources, togetherwith knowledge keep in files or databases, or being emitted by streaming knowledge systems.

Real-world knowledge is mussy and is usually created, processed and keep by a range of humans, business pro- cesses and applications. whereas it should be appropriate for the aim at hand, an information set could also be miss-ing individual fields, contain manual input errors, or have duplicate knowledge or totally different names to explain identical factor. In our project we have a tendency to pro-pose associate degree UI application that appearance for associate degree end-to-end resolution for knowledge pre- processing wherever we have a tendency to square mea- sure progressing to perform a number of the listed below following actions

- Missing Value Treatment
- Feature Scaling
- Outlier Treatment



2. LITERATURE REVIEW

Existing System

Many systems are available to process data. Mainly as Python scripts, Jupyter Notebooks, etc. These systems arequite useful and do solve their problems. These scripts canalso be modified as per the user's needs by adding customcode into those scripts. The problem with these scripts is that the user will need to use them from a computer whichhas python pre-installed as well as all the packages used in the script. These scripts are not user-friendly as they donot provide a user interface. Sometimes it can be difficult to execute these scripts due to some errors in the code, sothe user will have to either wait for the fixes or do it by themselves which will waste a significant amount of time.The existing system is

not much reliable as anyone cando changes in the code hence it increases the risk of com-promising sensitive data such as medical records, phone
numbers, etc.

Advantages Of Existing System

1. The existing system is highly efficient.
2. Easily customisable by the user
3. The existing system reduces errors.

Disadvantages Of Existing System

4. Large power consumption
5. Occupies large memory
6. The cost of installation is high
7. Wastage of memory.

3. SCOPE AND OBJECTIVE

Real world information area unit typically, Incomplete; lacking attribute values, lacking sure attributes of interest,or containing solely combination data; Noisy: containing errors or outliers; Inconsistent: containing discrepancies in codes or names. thus to beat these issues our main aim of this project is as given below:

- **Data Cleaning:** it's additionally referred to as scrub. This task involves filling of missing val- ues, smoothing or removing droning information and outliers beside resolution inconsistencies[6].
- **Data Integration:** This task involves integration in-formation from multiple sources like databases (re-lational and non-relational), data cubes, files, etc. the info sources will be consistent or heterogeneous.the info obtained from the sources will be structured, unstructured or semi-structured in format[4].
- **Data Transformation:** This involves social control and aggregation of knowledge consistent with the wants of the info set[5].
- **Data Reduction:** throughout this step information is reduced. the quantity of records or the quantity ofattributes or dimensions will be reduced. Reduction is performed by keeping in mind that reduced infor-mation ought to turn out a similar results as originalinformation.
- **Data Discretization:** it's thought-about as a vicinity of knowledge reduction. The numerical attributes area unit replaced with nominal ones[2].

4. MODELING AND ANALYSIS/ BENCHMARK AND DATASETS

- A dataset can be an excel (Xlsx) file, a Comma Sep-arated (CSV) file or a Database file.
- Firstly, a dataset is imported into the application andthe user is presented with tabular view of the data.
- From that view it makes it easier for the end user todetermine which actions will be useful for process-ing and cleaning data from various data anomalies such as outliers, missing values, etc.

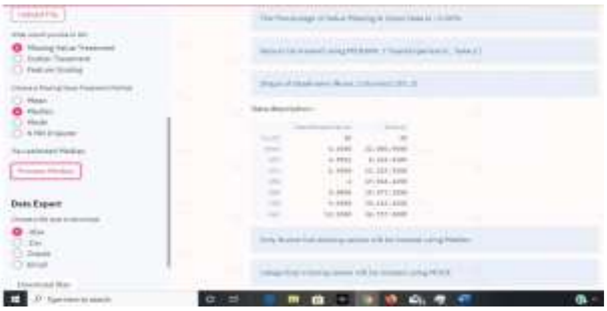
5. IMPLEMENTATION DETAILS

This web app is made using the python and it's libraries such as Pandas, Numpy , Seaborn. And the UI is designedwith the help of streamlit.

- Initially a data file is imported into the application and stored in temp memory to perform data clean- ing processes.
- User is presented with various processes to be per- formed on the dataset.
- **Missing price Treatment:** Missing values in knowledge may be a common development in uni- verse issues. Knowing a way to handle missing val-ues effectively may be a needed step to scale back bias and to provide powerful models. we tend to en-forced four treatment strategies that deals with miss-ing values in an exceedingly dataset. They are

1. **Mean:** Mean deviation from the mean tells North American nation however way, on av- erage, all values area unit from the center.
2. **Median:** Mean deviation from the mean tells North American nation however way, on aver-age, all values area unit from the center. The median formula is , wherever "n" is that the range of things within the set means that the
(n) range.
3. **Mode:** Mode price will typically be constant as mean and/or median, however not forever. The mode is incredibly helpful to search out out categorical knowledge.
4. **KNN Imputer:** knnImputation uses k-Nearest Neighbours approach to impute missing val- ues. What kNN imputation will in less com- plicated terms is as follows: for each obser-vation to be imputed, it identifies 'k' nighest observations supported the geometer distance and computes the weighted average (weightedsupported distance) of those 'k' obs.

- **Outlier Treatment:** An outlier is a data point that is distant from other similar points. They may be due to variability in the measurement or may indicate experimental errors[1]. If possible, outliers should be excluded from the data set. However, detecting that anomalous instances might be very difficult, and is not always possible. Our app deals with Out-liers in two ways :
-



1. Interquartile vary (IQR) : IQR is employed tolive variability by dividing a knowledge set into quartiles. the information is sorted in as- cending order and split into four equal com-ponents. Q1, Q2, Q3 referred to as 1st, sec- ond and third quartiles area unit the values that separate the four equal components. Q1 rep- resents the twenty fifth score of the informa- tion,Q2 represents the fiftieth score of the in-formation,Q3 represents the seventy fifth score of the information.
2. Z-score: Z-score is a very important idea in statistics. Z-score is additionally referred to asnormal score. This score helps to grasp if a knowledge price is larger or smaller than meanand the way remote it's from the mean. a lot of specifically, Z score tells what percentage normal deviations away a knowledge purposeis from the mean. Formula : (x -mean) / std. deviation

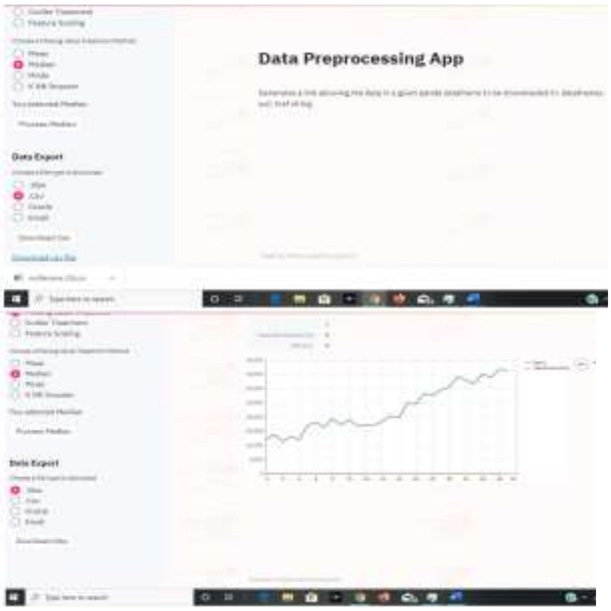
- **Feature Scaling :** Feature Scaling may be a technique to standardize the freelance options gift within the knowledge in an exceedingly fastened vary. it's performed throughout the information pre- processing to handle extremely varied magnitudes or values or units. If feature scaling isn't done, thena machine learning algorithmic rule tends to weigh larger values, higher and contemplate smaller values because the lower values, notwithstanding the unit of the values[3].

1. **Min-Max Normalization:** this system re-scales a feature or observation price with distribution price between zero and one.
2. **Standardization:** it's a really effective technique that re-

scales a feature price in order that it’s dis- tribution with zero norm and variance equals to one.

Data Export

After performing data pre-processing methods on thedataset it is time to export this processed data to the user. There are two ways either by downloading it directly intoyour local machine or by generating a link and emailingto either yourself or to another person. Dataset can bedownloaded in xlsx , csv, or in db file format.



6. CONCLUSION

An Efficient Data Pre-Processing Model for Machine Learning is a UI based system capable of filling the miss-ing values, smoothing or removing noisy data and outliers along with resolving inconsistencies. The System is ca- pable of processing industry standards data. We have development of our project to be highly scalable and ro- bust. All the module has been successfully implemented.

7. REFERENCES

[1] Irad Ben-Gal. Outlier detection. In *Data mining and knowledge discovery handbook*, pages 131–146. Springer, 2005.

[2] Ruoming Jin, Yuri Breitbart, and Chibuike Muoh. Data discretization unification. *Knowledge and Infor- mation Systems*, 19(1):1–29, 2009.

[3] Piotr Juszczak, D Tax, and Robert PW Duin. Feature scaling in support vector data description. In *Proc. asci*, pages 95–102. Citeseer, 2002.

[4] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246, 2002.

[5] S Manikandan. Data transformation. *Journal of Phar- macology and Pharmacotherapeutics*, 1(2):126, 2010.

[6] Erhard Rahm and Hong Hai Do. Data cleaning: Prob-lems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.