iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50406

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Web Based Machine Learning Automated Pipeline

Prof. Sachin Sambhaji Patil[1], Mahesh Manohar Sirsat[2], Ajitkumar Vishwakarma Sharma[3], Aashish Shahi[4], Omkar Maruti Halgi[5]
[1, 2, 3, 4, 5] *Computer Engineering, Zeal College Of Engineering And Research*

Abstract: ***With the increasing volume, velocity, veracity, and variety of data, it has become critical to have efficient techniques and tools for managing and analyzing data in machine learning.***
Abstraction is a powerful concept that allows users to interact with machine learning algorithms without understanding their technical implementation details. In this project the user will provide the dataset in .csv format the dataset is then processed further to different machine learning preprocessing steps like removing unwanted columns, handling missing values, label encoding, outlier detection and removal, normalization, model building, model prediction, and the result can be downloaded as pdf, tracable pdf and CSV, this all processes gives a result of different model and their respective accuracy so that we can choose the best model for that particular dataset. tracable pdf will be containing all the timestamp of the processes done with their respective result, Apart from client-server model user is also provided a api so that all processes can be implemented in different platforms like c++, java, ruby etc. Overall, this paper highlights the critical role of abstraction in managing the complexity of data and machine learning algorithms, enabling more efficient and effective analysis of large and complex datasets.
Keywords: ***Dataset, Dataset Filtering, Client Server, Pdf Generation, Data Preprocessing.***

I. INTRODUCTION

In today's world, information sharing needs to be fast and efficient. We need tools to take effectively collected data sets from various sources and present and present these visuals in the form of charts, patterns, etc. The tools created process datasets and automate the task of finding various patterns and decoding their semantic structure. The main purpose of integrating tools with datasets is to focus on how the functionality is used rather than how it is implemented to perform further analysis.
According to IDC's AI predictions for 2020 and beyond, IT must invest heavily in data integration, management, and cleansing to effectively use intelligent automation. Data professionals continue to be plagued by the tedious task of data cleansing. Organizations cannot achieve their digital transformation goals without an efficient way to automate data cleansing. [1] IDC Future Scape report finds solving historical data problems in legacy systems can be a significant barrier to entry, especially for large organizations, highlighting the challenges associated with adopting digital initiatives.
According to Morningstar, businesses have spent an estimated $1.3 trillion (USD) on digital transformation initiatives in the past year alone. McKinsey later reported that 70% of his programs were inadequate. Tracking it down at home, outages like this cost businesses over $900 billion.
Businesses cannot afford repeated failures in their digital transformation, regardless of the size of the investment lost. You need clean, standardized data to unlock the benefits of your digital transformation projects, but collecting the data you need in the way you need it can be tedious, expensive, and time consuming.

II. MOTIVATION

The Motivation Behind this project is that those who are beginner's Starting with machine learning they don't understand all that workflow of machine learning to overcome this problem we have made this tool so that the tool can guide or recommend the step by step process in every stage there will be multiple option provided to user according to need, atlast user will come to know how thing works.

Here are some Motivation that are are encorporated in this project:-4

1) *Improve Data Quality:* Outliers can significantly impact the quality and accuracy of data analysis. By removing outliers, your project can help improve the overall quality of the dataset and enable more accurate analysis and decision-making.

2) *Increase Efficiency:* Data preprocessing can be a time-consuming and labor-intensive task, especially when dealing with large datasets. By automating the process of outlier detection and removal, your project can help increase the efficiency of data preprocessing and enable analysts to focus on more complex tasks.
3) *Enhance Data Exploration:* Outliers can sometimes reveal interesting patterns or insights in data that would otherwise go unnoticed. However, outliers can also skew the overall distribution and make it difficult to visualize or explore the data. By removing outliers, your project can help enhance data exploration and visualization, enabling analysts to gain a better understanding of the underlying patterns and relationships in the data.
4) *Address Specific Use Cases:* Depending on the application, removing outliers may be critical for ensuring accurate and reliable results. For example, in finance, outlier detection is essential for identifying potentially fraudulent transactions, while in healthcare, outlier detection can help identify patients at risk for certain conditions. By addressing specific use cases, your project can provide value to a variety of industries and applications.

