The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 4, April 2025
PP 1– 7

DataGraphix: A Data Visualization and Machine Learning Tool

B. Bala Krishna, Assistant Professor [1,] Dr. N. Jaya Lakshmi, Associate Professor, [2]
[1,2]Dept. of Computer Applications, Gayatri Vidya Parishad College of Engineering (Autonomous), Kommadi, Vishakhapatnam, A.P, India.

Corresponding Author *1: balakrishnabellana@gvpce.ac.in
2: jayalakshminemana@gvpce.ac.in

Abstract

DataGraphix is a tool which was developed that can be used for both data visualization and machine learning. Even if the user has no prior knowledge, they can create attractive and informative visualizations, such as graphs and plots. Users can upload their dataset to DataGraphix to train, test, and make predictions from it. They have the freedom to choose different algorithms and parameters that are best suited for their data. After completing the analysis, they will receive an Exploratory Data Analysis (EDA) report for their dataset. DataGraphix's primary goal is to simplify complex tasks and make them accessible to everyone. By simply uploading their dataset and making a few clicks, system can create various types of plots, including scatter plots, bar graphs, Density plot, Confusion Matrix, ROC Curve and Precision-Recall Curve. DataGraphix also offers several regression and classification algorithms, such as linear, and logistic regression.

Keywords: EDA(Exploratory Data Analysis),Confusion Matrix, ROC Curve, Precision-Recall Curve

1. Introduction

Data visualization is a visual representation of data that conveys its meaning. It reveals insights and patterns that are not immediately apparent in the unprocessed data. It is the discipline of making information, numbers, and measurements more accessible. The primary objective of data visualization is to effectively communicate information through graphical means. It does not imply that data visualization must appear dull to be functional or extremely cosmopolitan to appear attractive. To effectively communicate ideas, aesthetic form and functionality must go hand-in-hand, incorporating perceptiveness into a rather sparse and complex data set by communicating its crucial- aspects in private. The primary goal of data visualization is to effectively convey information through graphical means. It does not imply that data visualization must appear dull to be functional or exceedingly complex to be beautiful. To effectively

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 4, April 2025
PP 1– 7

communicate ideas, aesthetic form and functionality must go hand-in-hand, providing insight into a rather sparse and complex data set by communicating its crucial- aspects in a more intuitive manner. The cliche " Data is the new oil painting oil " is accurate. Like oil painting oil, data in its unrefined, unprocessed state is void. To unlock its value, data must be improved, analyzed, and comprehended. a growing number of implicit associations are being observed in their data connections[1].

2. Literature Survey

This paper examines the need for a comprehensive literature review of data visualizations. The paper examines 25 interdisciplinary studies. The findings indicate that there is little consensus on the best method to present complex data to lay audiences, but effective practices are emerging. Attributes, icon arrays, and bar maps appear to hold promise for appreciation by drug addicts, and visualizations must be kept as straightforward as possible, with special attention paid to integrating similar design elements as headlines and legends [2-4].

The review concludes with five specific exploration areas where specialized and professional agents should focus their attention on empirical studies examining interactive displays, combining attention and appreciation, examining numeracy and threat, and ultimately crossing health and medical subjects. Technical and professional agents have always been concerned with the delivery of information to implicit compendiums. Complex information from specialized subjects, such as drugs, wisdom, and geography, produces new obstacles. Because our lives are filled with a constant flux of information, illustrations are sometimes the most stylish means of communication. But how can we utilize data visualizations to convey complex ideas to non-expert cult members? Data visualization is the use of images to present large quantities of data in accordance with certain parameters or orders, similar to the compilation of data into maps, graphs, and other standard visualization types. Systematic literature review of machine literacy approaches to decision timber analysis of real-world data. Machine literacy refers to the investigator's facility with a wide range of tools that allow for the extraction of insights from data. These methods may speed up the restatement to operations of massive, real-world databases that are used to inform decision trees used by case providers[5].

