BMS INSTITUTE OF TECHNOLOGY & MANAGEMENT

YELAHANKA, BENGALURU - 560064

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

PROJECT BASED LEARNING

2022-23 Even Semester

“Cleaner for Non-Image Dataset”

Submitted By

Aravind Suresh

M S Kaushik

Manish A S

Sandeep Arockia Samraj X

Under the guidance of

Dr.Rajesh I S
Assistant Professor

2022-2023

INSTITUTE VISION

To emerge as one of the finest technical institutions of higher learning, to develop engineering professionals who are technically competent, ethical and environment friendly for the betterment of society.

INSTITUTE MISSION

Accomplish stimulating learning environment through high quality academic instruction, innovation and industry-institute interface.

DEPARTMENT VISION

To develop professionals equipped to build sustainable and intelligent solutions that effectively interact with the natural intelligence towards creating a digitally empowered environment for future generations, safeguarding social ethics.

DEPARTMENT MISSION

- To enable students with the spirit and power of interdisciplinary acumen by integrating a world of knowledge into a world of intelligent systems and subsystems.
- Boost academic outcome through place-based education and collaborations with established research labs and industries.
- Encourage entrepreneurship efforts among students and develop them into great leaders.

Abstract: The data cleaning is the process of identifying and removing the errors in the dataset. While collecting and combining data from various sources into a data warehouse, ensuring high data quality and consistency. Without clean and correct data the data cannot be used in algorithms.

The project is aimed to develop a website that can clean the dataset. It can be accessed by the programmers. We are using ReactJS for developing the website. We use python to develop the backend and to perform the data manipulation.

Introduction: Data collection has become important for large organizations, record keeping, data analysis tasks are critical to the organization's success. Data analysis typically drives decision-making processes. Although data collection and analysis is important, data quality is not good and might contain missing values. The presence of incorrect or inconsistent data can significantly impact the results of analyses.

Data cleaning is part of data preprocessing before data mining. Data cleaning is also called data cleansing and deals with detecting and removing errors and inconsistencies from data in order to improve quality of data. The main objective of data cleaning is to reduce the time and complexity and increase the quality of data.

In recent years, data science, artificial intelligence and machine learning domains are the new trend in IT industries. These domains are mainly dependent on datasets and these datasets must be cleaned in order to arrive at accurate results.

The project deals with cleaning of data by removing the unnecessary features in the dataset and filling the missing values. It also has the option to provide the metrics of the dataset for a particular machine learning algorithm model.

A website is created in which there is a place to upload the unclean dataset and a button to start the cleaning process. Towards the right side there is an option to select the algorithm and request the metrics acquired by the model built by using the cleaned dataset. Once the dataset is cleaned it is made downloadable to the users.

Motivation: As cleaning a dataset consumes a lot of time. We thought of providing Machine Learning programmers clean datasets.

Existing System:.Existing system takes the dataset and displays it for modifications.Some tools fills the missing values and provides the rows count.

Limitations of Existing Systems:

- It consumes more time.
- Doesn't remove the unnecessary features.
- Manually cleaning the values of datasets may lead to removal of necessary data.

Proposed System:

- A website that cleans the dataset by removing the unnecessary features and optionally provides the metrics if requested by the user.

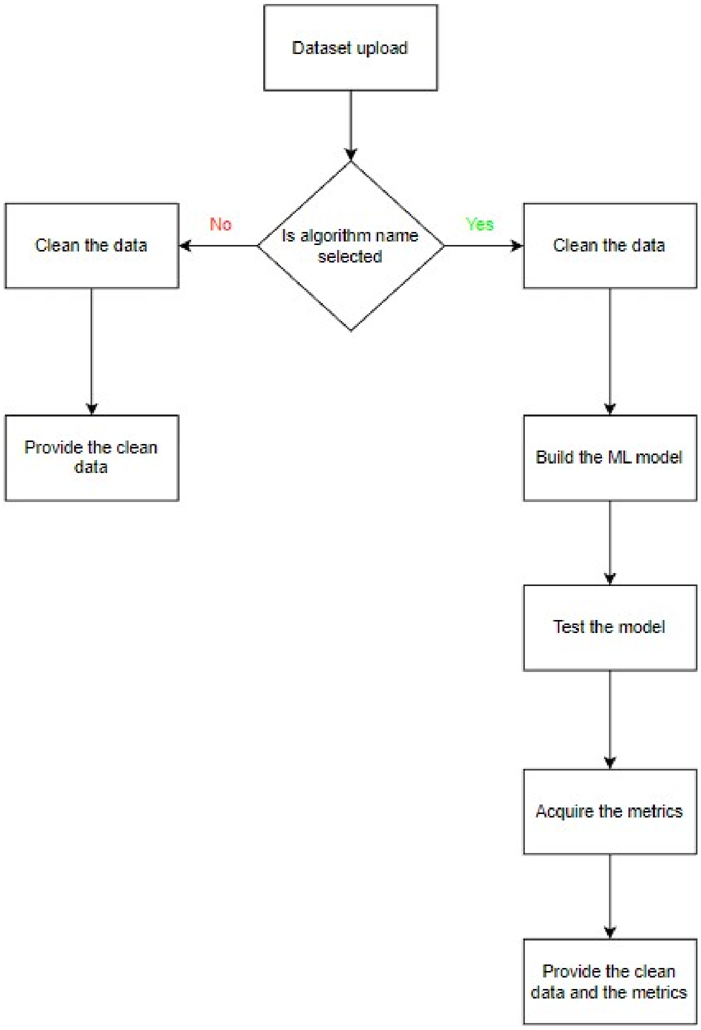
System Requirement Specifications (Functional & Non-Functional):

Functional Requirements: Upload unclean data, provide the algorithm for which the dataset will be used, download clean data, display the metrics.

Non-Functional Requirements: Usability, Performance

Proposed Methodology: The frontend of the website is created using ReactJS and for backend Python (Flask) is used. Then to fix the missing values, a decision will be made whether to remove the entire row of the missing field or to fill the values using mean, median or interpolation methods. Then to convert textual values to numerical we use the oneHotEncoding method. To select the required features we use Univariate Selection which is used to select those features that have the strongest relationship with the output variable.Finally the cleaned dataset will be available to download.

If the metrics are required by the user, then if the user must select the algorithm name then a model will be created and tested for its accuracy, precision, recall, etc and display the metrics along with the cleaned data.



Conclusion: Data cleaning is a very necessary part in any ML or DS projects. The data is manipulated so that it will be optimal to be used in a machine learning algorithm. This allows programmers to save time in cleaning the dataset. Thus it will be helpful to any project that uses datasets.The technique is suitable for one type of data cleaning and is not suitable for the other types.

References:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1198040/

https://www.researchgate.net/publication/278301609_Data_Cleaning_Current_Approaches_and_Issues

https://www.ijcsmc.com/docs/papers/March2015/V4I3201599a30.pdf

https://dsf.berkeley.edu/jmh/papers/cleaning-unece.pdf