BMS INSTITUTE OF TECHNOLOGY & MANAGEMENT

Yelahanka, Bengaluru – 560064

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

PROJECT BASED LEARNING (PBL)

2022-23 EVEN SEMESTER

DATASET CLEANER FOR NON-IMAGE DATASETS

Submitted by

Aravind Suresh

M S Kaushik

Manish A S

Sandeep Samraj X

DR. RAJESH IS

Assistant Professor, AI&ML

BMSIT&M, Yelahanka, Bengaluru – 560064

2022-23

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

CERTIFICATE

This is to certify that the project report entitled **DATASET CLEANER FOR NON-IMAGE DATASETS** is a Bonafide work carried out by **“Aravind Suresh ,**

M S Kaushik , Manish A S , Sandeep X , during the year 2022-2023. It is certified that all corrections / suggestions indicated have been incorporated in this report.

Signature of Guide **Signature of HOD**

BMS INSTITUTE OF TECHNOLOGY & MANAGEMENT

DECLARATION

We **Aravind Suresh, M S Kaushik, Manish A S and Sandeep Samraj X ,** pursuing BE degree from BMS Institute of Technology and Management, Yelahanka, Bengaluru, hereby declare that the report entitled **" Dataset Cleaner for Non-Image Datasets"** carried out is a Bonafide record of work done by us during the year 2022-23 and all the year 2022-23 and all the contents of the report and presented by us.

Place: Bengaluru Urban

Date: July 2023

ACKNOWLEDGEMENT

We are happy to present this project after completing it successfully. This project would not have been possible without the guidance, assistance and suggestions of many individuals. We would like to express our deep sense of gratitude and indebtedness to each and every one who has helped me to make this project a success.

We heartily thank our beloved **Principal, Dr. Mohan Babu G. N.** BMS Institute of Technology & Management for his constant encouragement and inspiration in taking up this project.

We heartily thank our beloved **Head, Dr. Anupama H S.** Department of Artificial Intelligence and Machine Learning, BMS Institute of Technology & Management for her constant encouragement and inspiration in taking up this project.

We gratefully thank the project guide, **Dr. Rajesh I S, Assistant Professor**, Department of Artificial Intelligence and Machine Learning, for their encouragement and advice throughout the course of the project work.

Special thanks to all the staff members of the Department of Artificial Intelligence and Machine Learning for their help and kind cooperation.

Lastly, we thank my parents and friends for their encouragement and support given to us in order to finish this precious work.

ABSTRACT

Detecting and repairing incorrect data is one of the most difficult challenges in data analytics, and failure to do so can result in inaccurate analytics and unreliable decisions. Over the past few years, there has been a surge of interest from both industry and academic domain on data cleaning problems including new abstractions, interfaces, approaches for scalability, and statistical techniques. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

CONTENTS

CHAPTER 1

INTRODUCTION

The availability of high-quality data is typically a prerequisite for developing high performing accurate Machine Learning (ML) applications. However, data is rarely clean in reality due to noisy inputs from manual data curation or inevitable flaws from automatic data collection or generation processes. For example, the inconsistency and incompleteness caused by broken sensors or human-error are common in real-world datasets, and can impact the machine learning systems built on top of them. To address this problem, we reviewed and propose several state-of-the-art data cleaning approaches based on literature review of publications in the data management systems field.

The availability of high-quality data is typically a prerequisite for developing high performing accurate Machine Learning (ML) applications. However, data is rarely clean in reality due to noisy inputs from manual data curation or inevitable flaws from automatic data collection or generation processes. For example, the inconsistency and incompleteness caused by broken sensors or human-error are common in real-world datasets, and can impact the machine learning systems built on top of them

Machine learning (ML) has gained widespread adoption in real-world problems that span business manufacturing , healthcare , agriculture , and more. ML relies on - and is “programmed" by - training data. Thus, the quality of the training data is a fundamental ingredient toward robust and accurate models, and ultimately toward useful and reliable ML-based applications. For this reason, data and ML engineers spend a tremendous amount of time—80% or more of a data scientist’s time on wrangling and cleaning the required datasets for their ML applications. Traditional data cleaning seeks to directly address data quality issues in a specific dataset. Given a structured dataset that potentially contains errors, it seeks to identify and/or repair those errors to derive a cleaned dataset that can be shared with the rest of the organization, or used by subsequent queries and applications without worry. Since cleaning occurs prior to, and often independent of the application, these techniques typically rely on error models to detecting duplicates or outliers, external constraint information (e.g., functional dependencies or integrity constraints), or human assessment and input (e.g., to recommend repairs or cleaning examples). The separation of data cleaning and the application is not optimal. For one, it is hard for users to define, or even assess, the correct integrity constraints for the application. It is also hard to anticipate the different ways that the cleaned dataset will later be used. Further, improving a dataset could in fact degrade the application . Thus, it is often unclear how a given cleaning intervention will affect the downstream application

In today's data-driven world, organizations rely heavily on accurate and reliable data for making informed decisions. However, data collected from various sources often suffer from issues like errors, inconsistencies, missing values, and outliers, which can significantly impact the quality and reliability of the insights derived from that data. To address these challenges, data cleaning has emerged as a critical step in the data preprocessing pipeline.

A data cleaner for non-image datasets is a specialized tool or software that is designed to identify and rectify errors, inconsistencies, and other quality issues present in non-image data. Unlike image data, non-image datasets comprise structured or unstructured data in various formats such as text, numerical values, categorical variables, time series, and more.

The primary objective of a data cleaner is to transform raw, messy data into a clean and usable format.

A dataset cleaner is a tool that helps you to remove unwanted or corrupted data from your dataset, such as duplicates, outliers, missing values, etc. This can improve the quality and accuracy of your data analysis or machine learning models.

There are different ways to clean a dataset, depending on the type and format of the data. For non-image datasets, such as tabular or text data, you can use various techniques such as:

- **Exploratory data analysis**: This involves using statistical methods and data visualization to understand the structure, distribution, and relationships of the data. You can use libraries such as pandas, numpy, matplotlib, seaborn, etc. in Python to perform this task.
- **Data preprocessing**: This involves transforming the data into a suitable format for analysis or modeling, such as encoding categorical variables, scaling numerical variables, handling missing values, etc. You can use libraries such as sklearn, scipy, etc. in Python to perform this task.
- **Data validation**: This involves checking the consistency and integrity of the data against some rules or criteria, such as data types, ranges, formats, etc. You can use libraries such as pandas-validator, cerberus, schema, etc. in Python to perform this task.

The project deals with cleaning of data by removing the unnecessary features in the dataset and filling the missing values.

CHAPTER 2

PROBLEM STATEMENT AND OBJECTIVE

2.1 Problem Statement

Cleaning a dataset is time consuming and only small datasets can be cleaned manually.

2.2 Objective

- This project aims to reduce the time to clean the dataset.
- Also to provide a general idea on how good the dataset is for a particular machine learning algorithm.
- To improve the quality of the dataset.

CHAPTER 3

LITERATURE SURVEY

SL No.	Paper	Work Carried out	Limitations
1.	Jesmeen, M.Z.H., Hossen, J "A survey on cleaning dirty data using machine learning paradigm for big data analytics." [1]	An overview is initiated to identify the potential of data cleaning in big data analytics in the process of gathering, arranging and processing information. A comparison of commercialized tools is presented by obtaining comments from different customers.	● Scalability ● User Engagement ● Manual
2.	Neutatz. F, Chen. B "From Cleaning before ML to Cleaning for ML" [2]	The data is prepared by canonicalizing the user ids, extracting product information from the browsed pages, and ensuring that the expected attributes appear in each data record. Separate data science teams develop two models: the first estimates the likelihood that a given user will leave the service (churn), and the second estimates a user's preference for different products (affinity).	● Unstructured data ● Population-level Errors ● Model entanglement

3.	Ga Young Lee, Lubna Alzamil “A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance”[3]	Survey paper presents several approaches to data cleaning developed by reputable researchers and developers in the database management systems field to solve the challenge with data cleaning. Various data cleaning techniques are analyzed and compared, a general trend observed from the literature review is a transition toward more scalable and efficient frameworks that aims to reduce the human overhead cost in the development of more accurate and representative data instances	● Lack of Optimizer ● Tradeoff Between Efficiency and Coverage ● Limited Generalizability
4.	Mohammad Mahdavi , Felix Neutatz “Towards Automated Data Cleaning Workflows” [4]	For unknown datasets, it is unrealistic to know the data quality problems upfront and to formulate all necessary quality constraints in one shot. Pragmatically, the user solves data quality problems by implementing an iterative cleaning process. This incremental approach poses the challenge of identifying the right sequence of cleaning routines and their configurations.	● Scalability ● Lack of Optimizer ● Model entanglement
5.	David Camilo Corrales, Juan Carlos Corrales “How to Address the Data Quality Issues in Regression Models: A Guided Process for Data Cleaning” [5]	DC-RM provides support to methodologies from data mining and machine learning. For instance, in Knowledge Discovery in Databases, DC-RM can support the Preprocessing and Data Cleaning, Data Reduction, and Projection phases. In Cross Industry Standard Process for Data Mining, DC-RM gives support to Verify Data Quality and Clean Data steps. and, in Data Science Process into the Clean Data phase	● Outliers Detection ● Density-Based Spatial Clustering

Fig3.1: Literature Survey

CHAPTER 4

PROPOSED METHODOLOGY

The frontend of the website is created using ReactJS and for backend Python (Flask) is used.

The Fig 4.1 shows the work flow of the dataset cleaning process.

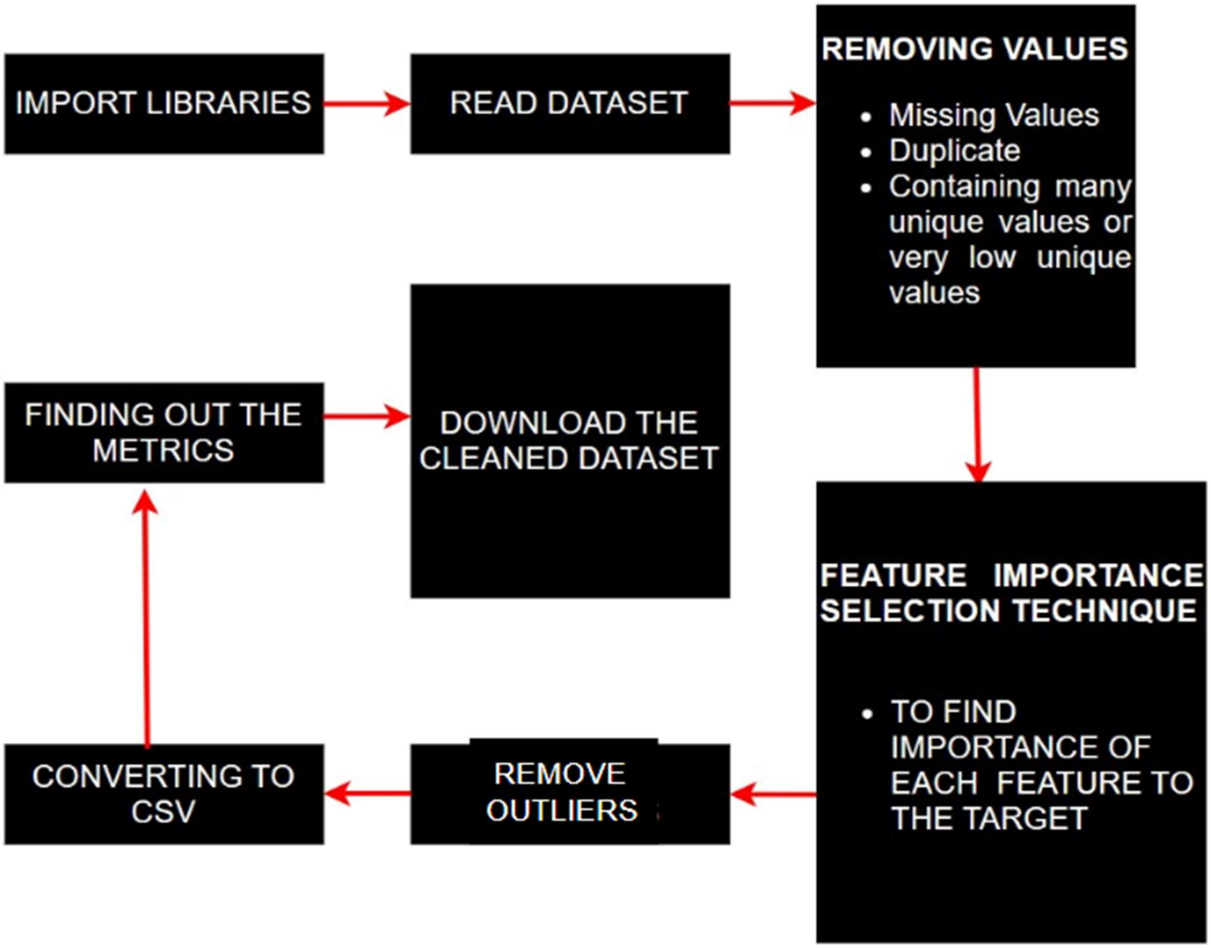

IMPORT LIBRARIES
READ DATASET
REMOVING VALUES
• Missing Values
• Duplicate
• Containing many unique values or very low unique values
FEATURE IMPORTANCE SELECTION TECHNIQUE
• TO FIND IMPORTANCE OF EACH FEATURE TO THE TARGET
REMOVE OUTLIERS
CONVERTING TO CSV
FINDING OUT THE METRICS
DOWNLOAD THE CLEANED DATASET

Fig 4.1: Flowchart for the project

The steps involved to clean the non - image dataset are:

- **Data diagnosing**: This involves identifying and locating the errors or anomalies in the data, such as duplicates, outliers, missing values, invalid values, etc. You can use libraries such as sklearn, scipy, etc..
- **De-duplication**: This involves removing or resolving the duplicate records in the data. You can use libraries such as pandas-dedupe, recordlinkage, fuzzywuzzy, etc.
- **Invalid data**: This involves removing or correcting the invalid values in the data that do not conform to the expected format or range. You can use libraries such as pandas, numpy, etc.
- **Missing data**: This involves handling the missing values in the data by deleting them, imputing them, or ignoring them. You can use libraries such as sklearn.impute, fancyimpute, missing values, etc.
- **Outliers**: This involves detecting and handling the extreme values in the data that deviate significantly from the rest of the data. You can use libraries such as scipy.stats, sklearn.covariance, pyod, etc.

CHAPTER 5

SYSTEM REQUIREMENTS

Hardware Requirements:

1. Windows 8 and above
2. RAM minimum of 2GB, 4GB is recommended.

Software Requirements:

1. Client: Operating System
2. Software: Spyder, VS code,
3. Language: Python, ReactJs

CHAPTER 6

IMPLEMENTATION

We are using Python (FLASK) AND React Js for implementing this. project.

Python:

Python is an interpreted, object-oriented, high-level programming language. Its high-level built in Flask is a web framework, it's a Python module that lets you develop web applications easily.

REACT JS:

ReactJS is a declarative, efficient, and flexible JavaScript library for building reusable UI components. It is an open-source, component-based front end library which is responsible only for the view layer of the application. It was initially developed and maintained by Facebook and later used in its products like WhatsApp & Instagram.

To implement the process of automated cleaning of dataset, the following steps are performed:

1.**Removing missing values**: First the columns with multiple null values are removed.Then the rows with null values are removed.

Removing missing values is an essential step in data preprocessing to ensure the accuracy and reliability of data analysis. Missing values, represented as NaN (Not a Number) or NULL, can occur due to various reasons such as data collection errors, incomplete records, or system issues. Handling missing values requires careful consideration, and there are several approaches available.

One common method is to remove the rows or columns containing missing values entirely. This approach is suitable when the missing values are limited and do not significantly affect the overall dataset. However, this method may lead to a loss of valuable information if the missing values are widespread.

Another approach is imputation, where missing values are replaced with estimated or predicted values. This can be done using statistical measures such as mean, median, or mode of the available data. Imputation helps retain the dataset's size and structure, but it introduces potential bias, especially if the missing values are not randomly distributed.

Advanced imputation techniques like regression imputation or machine learning algorithms can be employed when the dataset's complexity and dependencies are high

2.**Removing duplicates**: The rows which are repeated are deleted.
Removing duplicates from a dataset is an important step in data preprocessing to ensure data integrity and avoid bias in subsequent analyses. Duplicate values can arise from various sources, such as data entry errors, merging multiple datasets, or system glitches. Removing duplicates involves identifying and eliminating identical or highly similar records.

One common method for duplicate removal is to compare records across one or multiple columns and remove exact duplicates. This can be done by using functions or methods provided by programming languages or data analysis tools. By removing identical records, data integrity is preserved, and subsequent analyses are not skewed by duplicate entries.

In some cases, duplicates may not be exact but exhibit similarity across certain attributes.

3.**Uniques columns**: The columns with multiple unique values as well as the columns with very few unique values are removed.

Unique columns in a dataset refer to those columns that contain distinct or non-repetitive values. These columns play a crucial role in data analysis as they provide information that is different for each record, thus offering diverse perspectives and insights. Identifying unique columns is important for understanding the uniqueness and variability within the dataset.

To determine unique columns, one can iterate through each column and check for the absence of duplicate values. This can be accomplished using various programming languages or data analysis tools that provide functions or methods for identifying unique values. By doing so, one can isolate columns that contain a single value per record, indicating uniqueness.

Unique columns can be particularly informative in several scenarios. For categorical variables, unique columns may indicate distinct categories or labels that differentiate records. In such cases, these columns can be useful for grouping, filtering, or segmenting data based on unique attributes.

For numerical variables, unique columns can provide insights into individual data points that are non-repetitive or have unique characteristics. Analyzing these unique values can help identify outliers, anomalies, or specific patterns within the dataset.

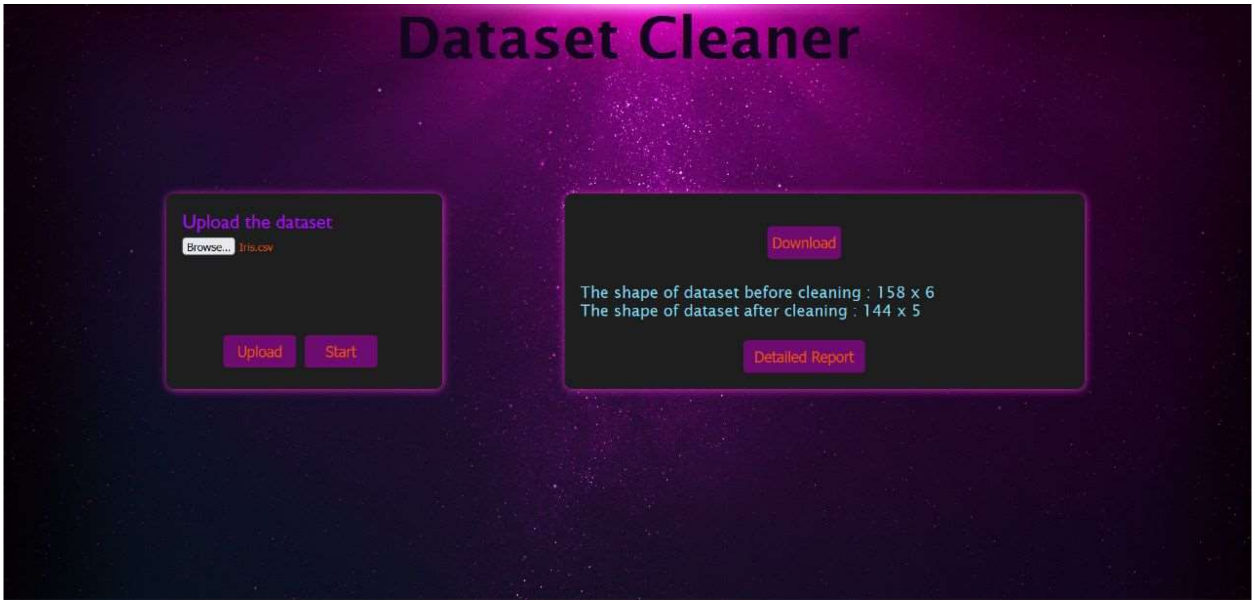
4.**Feature importance**: First the textual data is converted into numbers using Integer Encoding.

Feature importance of each feature is found by using ExtraTreesClassifier. Then the columns with lesser importance are removed. Feature importance refers to the process of determining the relevance or contribution of different features (variables) in a dataset towards predicting the target variable. It helps identify the most influential features that have a significant impact on the model's performance or the outcome of a problem. Feature importance can be estimated using various techniques such as statistical measures, feature selection algorithms, or machine learning models. Understanding feature importance allows researchers or data scientists to prioritize and focus on the most informative features, leading to improved model accuracy, interpretability, and better decision-making

5.**Remove Outliers**:Removing outliers is a crucial step in data preprocessing to improve the accuracy and reliability of data analysis. Outliers are data points that significantly deviate from the majority of the dataset. To remove outliers, various methods can be applied, such as using statistical measures like the z-score or the interquartile range (IQR). These methods identify data points that fall outside a defined threshold and can be subsequently removed or adjusted. Removing outliers helps to mitigate the impact of extreme values on statistical measures and ensures that subsequent analyses are not biased or skewed by the presence of outliers.

CHAPTER 7

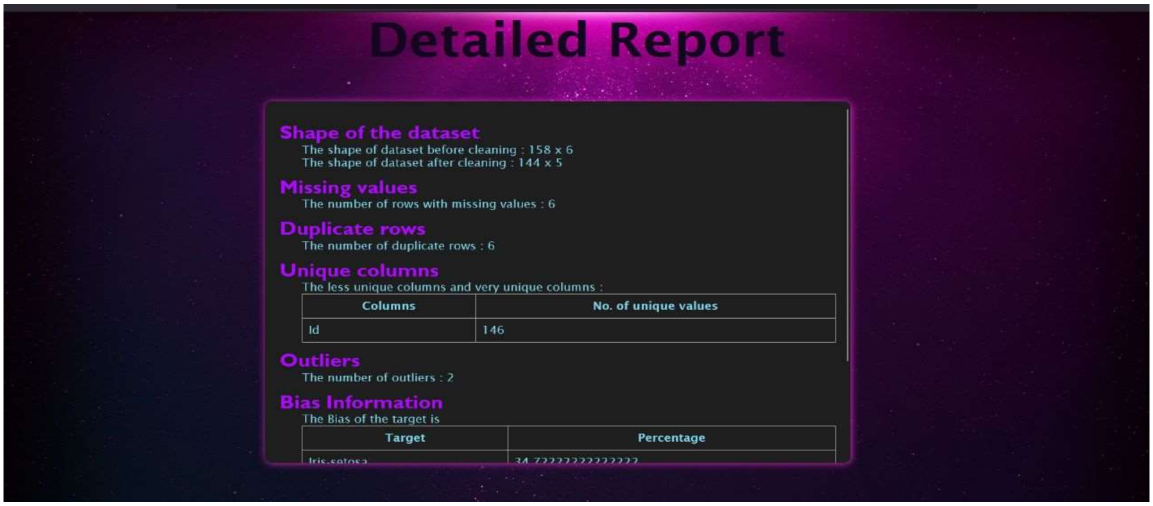
RESULT



The shape of dataset before cleaning : 158 x 6
The shape of dataset after cleaning : 144 x 5

Fig 7.1

Dataset before cleaning and after cleaning .The image show the no of rows and columns before cleaning and the no of rows and columns after cleaning



Shape of the dataset

The shape of dataset before cleaning : 158 x 6
The shape of dataset after cleaning : 144 x 5

Missing values

The number of rows with missing values : 6

Duplicate rows

The number of duplicate rows : 6

Unique columns

The less unique columns and very unique columns :

Columns	No. of unique values
Id	146

Outliers

The number of outliers : 2

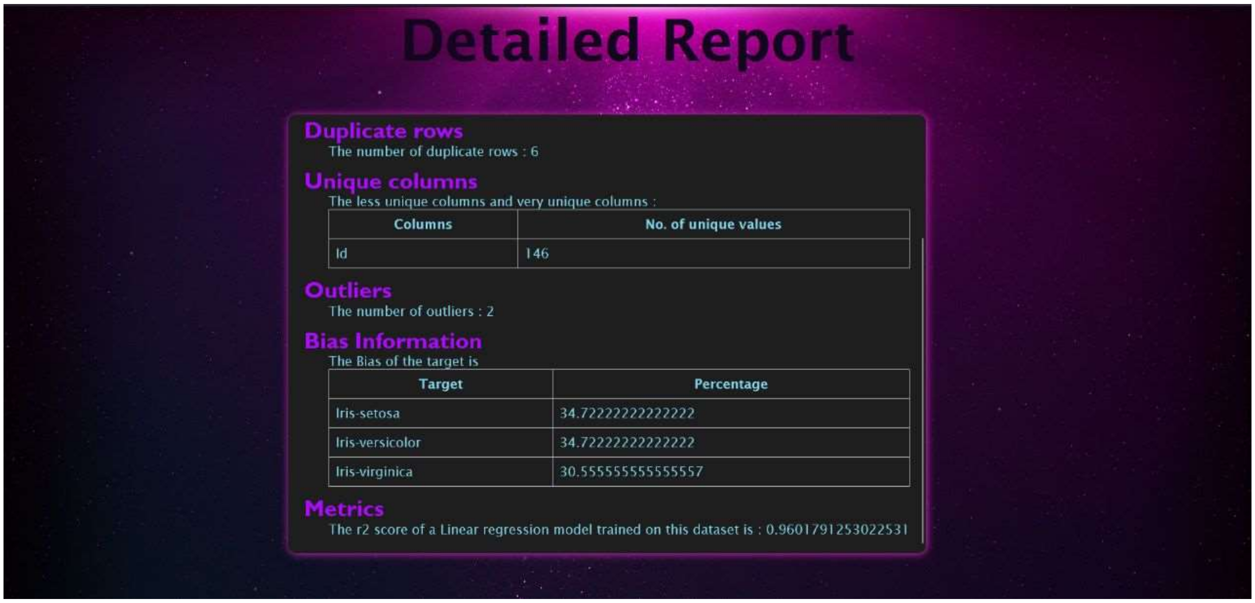
Bias Information

The Bias of the target is

Target	Percentage
Iris-setosa	34.72222222222222

Fig 7.2

Shows the detailed report of missing values , duplicated rows , unique columns and outliers and bias information



Detailed Report

Duplicate rows

The number of duplicate rows : 6

Unique columns

The less unique columns and very unique columns :

Columns	No. of unique values
Id	146

Outliers

The number of outliers : 2

Bias Information

The Bias of the target is

Target	Percentage
Iris-setosa	34.72222222222222
Iris-versicolor	34.72222222222222
Iris-virginica	30.555555555555557

Metrics

The r2 score of a Linear regression model trained on this dataset is : 0.9601791253022531

Fig 7.3

Shows the download report and bias information

CHAPTER 8

CONCLUSION AND FUTURE WORK

We conclude that we accomplished our objectives by reducing the time to clean the dataset and improving the quality of the dataset with max r2_score possible.

The limitations of our project will be our challenge as the future work. They are:

- Websites cannot handle heavy datasets.
- Low webpage security.
- Advancing the field of data cleaning for non-image datasets through improved techniques, automated approaches, and domain-specific considerations can further enhance the quality and usability of the data
- Enabling more reliable and meaningful data analysis and decision-making
- Additionally, developing automated approaches or algorithms that can identify and handle outliers in a more robust and adaptable manner would be beneficial
- Incorporating domain-specific knowledge and expertise into the data cleaning process can enhance the accuracy and relevance of the cleaning techniques

REFERENCES

1. Jesmeen, M.Z.H., Hossen, J., Sayeed, S., Ho, C.K., Tawsif, K., Rahman, A. and Arif, E., 2018. A survey on cleaning dirty data using machine learning paradigm for big data analytics. *Indonesian Journal of Electrical Engineering and Computer Science*, *10*(3), pp.1234-1243.
2. .Neutatz, F., Chen, B., Abedjan, Z. and Wu, E., 2021. From Cleaning before ML to Cleaning for ML. *IEEE Data Eng. Bull.*, *44*(1), pp.24-41.
3. Ga Young Lee, Lubna Alzamil, Bakhtiyar Doskenov, Arash Termehchy Cite As: arXiv:2109.07127 [cs.DB]
4. Mahdavi, M., Neutatz, F., Visengeriyeva, L. and Abedjan, Z., 2019. Towards automated data cleaning workflows. *Machine Learning*, *15*, p.16.
5. Corrales, D.C., Corrales, J.C. and Ledezma, A., 2018. How to address the data quality issues in regression models: a guided process for data cleaning. *Symmetry*, *10*(4), p.99

Subject Mapping

Machine Learning **– 18AI61 -Course outcomes (Cos) with respect to this PBL**	
CO #1	Choose the learning technique with this basic knowledge
CO #2	Apply effectively ML algorithms for appropriate applications
CO #3	Apply bayesian techniques and derive effectively learnings rules

Project to Program Outcomes (PO) Mapping

Course	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO110	PO111	PO112
Machine Learning	✓	✓	✓	-	-	-	-	✓	✓	✓	✓	✓

Program Outcomes (PO)	
PO1	**Engineering Knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
PO2	**Problem Analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
PO3	**Design/Development of Solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
PO4	**Conduct Investigations of Complex Problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
PO5	**Modern Tool Usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
PO6	**The Engineer and Society:** Apply reasoning informed by contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to professional engineering practice.
PO7	**Environment and Sustainability:** Understand the impact of professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO8	**Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of engineering practice.
PO9	**Individual and Teamwork:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
PO10	**Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
PO11	**Project Management and Finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
PO12	**Life-long Learning:** Recognize the need for and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.