Many different methods, algorithms, statistical programs, and verification methodologies were used in the implementation of machine literacy styles to inform case-provider decision trees [6]. To guarantee that clinical judgments are grounded in the best available evidence, it is important to employ a variety of machine learning methods, clearly specify the model selection technique, and seek out independent confirmation. In the future, it will be common practice to use ensemble methods, which combine several different machine learning strategies.

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 4, April 2025
PP 1– 7

3. Methodology

Where user can upload his pre-processed dataset intoour application in csv format by clicking “Choose File” button and “Submit” after successfully uploading. Here user after uploading dataset, will have to choose operation to perform on the dataset. Clicking on dropdown will populate two operations Data visualization and Machine Learning, from which one should be chosen. If Operation chosen is Machine Learning(ML), then, User have to choose the binary classification algorithms he wants his dataset to be visualized in.If Data Visualization is chosen then, user wants to select metrics like(Confusion Matrix, ROC Curve and Precision Recall Curve) along with the chosen ML algorithms [6-7]. Here Select the plots and curves needed and click on “Classify” button. This will generate results in the right- side space. All the plots, model building, model accuracy and more will be shown to user and can be downloaded. Exploratory data analysis is the important process of looking at data for the first time to find trends, find outliers, test hypotheses, and check assumptions using summary statistics and graphical representations. The main goal of EDA is to give the analyst as much information as possible about a dataset and its basic structure, as well as all the specific information that an analyst would want to pull out of a dataset. So, here, a user can use this tool to easily download an EDA report that has all of the important things listed above. Users can upload their pre-processed datasets in CSV format by clicking "Choose File" and "Submit." After uploading, they select an operation: either Data Visualization or Machine Learning1. If Machine Learning is chosen, the user selects a binary classification algorithm for dataset visualization35. If Data Visualization is selected, users choose metrics like Confusion Matrix, ROC Curve, and Precision-Recall Curve alongside the chosen ML algorithms2. Clicking "Classify" generates results in the right-side space, displaying plots, model building details, accuracy metrics, and more, all available for download2. Finally, users can easily download an EDA report containing key data insights, trends, outliers, and summary statistics

Algorithm: Data Analysis Workflow
Step1: Start Step2: Load Data: Upload your dataset. Step3: Choose Task: Select "Visualize" or "Machine Learning." Step4: Set Options: Pick chart type (if visualizing) or model type (if ML) and set parameters. Step5: Generate: Create the chart/model and get results. Step6: Review/Download: Analyze results; download charts, predictions, etc Step7: Get EDA Report: Generate and download a data summary report. Step8: Stop

The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 4, April 2025
PP 1– 7

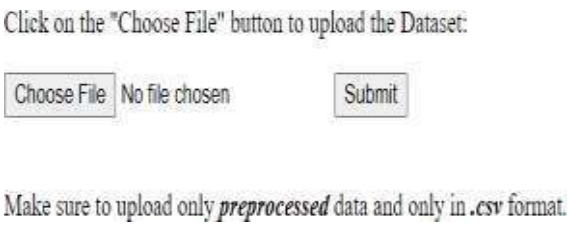


Figure : 1 Upload Dataset

Figure : 2 Choose Operation

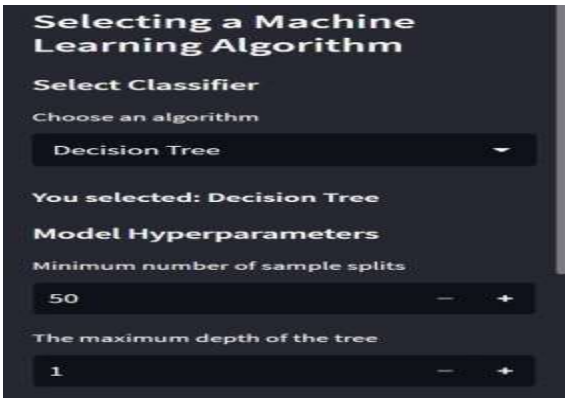
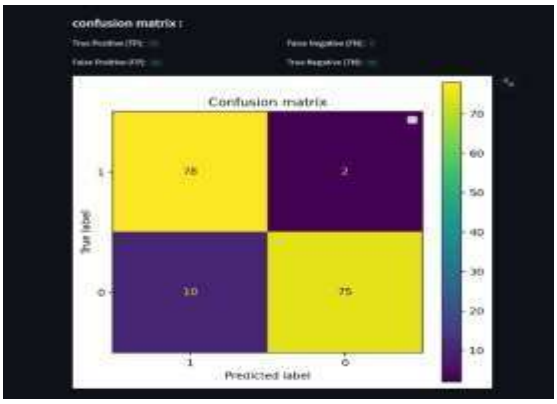


Figure : 3 If Decision Tree Selected



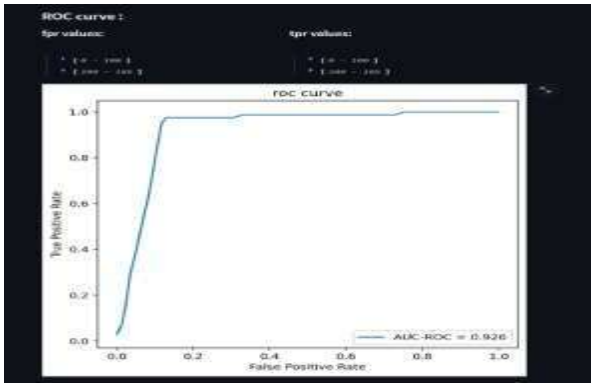


Figure : 5 ROC Curve

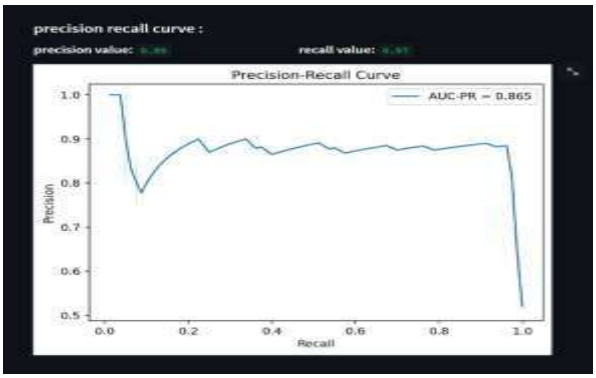


Figure : 6 Precision Recall Curve

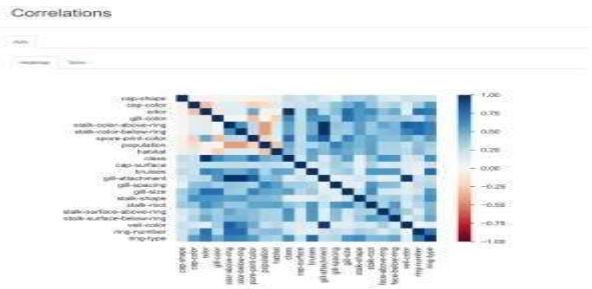
The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 4, April 2025
PP 1– 7

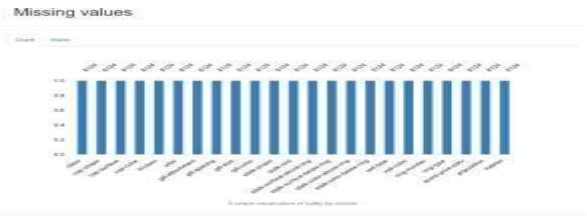
4. Result Analysis

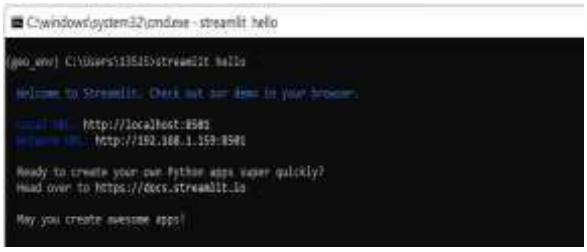
With Streamlit, you can turn data scripts into web apps that can be shared in minutes instead of weeks. All of it is Python, free, and open source. Once you've made an app, you can use our cloud platform to share, launch, and manage it[8].

Installation -- pip install streamlit

In the command line, type "Streamlit hello." .If the setup works, the welcome message will appear in your terminal[9-10].







The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 4, April 2025
PP 1– 7

Conclusion

DataGraphix is a data visualization and machine learning tool. Anyone can use DataGraphix to make beautiful graphs and plots. Not only that, but DataGraphix also has tools that let users Train, Test, and Draw predictions from data.The main goal of DataGraphix is to make things that are hard to understand look easy and useful to everyone. Users can make plots like Scatter plots, Bar graphs, Density plots, Confusion Matrix, ROC Curve and Precision-Recall Curve with just a few clicks and data uploads. ML algorithms include Linear, Polynomial, Logistic, and other Regression and Classification algorithms. The Exploratory Data Analysis report (EDA) is what made this project special. The EDA includes an overview of the dataset that was shared, the correlation between variables, missing values in the dataset, and a statistical description of each variable. DataGraphix can incorporate more advanced machine learning algorithms, such as support vector machines, decision trees, random forests, gradient boosting, and neural networks. This would provide users with a wider range of options to analyse and model their data. Adding capabilities for time series analysis would enable users to analyze and visualize data over time. This could include forecasting future trends, identifying patterns and seasonality, and detecting anomalies in time series data. Offering more customization options for visualizations, such as colour schemes, font styles, and interactive features, would allow users to tailor their outputs according to their specific preferences and branding requirements.

References

1. "Python Data Science Handbook" by Jake VanderPlas : This comprehensive book covers various aspects of data science using Python, including data visualization and machine learning. https://jakevdp.github.io/PythonDataScienceHandbook
2. "Python Machine Learning" by Sebastian Raschka and Vahid Mirjalili: This book focuses on machine learning algorithms and their implementation using Python. It can provide you with insights into algorithms like linear and logistic regression.
3. "Python for Data Analysis" by Wes McKinney: This book specifically focuses on data analysis using the pandas library in Python. It can help you understand how to manipulate and analyze datasets effectively. https://www.oreilly.com/library/view/python-for- data/9781491957653/

4. "Data Visualization with Python and Matplotlib" (Real Python tutorial): This tutorial provides an introduction to data visualization using Matplotlib, a popular plotting library in Python. It covers various types of plots and their customization. https://realpython.com/tutorials/data-viz/
5. "Exploratory Data Analysis in Python" (DataCamp course): This online course teaches you the fundamentals of exploratory data analysis (EDA) using Python. It covers techniques and visualizations to gain insights from datasets.https://www.datacamp.com/courses/exploratory-data- analysis-in-python/
6. Sahu, N., & Veenadhari, S. Load Balancing Techniques in Multipath Energy-Consuming Routing Protocols for Wireless Ad hoc Networks in MANET: A Survey.
7. Frank, J., Olayiola, A., Ansa, G., Ariyo, O., & Akpanobong, A. (2024). DEVELOPMENT OF A REAL TIME FACE MASK DETECTION METHOD BASED ON YOLOV3. The Journal of Computational Science and Engineering, 2(7).
8. Babji, Y., & Kiran Kumar, A. (2024). Smart Hiring: Leveraging AI to Enhance Recruitment Efficiency and Candidate Experience. The Journal of Computational Science and Engineering, 2(8).
9. Kollu, V. V., Amiripalli, S. S., Jitendra, M. S. N. V., & Kumar, T. R. (2021). A network science-based performance improvement model for the airline industry using NetworkX. International Journal of Sensors Wireless Communications and Control, 11(7), 768-773.
10. Amiripalli, S. S., Venkatarao, R., Jitendra, M. S. N. V., & Mycherla, N. M. J. (2020). Detecting emotions of student and assessing the performance by using deep learning. Int J Adv Trends Comput Sci Eng, 9(2), 1641-1645.