III. PROPOSED SYSTEM

The system then generates different models from the preprocessed dataset and calculates their respective accuracy. The user can choose the best model for their particular dataset based on the generated results. The system also provides the user with the option to download the results in different formats, including PDF and CSV.

In addition, the proposed system includes a client-server model that allows the user to interact with the system through a web interface.

The system also provides an API that allows the user to implement the preprocessing steps and model building process in different programming languages, including C++, Java, and Ruby. Finally, the system generates a traceable PDF report that includes the timestamp of each process and their respective results.

IV. LITERATURE REVIEW

Automated machine learning (AutoML) has become one of the most dynamic sub-areas in the data science field. Sounds great to a non-machine learning expert, but to a practicing data scientist, it sounds terrifying.

AutoML seems to be able to completely change the way models are generated by removing the need for data scientists based on how they are portrayed in the media.

While some companies like DataRobot want to fully automate the machine learning process, much of the industry is using AutoML as a tool to augment the capabilities of today's data scientists and expand the field.

We make it easy for people just starting out. Now that everything about the system is automated, what is left for the users of these systems? Just get the dataset and check the results.

This level of automation poses potential problems not only for the model, but also for the users who interpret it. Three things turned out to be the most common when studying to become a data scientist.

Here are some gaps that are researched before the project is being carried out, all the gaps are overcomed in this project:-

1) *Overview of Data Preprocessing:* A brief introduction to data preprocessing, which includes the various steps involved in the process such as data cleaning, data transformation, data reduction, and data integration.
2) *Importance of Outlier Detection and Removal:* An explanation of why detecting and removing outliers is an essential step in data preprocessing. You could also include some examples of how outliers can negatively impact data analysis.
3) *Existing Methods for Outlier Detection and Removal:* A review of existing methods for outlier detection and removal, including statistical methods like Z-score and IQR, and machine learning-based methods such as clustering, SVM, and neural networks. You could also discuss the advantages and limitations of each method.
4) *Evaluation Metrics for Outlier Detection and Removal:* A review of evaluation metrics that can be used to assess the effectiveness of outlier detection and removal methods, such as accuracy, precision, recall, F1-score, and AUC.
5) *Applications of Outlier Detection and Removal:* A discussion of various applications of outlier detection and removal techniques in different fields, such as finance, healthcare, and social sciences

V. SYSTEM ARCHITECTURE

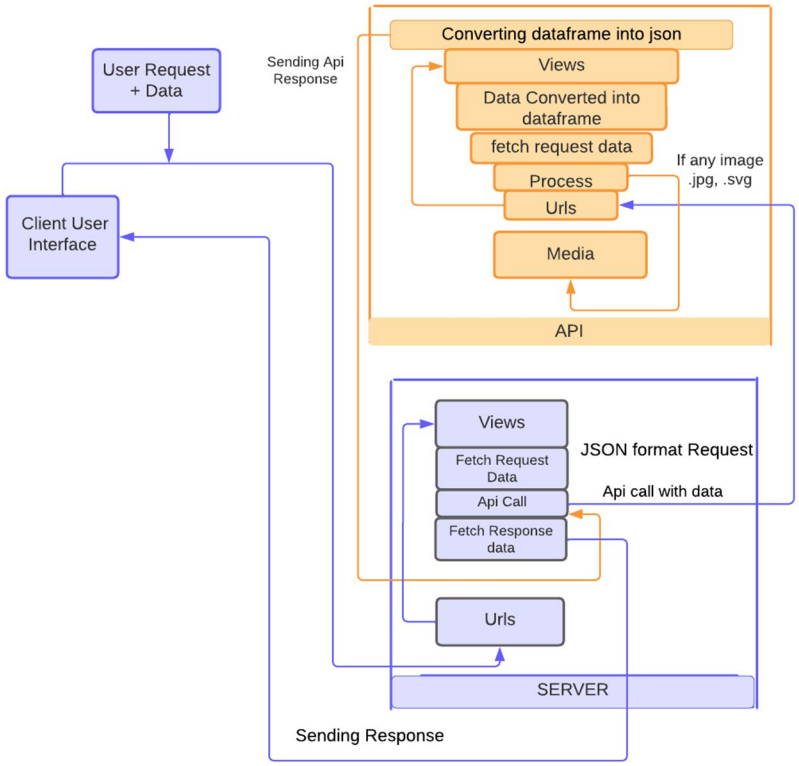


Figure 1 The architecture of the proposed system.

1) *Client:* At first client will load dataset using UI feature provided , function .Which processing are required are managed at that time itself. The Client can give the csv as the input to the system so that the further prcoess can be carried out. The user request and the data is sent as input to the server which is received by the Urls.py file which manages all the urls from the websites.
2) *API:* API will handle all request which are forwarded by client to server or processor to perform analysis.rendering data in form of html is handled by same to provide functionality of dynamic rendering. through the api call all the request is send through the server to the api's urls.py file to manage all the links. view.py convert that data into the dataframe as it is in the json format. all the preprocessing steps is performed in the api like deleting column from the data or table, deleting missing values, encoding with label encoder, performing normalization, removing outliers, building models with different regressor and classifiers in the applications, after that we perform the prediction and compare the results of the model result and compute our results. the preprocessed data is converted into the .json file which is sent to the media file which is the database of the application for temporary purpose. after that all the response is sent to the api call of the server and the response is sent to the client as the output.
3) *Server:* Server is core component of architecture which will process the required dataset into desired one by performing some preprocessing technique, Fetching the data from storage and utilizing it for further analysis of data. In server there is urls.py which collects the dataset from the user and sent it to the views.py in the server part, this data is then transferred to the api through an api call presented at server the api does it work and sent the response in the form of data required where user can fetch in java, python, ruby, c++ etc.
4) *PDF:* At last after processing is finished if user performing all this operations using UI interface) pdf report is send back to client which will record all analysis made and logs of command and also provide visualization to it. There is the csv as we get as the output which is fully tracable that means we can compute the result of the first prediction with the another one so that the result can be varied and hence we can achieve the better results. Further this pdf is static which is begineer friendly and show all the time stamp of the work done in the application with its result.

VI. FUTURE WORK

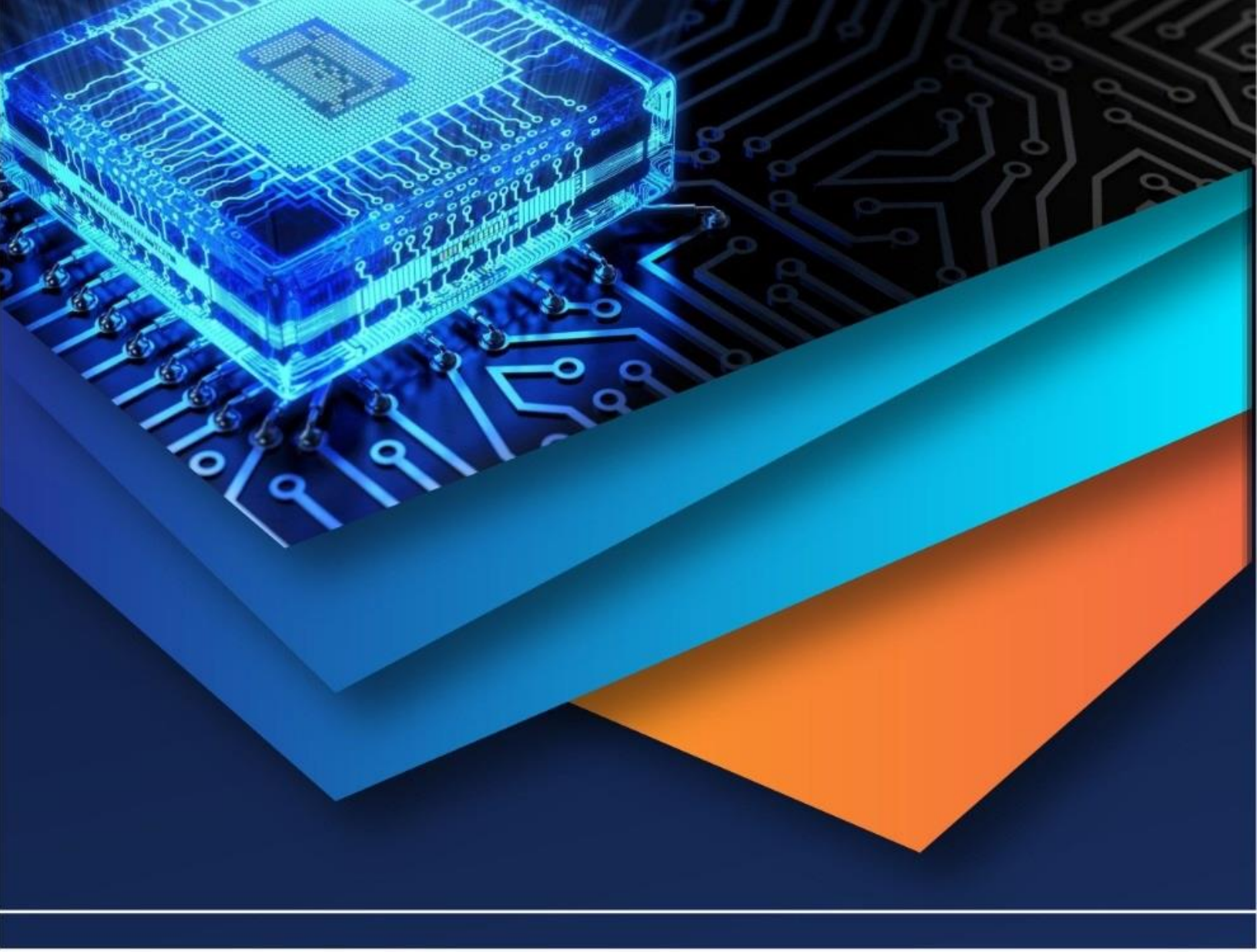
1) *Scaling:* Depending on the size of the CSV data, the project may need to scale to handle large datasets. This can be achieved by optimizing the code, using parallel processing, and leveraging cloud computing.
2) *Automation:* If the project is successful, it may be integrated into a larger system for automation. This could include automating the data collection process, scheduling the pipeline, and integrating the pipeline with other systems.
3) *Integration:* The project can be integrated with other tools such as data visualization tools, dashboards, and APIs to provide more value to the end-users.
4) *Optimization:* The project can be optimized to improve the accuracy of the machine learning models. This can include optimizing hyperparameters, feature selection, and data cleaning.
5) *Extension:* The project can be extended to handle different types of data inputs and machine learning models. This can help to expand its applicability to different use cases and domains.

VII. CONCLUSION

As the project progresses, it effectively improves the visualization of the material. The integrity and consistency of the material is maintained throughout the request response cycle. Provides an important feature that is easy to use when analyzing data and provides meaningful information when extracting data. We present a flow-based view of services through case studies and an overview of the business. One may argue that at the first stages of design, flow-based conceptualization promises to provide Web application development with a more stable basis. This flow-based approach may be used with modern software development methodologies.

REFERENCES

[1] IEEE, “A dataset of attributes from papers of a machine learning conference Algorithm,” 2019.
[2] IEEE, “Missing Data Analysis in Regression,” 2022.
[3] IEEE, “A survey on outlier explanations,” 2022.
[4] I. F. Qayyum and D.-H. Kim, “A Survey of datasets, preprocesssing, modelling mechanisms,” 2022.
[5] C. Fan and M. Chen, “A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data,” 2021



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY