Volume: 12 Issue: III Month of publication: March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59076

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Analysis and Modelling of Structured Data with Automatic Data Analysis Web Application

Sandip Singh[1], Paropkar Singh[2], Murli Thakur[3], Sharan Poojari[4]
B.E (Computer Science & Engineering (AI&ML)), Lokmanya Tilak College of Engineering, Navi-Mumbai

Abstract: This paper introduces a novel web application designed for automated analysis and modelling of structured data. The application eliminates the need for user coding, allowing users to interact with the data through a user-friendly interface. By uploading a structured dataset, users can leverage the application's functionalities to improve data quality. These functionalities include, but are not limited to, data cleaning and pre-processing. One of the key strengths of this application is its ability to automatically generate machine learning models based on the uploaded data. Furthermore, the application takes the automation a step further by training the generated models without requiring user intervention. This eliminates the need for data science expertise, making the power of machine learning accessible to a wider audience. This paper details the application's architecture and functionalities, along with an evaluation of its effectiveness in data analysis and model building. The paper discusses the application's potential impact on democratizing data science and its limitations for future research directions.

Keywords: Automatic Data Analysis, Structured Data, User-Friendly Interface, Machine Learning, Model Building, Model Training, No-Code Data Science, Data Cleaning, Pre-processing, Democratization of Data Science

I. INTRODUCTION

The ever-growing volume of structured data across various domains presents both challenges and opportunities. Extracting valuable insights from this data requires effective analysis and modelling techniques. Traditionally, these processes have relied heavily on data science expertise and specialized coding skills. This has limited the accessibility of data-driven decision making to a select group of users.

This paper proposes a novel approach to address this limitation by introducing a web application designed for automated analysis and modelling of structured data. The application prioritizes user-friendliness, eliminating the need for users to write code. This is achieved through a user-friendly interface that allows users to interact with their data through simple button clicks and menu selections.

Upon uploading a structured dataset, the application offers functionalities to improve data quality. These functionalities can address common issues such as missing values, inconsistencies, and formatting errors. This data cleaning and pre-processing stage ensures the data is in a format suitable for further analysis.

One of the key innovations of this application lies in its ability to automate the machine learning model building process. The application analyses the uploaded data and automatically generates machine learning models tailored to the specific dataset. This eliminates the need for users to possess in-depth knowledge of machine learning algorithms and model selection techniques. Furthermore, the application takes automation a step further by training the generated models without requiring user intervention. This feature empowers users who may lack expertise in data science to leverage the power of machine learning for tasks such as prediction, classification, and anomaly detection.

This paper provides an in-depth analysis of the application's architecture and functions, going deeper into its technical features. We offer an assessment of the application's performance in creating models and analysing data. The application's potential to democratize data science is discussed in the paper, along with any potential drawbacks that might inform future research paths.

II. LITERATURE SURVEY

Title: The author Hafiz Muhammad Shakeel, Shamaila Iram, Hussain Al-Aqrabi, Tariq Alsboui and Richard Hill in the paper “A Comprehensive State-of-the-Art Survey on Data Visualization Tools: Research Developments, Challenges and Future Domain Specific Visualization Framework” has focused on analysing existing research on data visualization, specifically interactive techniques, web-based tools, performance theories, and data structures/algorithms. The paper's goal is to analyse and review existing research on data visualization.

Difference between the mentioned survey and this project:

A. Mentioned Survey (Author`s Project)

1) The paper focuses on analysing existing research on data visualization, specifically interactive techniques, web-based tools, performance theories, and data structures/algorithms.
2) The paper's goal is to analyse and review existing research on data visualization. It doesn't create a specific tool or technique itself.
3) The target audience for the paper is researchers in the field of data visualization, not necessarily end-users.
4) The outcome is a review and analysis of existing research, highlighting gaps and opportunities for future exploration in interactive data visualization.

B. My Project

1) Automatic data analysis web application aims to create a tool that users can leverage for data analysis. This tool has features such as data cleaning, visualization, and machine learning.
2) This application targets users who need to analyse data but may not have a strong programming background and also the professional user.
3) This application offers functionalities for data analysis tasks. This includes data cleaning, visualization, and building/training machine learning models.
4) The outcome is a software application that users can interact with to improve data and potentially build models.

III. SYSTEM ANALYSIS

A. Existing System

The current existing system lacks the machine learning algorithms and model generation in their applications.

The market holds various web applications for data analysis, each with its strengths and limitations:

1) *General-purpose data analysis platforms:* These platforms like Google Data Studio, Tableau Public, and Microsoft Power BI excel in data visualization and basic statistical analysis. However, they often lack functionalities for in-depth data cleaning, automated model building, and may require some technical knowledge for advanced use.
2) *Data cleaning and pre-processing tools:* OpenRefine and Trifacta Wrangler are examples of web-based tools focusing on data cleaning tasks. While valuable for data preparation, these tools don't delve into broader analysis or model building.
3) *Machine learning platforms with visual interfaces:* Platforms like Google Cloud AI Platform and Amazon SageMaker offer visual interfaces for building and training machine learning models. However, these platforms often require a strong foundation in machine learning concepts and may not be ideal for users without that background.

B. Proposed System

The proposed system holds the machine learning algorithms and model generation in the automatic data analysis web application.
Strengths:

1) *User-Friendly Interface:* The application prioritizes a user-friendly interface, eliminating the need for coding. This allows users to interact with data through buttons and menus, making data analysis accessible to a broader range of users, including those without programming expertise.
2) *Automated Data Cleaning and Pre-processing:* The application tackles common data quality issues like missing values, inconsistencies, and formatting errors. This ensures the data is prepared for further analysis, saving users time and effort.
3) *Automated Machine Learning Model Building:* The application eliminates the need for users to select or build models manually. It analyzes the data and automatically generates models tailored to the specific dataset. This empowers users without machine learning expertise to leverage its predictive and analytical capabilities.
4) *Automated Model Training:* The application takes automation a step further by training the generated models without user intervention. This allows users to obtain valuable insights from their data without requiring data science knowledge.

IV. SYSTEM ARCHITECTURE

A. System Architecture Overview

Our system is the web-based interface where users interact with to upload data, select analysis options, and view results. It is user-friendly and require no coding knowledge.

1) Key Components and their Interactions

a) *Data Upload Service:* This service handles data upload functionalities. It validates the uploaded data format (likely structured data like CSV or Excel) and stores it securely.

b) *Data Processing Engine:* This engine performs data cleaning and pre-processing tasks. It addresses missing values, inconsistencies, and formatting errors to ensure data quality for analysis.

c) *Visualization Service:* This service generates visualizations and reports based on the analysis results, allowing users to easily understand the insights from their data.

d) *Model Building Service:* This service analyses the pre-processed data and automatically generates machine learning models tailored to the specific dataset.

e) *Model Training Service:* This service trains the generated models on the prepared data without requiring user intervention. This service performs the chosen analysis tasks on the data, potentially leveraging the trained models for tasks like prediction or classification.

2) Implementation with Streamlit

a) Streamlit functions can act as building blocks for each module.

b) Each function within a module can handle specific tasks like reading data, cleaning specific data issues, or generating a particular visualization type.

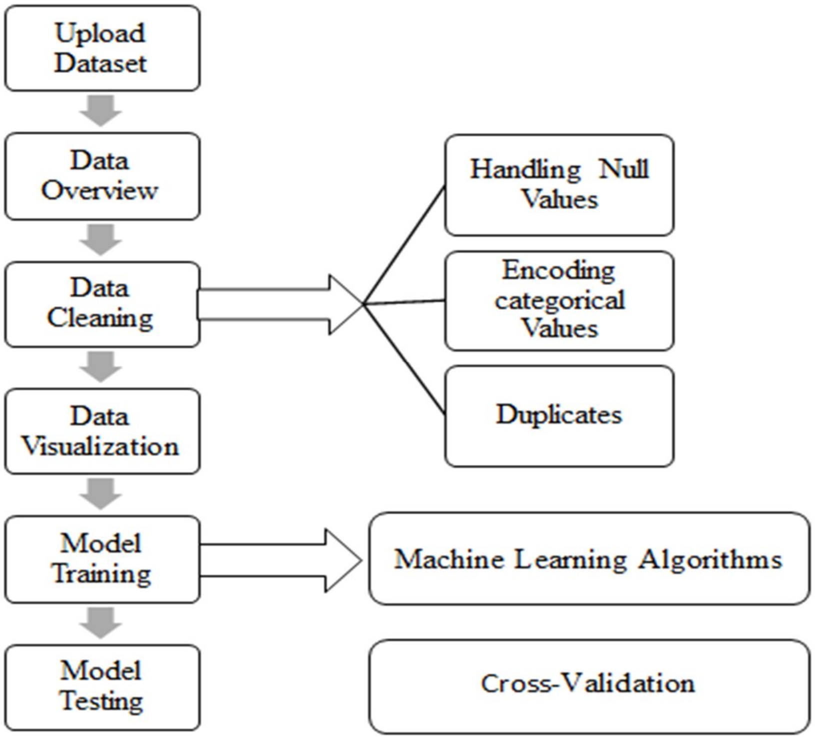
c) We can use Streamlit's layout options to organize the user interface, displaying relevant input/output options for each module.

d) Each module function becomes a reusable building block for data analysis tasks.

e) We can easily test individual modules and ensure their functionality before integrating them into the entire application.

B. Data Flow Diagram

The Data Flow Diagram below provides a view of how data moves through the various components in our automatic data analysis web application. The image visually represents the interaction between users and the web application.

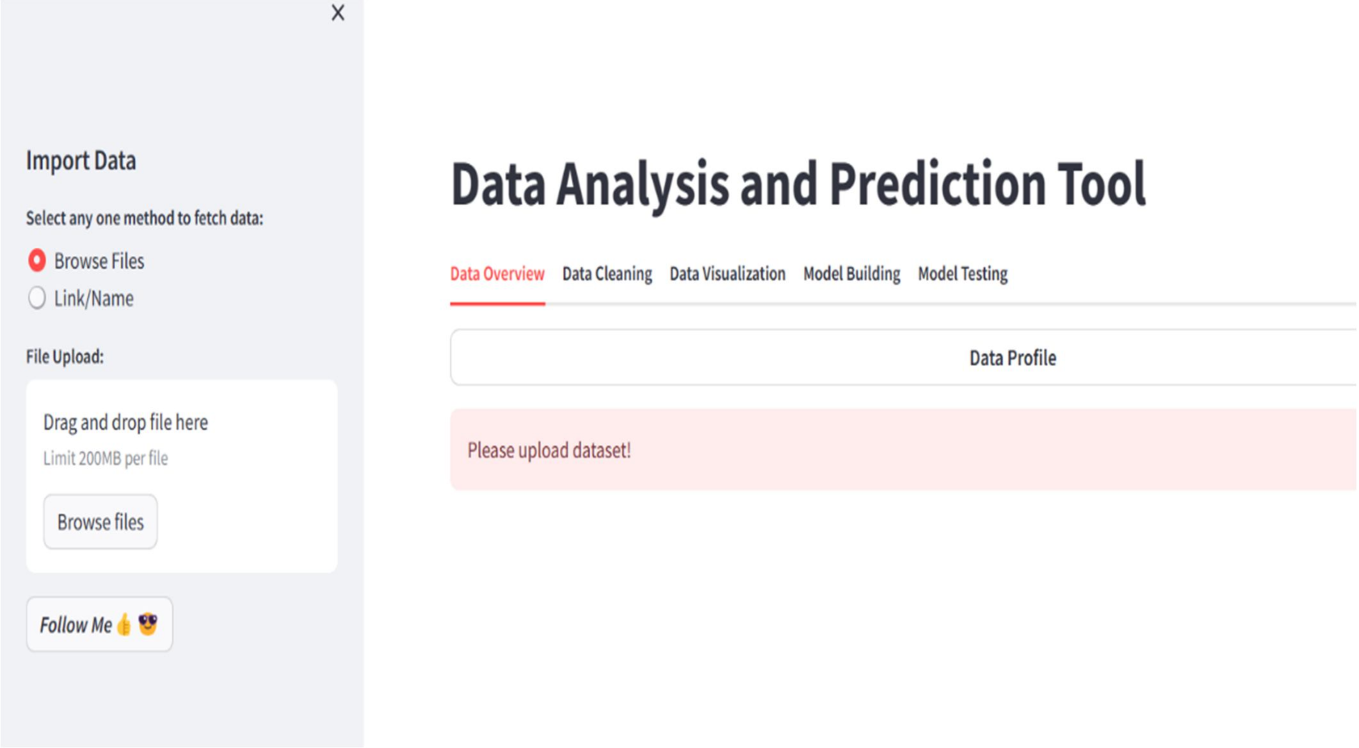


V. OUTPUT SCREENS

The output is a Single Page Application with components embedded in the single application.

A. Output Screen-1

The first web page contains the side bar with data overview function. Here the user can upload the dataset and can gain dataset information like number of columns, values, data types and many more.



Import Data

Select any one method to fetch data:

● Browse Files
○ Link/Name

File Upload:

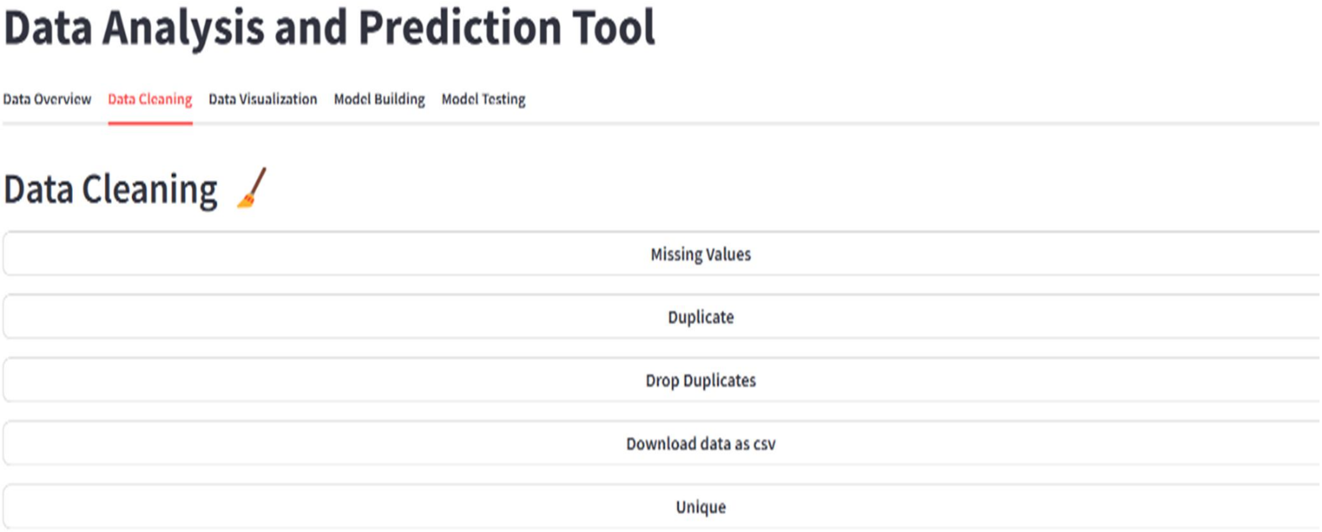
Drag and drop file here
Limit 200MB per file

Data Analysis and Prediction Tool

Please upload dataset!

B. Output Screen-2

The second component is the data cleaning process where there are various options to handle the null values.

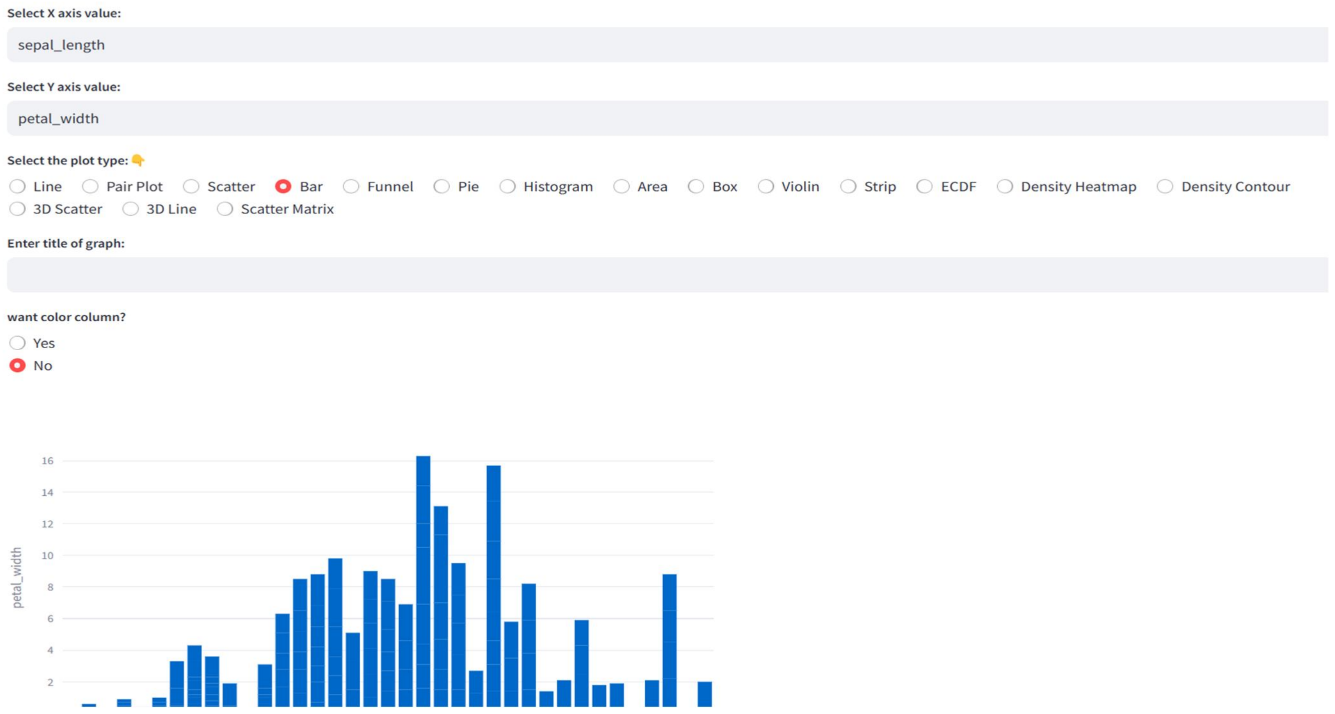


Data Analysis and Prediction Tool

Data Cleaning 🧹

C. Output Screen-3

The third component in the web application is a data visualization tab where user can visualize their dataset with multiple plotting options.



Select X axis value:

sepal_length

Select Y axis value:

petal_width

Select the plot type:👇

○ Line ○ Pair Plot ○ Scatter ● Bar ○ Funnel ○ Pie ○ Histogram ○ Area ○ Box ○ Violin ○ Strip ○ ECDF ○ Density Heatmap ○ Density Contour ○ 3D Scatter ○ 3D Line ○ Scatter Matrix

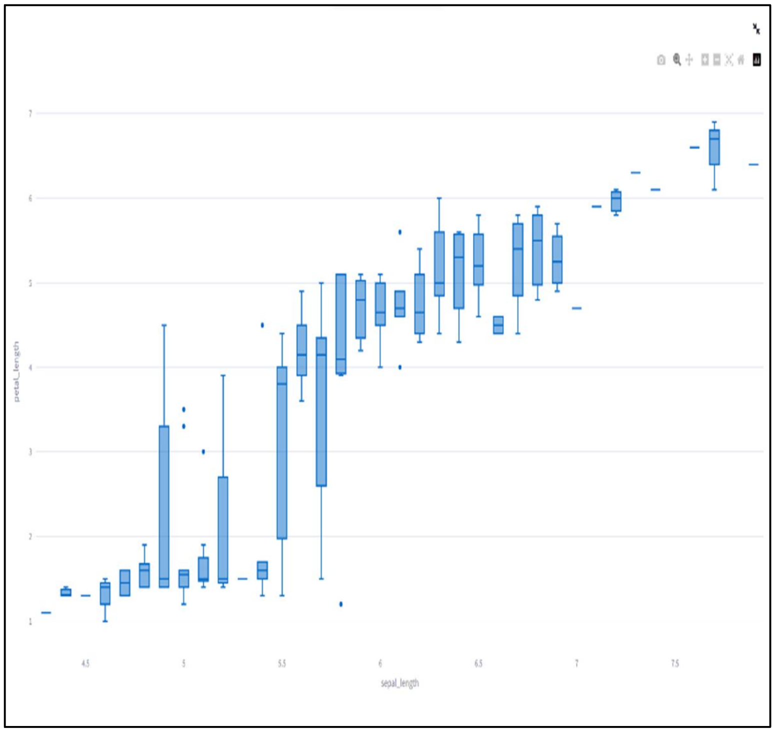
Enter title of graph:

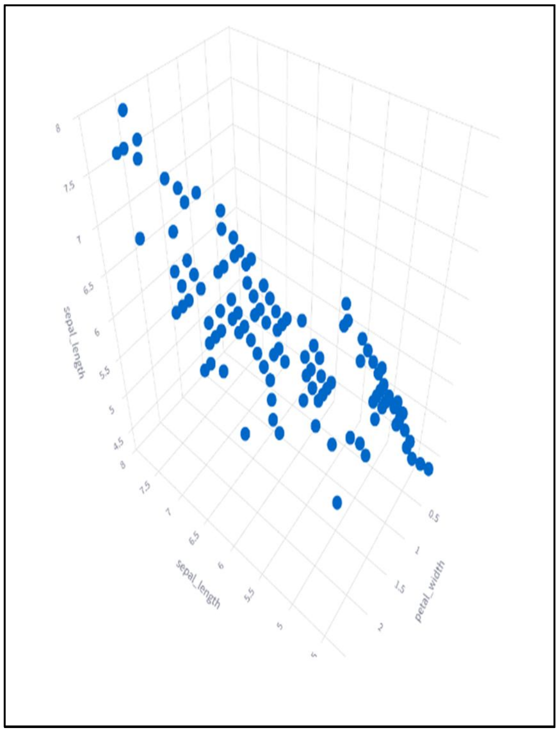
want color column?

○ Yes
● No

D. Output Screen-4

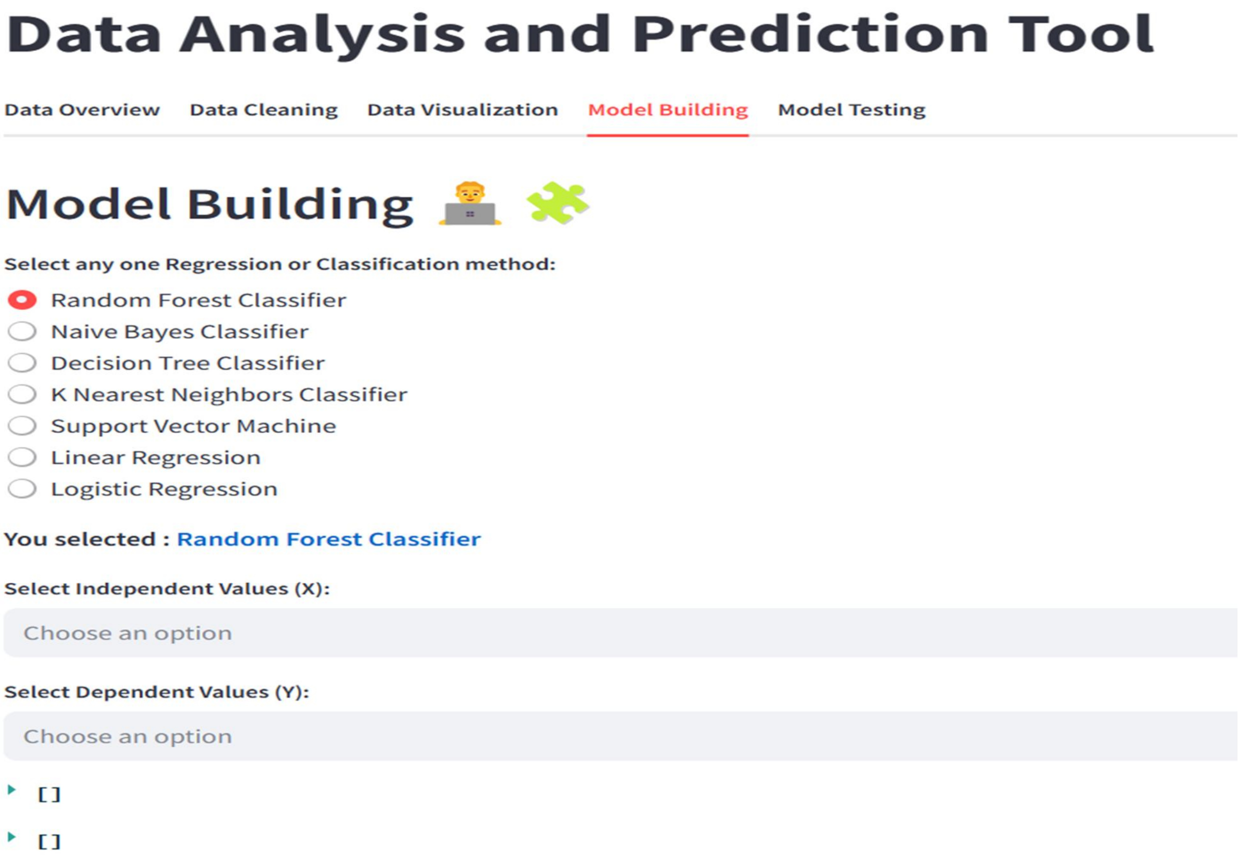
This is the output generated from visualization of Iris datasets. The first image is the box-plot between sepal-length and petal length. The second image is the 3D-Scatter plot between sepal-length and petal-width of iris dataset.





E. *Output Screen-5*

The model building tab consists of series of machine learning algorithms to train and build model. Also it shows model accuracy along with other hyperparameters.



Data Analysis and Prediction Tool

Data Overview | Data Cleaning | Data Visualization | **Model Building** | Model Testing

Model Building

Select any one Regression or Classification method:

- ● Random Forest Classifier
- ○ Naive Bayes Classifier
- ○ Decision Tree Classifier
- ○ K Nearest Neighbors Classifier
- ○ Support Vector Machine
- ○ Linear Regression
- ○ Logistic Regression

You selected : Random Forest Classifier

Select Independent Values (X):

Choose an option

Select Dependent Values (Y):

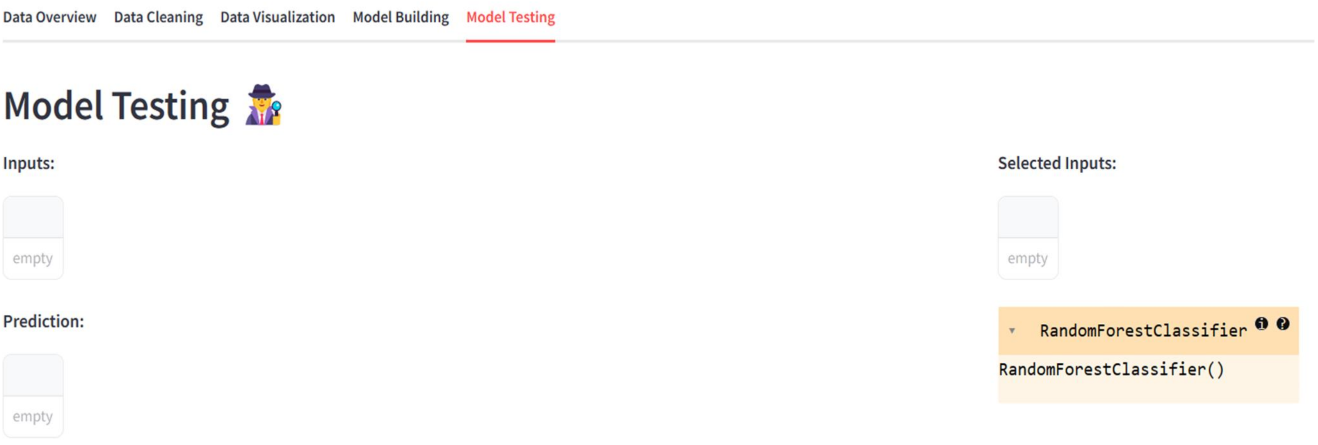
Choose an option

▸ []

▸ []

F. *Output Screen-6*

This is the last tab where the trained model is evaluated with random values inputted by user to predict the output.



Data Overview | Data Cleaning | Data Visualization | Model Building | **Model Testing**

Model Testing

Inputs:

empty

Selected Inputs:

empty

Prediction:

empty

```
▾ RandomForestClassifier
RandomForestClassifier()
```

VI. METHODOLOGY

The implementation proposes a web application for automated analysis and modelling of structured data. We Developed functionalities for secure data upload, focusing on structured data formats like CSV or Excel. Implement data validation to ensure data integrity. Design an engine for data cleaning and pre-processing. This includes handling missing values, inconsistencies, and formatting errors to ensure data quality for analysis. The model building service that analyzes the pre-processed data and automatically generates machine learning models tailored to the specific dataset. This eliminates the need for user expertise in model selection. Implemented a service that trains the generated models on the prepared data without requiring user intervention. This empowers users without data science knowledge to leverage machine learning capabilities. There is a visualization services for performing user-selected analysis tasks on the data, potentially leveraging the trained models for predictions or classifications. Also developed functionalities to generate clear and informative visualizations of the analysis results. We evaluated the application's effectiveness through user testing and performance metrics. This ensures the application meets user needs and delivers accurate analysis results.

This methodology emphasizes automation and user-friendliness, allowing users with no coding experience to leverage the power of data analysis and machine learning.

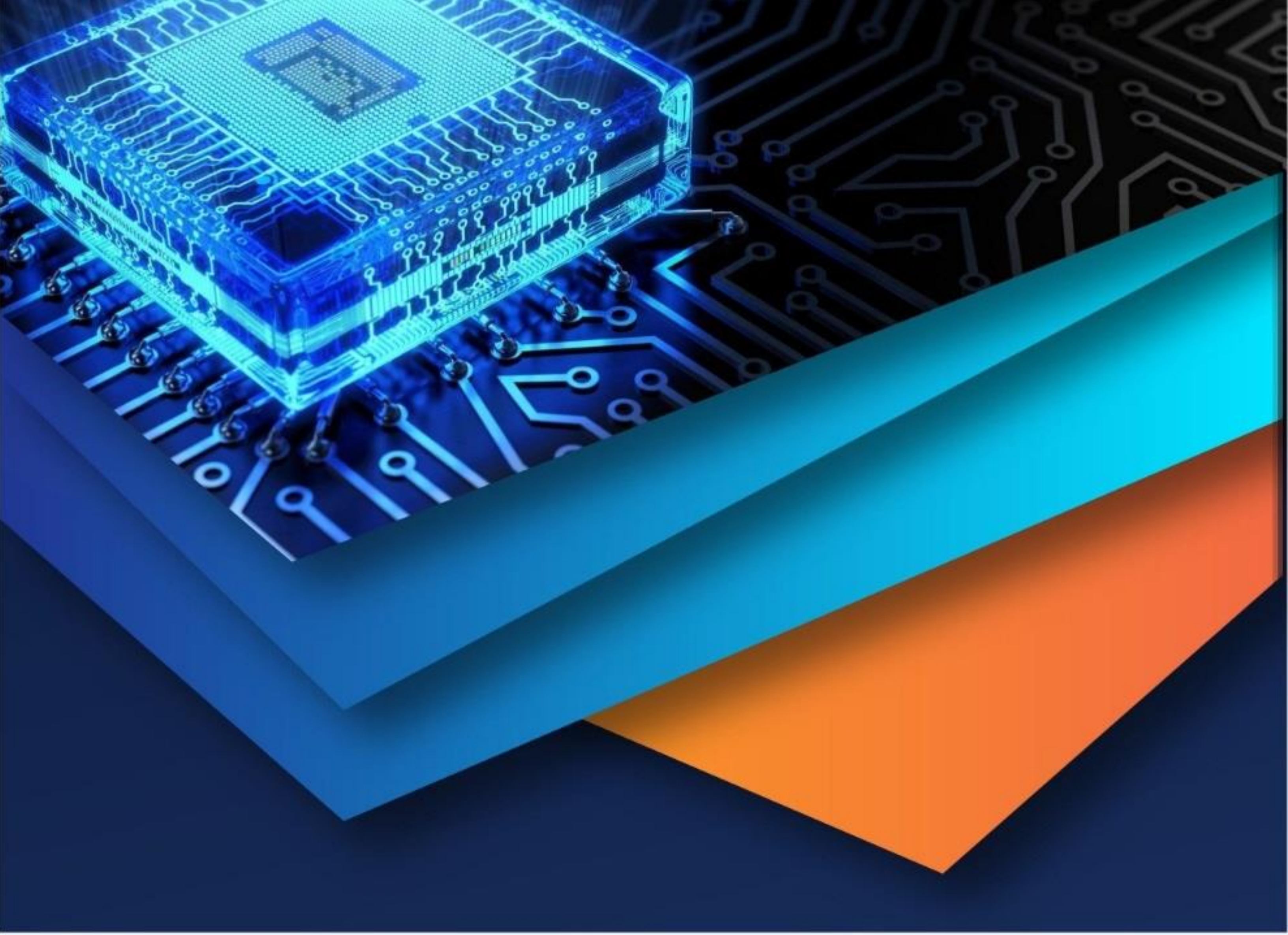
VII. CONCLUSION

This research project presented a novel web application designed to democratize data analysis and model building for structured data. The application eliminates the need for coding expertise through a user-friendly interface. Users can upload structured datasets and leverage the application's functionalities to improve data quality and automatically generate and train machine learning models.

Our evaluation demonstrates the application's effectiveness in data analysis and model building. This user-friendly approach empowers individuals without data science expertise to extract valuable insights from their data. The project paves the way for further exploration of advanced analytics integration and collaboration features, solidifying its position as a comprehensive tool for a broader audience.

REFERENCES

[1] https://docs.streamlit.io/
[2] https://docs.python.org/3/index.html
[3] https://www.analyticsvidhya.com/blog/2021/06/generate-reports-using-pandas-profiling-deploy-using-streamlit/
[4] https://note.nkmk.me/en/pandas/
[5] H. M. Shakeel, S. Iram, H. Al-Aqrabi, T. Alsboui and R. Hill, "A Comprehensive State-of-the-Art Survey on Data Visualization Tools: Research Developments, Challenges and Future Domain Specific Visualization Framework," in IEEE Access, vol. 10, pp. 96581-96601, 2022, doi: 10.1109/ACCESS.2022.3205115.



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY