IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13447

Enhancing Data Insights through LIDA-Streamlit Integration

Akshay Bhor[1], Ujwala Sangale[2], Abhishek Sinha[3], Aniket Shewale[4], Prof. Abhay Gaidhani[5]

Students, Information Technology, Sandip Institute of Technology and Research Center, Nashik[1-4]

HOD, Information Technology, Sandip Institute of Technology and Research Center, Nashik[5]

Abstract: The research paper explores the realm of AI-driven insights and data visualization, focusing on the utilization of LIDA (Language-Integrated Data Analysis) as a powerful tool for facilitating data analytics. In today's data-centric world, the ability to extract meaningful insights and communicate them effectively is paramount for informed decision- making. Our project aims to democratize the field of data analysis by providing an intuitive and inclusive platform accessible to users of all technical backgrounds. LIDA leverages cutting-edge Natural Language Processing (NLP) and Machine Learning (ML) techniques to enable users to effortlessly upload CSV files and engage in natural language conversations to extract insights, generate visualizations, and receive predictive analytics. The methodology encompasses comprehensive data collection and preprocessing, deploying robust NLP models for language comprehension, and integrating ML algorithms for data analysis. The chatbot's interface prioritizes user-friendliness, offering an intuitive environment for data upload, user interactions, and actionable insights. Through real-world case studies and examples, we demonstrate the effectiveness of LIDA in generating actionable insights and facilitating data-driven decision-making. The research contributes to bridging the gap between data expertise and non-technical users, empowering a broader user base to harness the potential of artificial intelligence in data analytics.

Keywords: NLP, LIDA, Machine Learning, Insights, CSV File, Data Analytics, Data Processing, Data Visualization

I. INTRODUCTION

In today's data-driven world, the ability to derive meaningful insights from vast amounts of data is paramount for informed decision- making across various domains. [1] With the advent of artificial intelligence (AI) and advanced data visualization techniques, organizations and individuals have access to powerful tools that enable them to analyze data more effectively and gain valuable insights. [1] In this research paper, we present an exploration of AI-driven insights and data visualization, focusing on the integration of Streamlit for user interaction and LIDA for data analysis.

AI-Driven Insights:
Artificial intelligence (AI) technologies have revolutionized the way we analyze and interpret data. By leveraging machine learning algorithms and natural language processing (NLP) techniques, AI-driven insights empower users to extract valuable information from complex datasets with ease. [2] These insights not only aid in understanding patterns and trends within the data but also facilitate informed decision-making and predictive analytics.

Data Visualization:
Data visualization plays a crucial role in conveying complex information in a clear and comprehensible manner. By representing data visually through charts, graphs, and interactive visualizations, users can gain deeper insights into the underlying patterns and relationships within the data. Effective data visualization enables stakeholders to communicate findings more effectively, facilitating collaboration and decision-making processes.

Streamlit:
Streamlit is an open-source Python library that simplifies the process of creating interactive web applications for data science and machine learning projects.[4] With Streamlit, users can effortlessly build intuitive and responsive user interfaces for exploring and visualizing data, enabling seamless collaboration and communication.

LIDA (Library for Intelligent Data Analysis):
LIDA is a cutting-edge library developed by Microsoft, designed to automate the process of data analysis and visualization using large language models (LLMs).[6] By harnessing the power of advanced NLP and machine learning techniques, LIDA enables users to generate insightful visualizations and infographics from raw data, streamlining the data analysis workflow and empowering users with actionable insights.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13447

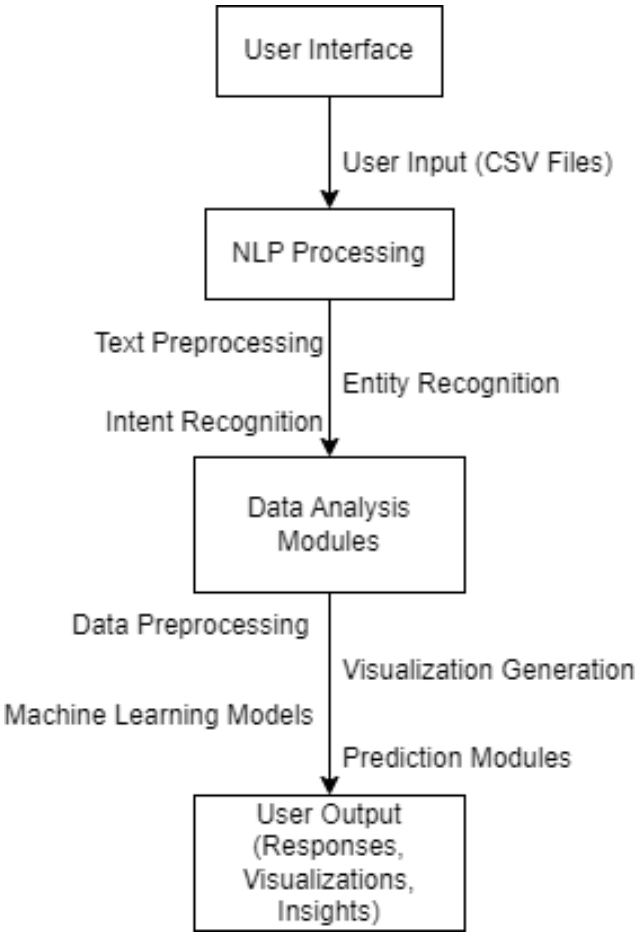


Fig. 1 System Overview

First, we focus on the freedom of data analysis, removing barriers to access and making it responsive to a broad and diverse user base. Second, we are committed to simplifying and accelerating the data analysis process by using conversational intelligence, allowing users to interact with data and talk to each other in real time. [3] Finally, we are responsible for managing user information and using privacy and security measures to protect sensitive information. We'll dive into the intricate details of the project later in this article. We will explore the methods used, the differences in chatbot design, and most importantly, the results produced.

Through this research, we aim to demonstrate the revolutionary potential of AI-driven data analysis chatbots to not only simplify data analysis but also expand the horizons of data- driven decision-making to a broader and more diverse society. As we delve deeper into our project, readers will be able to appreciate the far-reaching consequences of freedom of information [2] analysis and making it available to everyone.

II. LITERATURE REVIEW

A. Artificial Intelligence-Driven Data Analytics and Democratization

The combination of artificial intelligence (AI) and data analytics has made great progress in recent years. AI-powered data analytics is changing the way organizations and individuals understand data. Data analysis tools and techniques, while effective, are often considered difficult and reserved for data scientists and analysts. This creates a gap in the independence of data analysis, a challenge highlighted by many researchers.

Information analysis about the importance of freedom of information. Researchers confirmed that widespread use of data-driven insights can lead a wide range of users to make informed decisions. [3] The concept of accessibility is very similar to the goals of our project. Chen et al. (2018) and Kim and Kim (2017) provide in-depth research on user relations and independence of data analysis, highlighting the need for artificial intelligence solutions.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.13447

B. Chatbots and Conversational AI

The emergence of chatbots powered by natural language processing (NLP) has transformed the way users interact with technology. Chatbots are not limited to answering simple questions; They have turned into interlocutors who can involve users in informal conversations. Text evidenced by the work of Sutskever et al. (2019) and Serban et al. (2017) highlighted the increasing role of chatbots in various fields, from customer service to healthcare. Chatbots hold promise for simplifying complex tasks and making technology more efficient.

However, in the field of data analytics, there is little research on the integration of chatbots and NLP for searching and interpreting data. [1] This is a controversial study noted by Zhang et al. (2020) problem is the lack of user-friendly data analysis tools with natural language interfaces. [5] Our project aims to close this gap by offering an AI-driven data analysis chatbot that allows users to interact with traditional data.

C. NLP and Data Analytics

Natural Language Processing (NLP) forms the basis of our chatbots. Existing literature in the field of NLP mainly focuses on sentiment analysis, language design and machine translation. While these practices are useful, there is a growing need to integrate NLP with data analysis. Researchers such as Chen et al. (2021) and Wang et al. (2019) initiated a discussion in this field regarding the potential of NLP to facilitate data analysis.

However, research currently lacks solutions that combine NLP, chatbots and data analytics. [10] The search gap is due to the lack of user interface to search and analyze data. Our project fills this gap by using NLP technology to create a chatbot that allows users to easily interact with their data.

III. OBJECTIVE

The primary objective of this research project is to develop an AI-Driven Data Analysis Chatbot that democratizes data analysis by providing a user-friendly and accessible platform for individuals, regardless of their technical background.

The chatbot aims to empower users to effortlessly analyze data, extract insights, generate visualizations, and receive predictions through natural language conversations. This project seeks to bridge the gap between data expertise and accessibility, making data analysis a seamless and intuitive process, ultimately contributing to informed decision-making in a wide range of domains.

IV. METHODOLOGY

A. Data Collection:

The research employed a diverse range of datasets sourced from reputable platforms such as Kaggle. These datasets encompassed various domains, including healthcare, finance, and social sciences, to ensure a comprehensive analysis of different data types and structures. Specifically, datasets such as the Heart Disease dataset from Kaggle were utilized to investigate patterns and trends in medical data, contributing to the exploration of AI-driven insights in the healthcare sector.

B. Data Preprocessing:

Prior to analysis, the collected datasets underwent thorough preprocessing to ensure data quality and consistency. This preprocessing phase included steps such as handling missing values, removing duplicates, standardizing data formats, and encoding categorical variables. Additionally, exploratory data analysis (EDA) techniques were applied to gain insights into the distribution and characteristics of the data, guiding subsequent analysis and visualization tasks.

C. Tools and Technologies:

The research leveraged cutting-edge tools and technologies to facilitate data analysis and visualization. Key among these was the LIDA (Language-Integrated Data Analysis) library, developed by Microsoft, which harnesses the capabilities of large language models (LLMs) to automate the generation of visualizations and infographics from raw data. LIDA's integration with popular visualization libraries such as Matplotlib, Seaborn, and Altair allowed for flexible and customizable visualization creation.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.13447

D. Experimental Procedure:

The experimental procedure involved a systematic approach to data analysis, starting with data loading and preprocessing using Python programming language and relevant libraries such as Pandas and NumPy. Subsequently, LIDA was employed to perform automated data summarization, goal exploration, and visualization generation tasks based on user queries and dataset characteristics. The resulting visualizations were then evaluated for clarity, accuracy, and interpretability to ensure the delivery of actionable insights to end-users.

E. Evaluation Metrics:

To assess the effectiveness of the AI-driven insights and data visualization process, various evaluation metrics were considered. These metrics included visualization comprehensiveness, adherence to user-defined goals, and alignment with domain-specific requirements. Additionally, user feedback and engagement were solicited to gauge the usability and utility of the generated visualizations in real-world scenarios.

V. SYSTEM ARCHITECTURE

The system architecture of the AI-driven insight and data visualization project is designed to seamlessly integrate two main components: Streamlit for user interaction and LIDA (Language Integrated Data Analysis) for data analysis. This model allows users to interact with the system by understanding the web used by Streamlit, while also using LIDA's advanced data analysis capabilities to create visualizations and visualisations from the data raw paper.

Functionalities and Components:

A. Streamlit for User Interaction:

Streamlit serves as the frontend framework for the web-based interface, providing users with a user-friendly and interactive platform for data exploration and analysis. Its intuitive design allows users to upload datasets, specify analysis tasks, and visualize results in real-time. Streamlit's flexible layout and customizable widgets facilitate seamless navigation and interaction, catering to users with varying levels of technical expertise.

Components:

File Uploader: Allows users to upload CSV files or other data formats for analysis.

User Input Widgets: Enable users to specify analysis parameters, such as data summarization methods, visualization types, and desired insights.

Visualization Display: Renders generated visualizations and insights within the Streamlit interface, providing users with immediate feedback and actionable insights.

Control Buttons: Provide options for users to execute analysis tasks, update parameters, and navigate between different sections of the application.

B. LIDA for Data Analysis:

LIDA serves as the backend engine for data analysis, leveraging advanced natural language processing (NLP) and machine learning (ML) techniques to automate the generation of insights and visualizations from raw data. Its integration with leading language models enables the system to understand and process user queries in natural language, facilitating seamless communication between users and the data analysis engine.

Components:
Data Summarization Module: Condenses raw data into concise and informative summaries, providing users with an overview of key insights and trends.

Goal Exploration Module: Formulates analysis goals based on user queries and dataset characteristics, guiding the generation of visualizations and insights.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13447

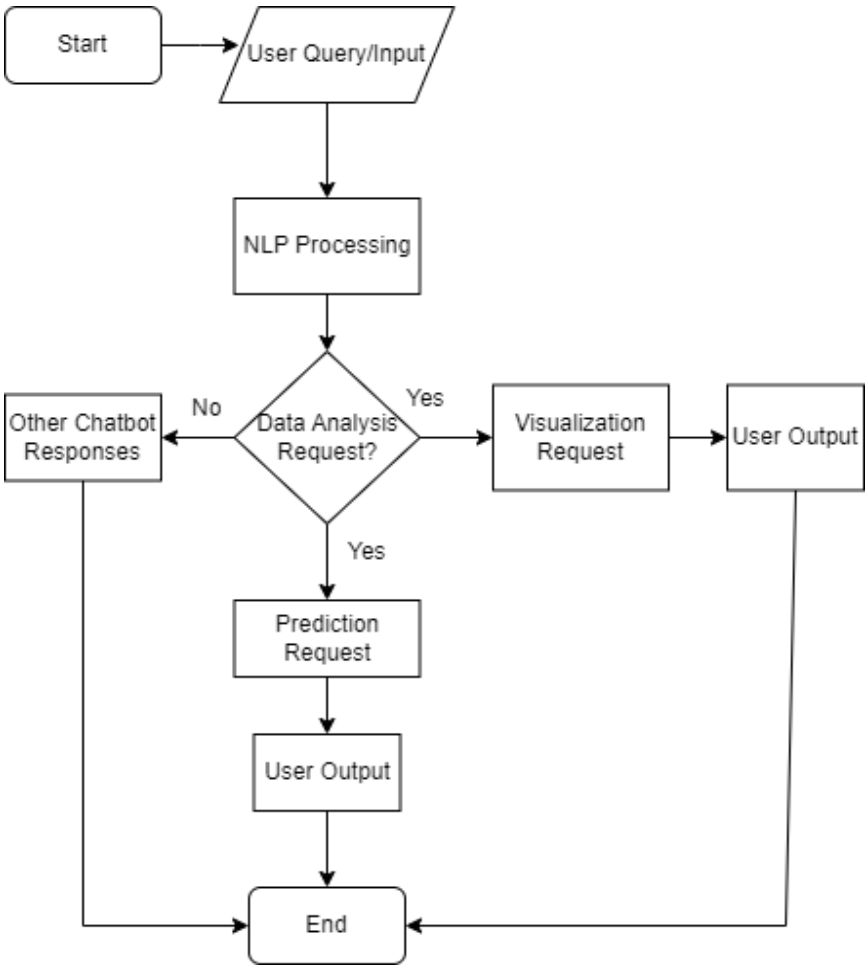


Fig. 2. Flowchart

Visualization Generation Module: Utilizes code generation techniques to create interactive visualizations tailored to user specifications, such as chart types, data attributes, and formatting preferences.

Infographics Generation Module (Work in Progress): Transforms data into stylized infographics, enhancing the presentation and interpretation of insights for diverse audiences.

C. Interaction Flow:

The interaction flow within the system follows a user-centric approach, where users initiate analysis tasks through the Streamlit interface, which in turn communicates with the LIDA backend for data processing and visualization generation. Users can upload datasets, input analysis parameters, and explore generated insights within the Streamlit environment, fostering an iterative and interactive data analysis experience.

VI. RESULTS AND ANALYSIS

User query: What is the distribution of age in the dataset? Histogram of Age

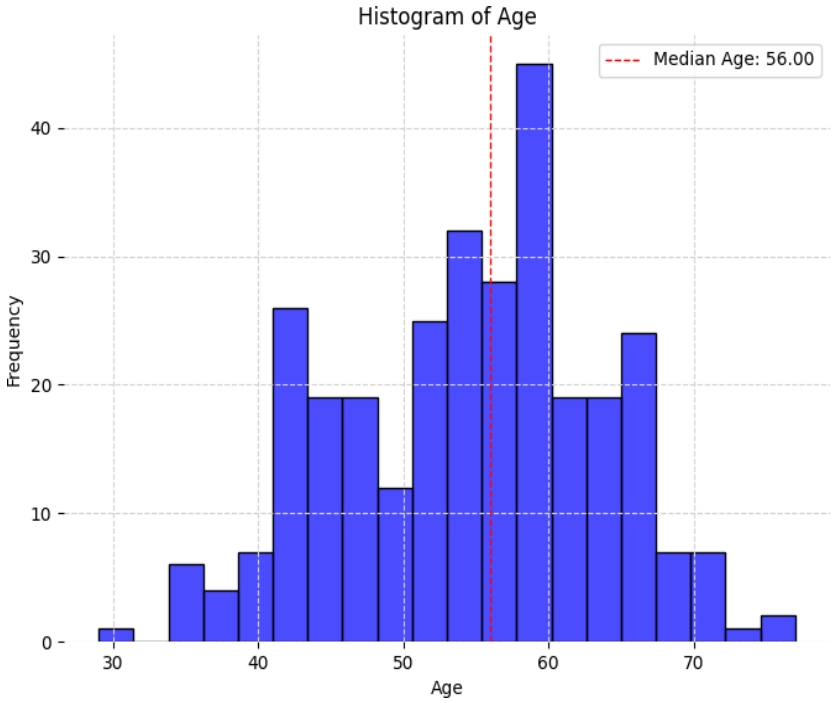
Explanation: The plot is a histogram of the 'Age' column in the provided data. The chart is created using the seaborn library with blue color and alpha value of 0.7. The chart also includes a vertical red dashed line representing the median age of the data.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

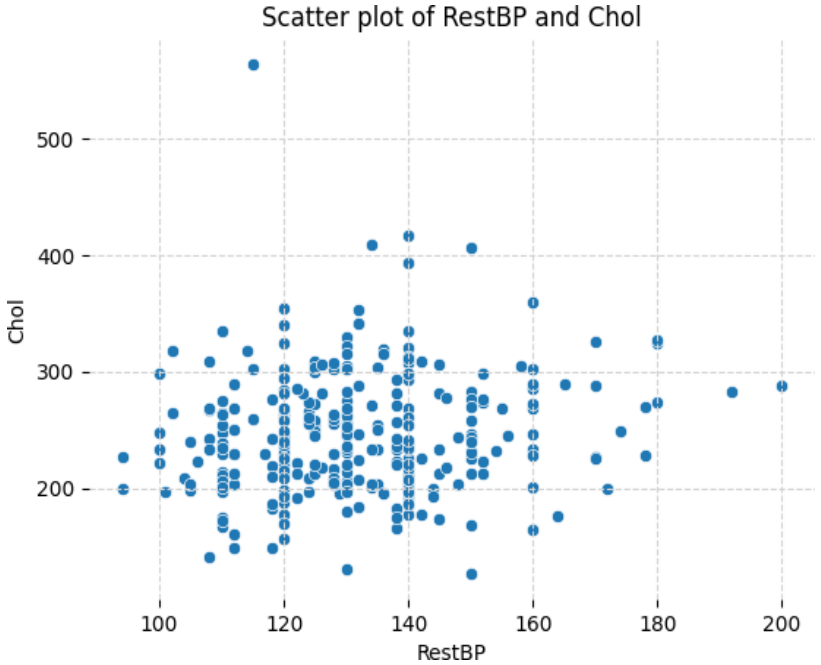
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13447



User query: What is the relationship between RestBP and Chol? Scatter plot of RestBP and Chol

Explanation: The chart created by the code is a scatter plot with RestBP on the x-axis and Chol on the y-axis. The plot uses the seaborn library and has default colors and markers. The goal of the chart is to visualize the relationship between RestBP and Chol. The main insight from the chart is that there is no clear linear relationship between RestBP and Chol.



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

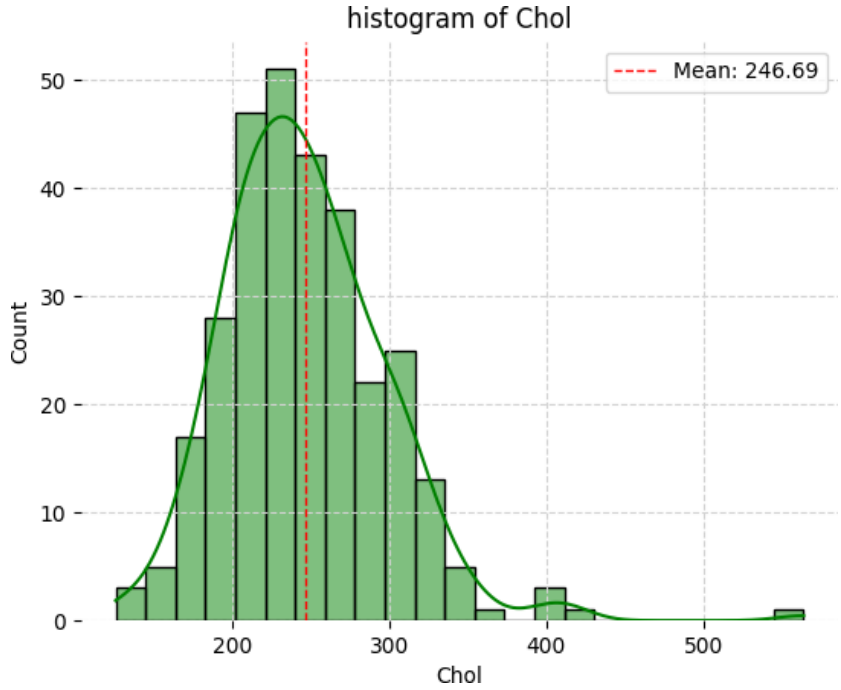
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13447

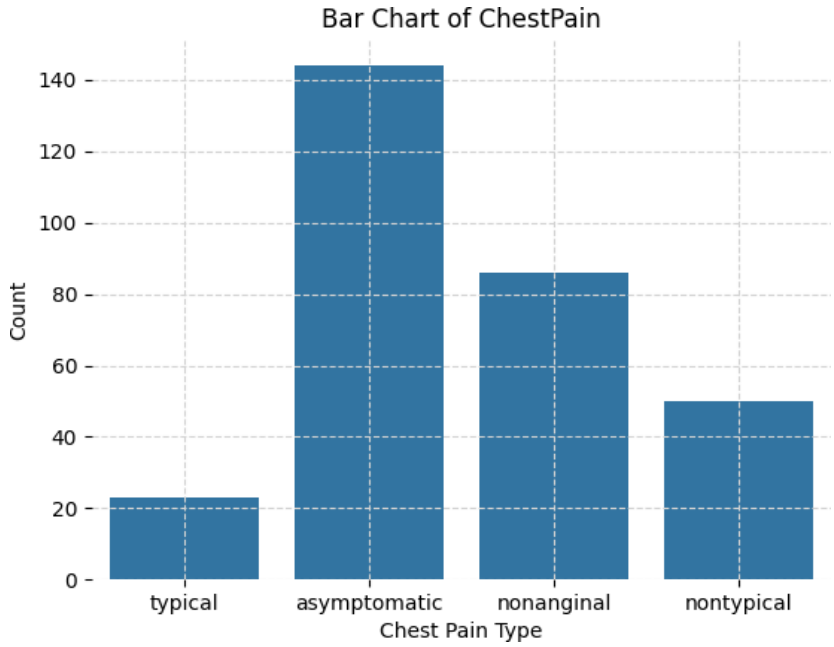
User Query: What is the distribution of cholesterol levels in the dataset? histogram of Chol

Explanation: The chart is a histogram with a green color scheme and a red dashed line indicating the mean value of the 'Chol' column. The goal of the chart is to visualize the distribution of cholesterol levels in the dataset. The main insight from the chart is the shape of the distribution and the location of the mean value.



User Query: What is the distribution of chest pain types in the dataset? bar chart of ChestPain

Explanation: The goal of this chart is to show the count of different types of chest pain. The chart is a bar chart with the x- axis representing the different types of chest pain and the y-axis representing the count. The chart has a title 'Bar Chart of ChestPain', x-axis label 'Chest Pain Type', and y-axis label 'Count'. The chart is created using the seaborn countplot function and then the title, x-axis label, and y-axis label are added using matplotlib.



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13447

VII. CONCLUSION

Our research paper has introduced an innovative approach to data analysis and visualization through the integration of Streamlit for user interaction and LIDA for automated insights generation. Through this endeavor, we aimed to democratize the field of data analysis by providing a user-friendly platform accessible to individuals with varying levels of technical expertise.

Our study has demonstrated the effectiveness of this approach in facilitating data-driven decision-making processes across different domains. By harnessing the power of natural language processing and machine learning, our platform enables users to upload datasets, engage in natural language conversations, extract insights, generate visualizations, and receive predictive analytics seamlessly.

REFERENCES

[1]. Akshay Bhor, Ujwala Sangale, Abhishek Sinha, Aniket Shewale, Prof. Abhay Gaidhani" Al-Driven Insights and Data Visualization ", IJARCCE International Journal of Advanced Research in Computer and Communication Engineering, 2023

[2]. Jash Doshi, “Chatbot User Interface for Customer Relationship Management using NLP models”,International Conference on Artificial Intelligence and Machine Vision (AIMV), 2024

[3]. Bhupesh Rawat, Ankur Singh Bist, “Recent Deep Learning Based NLP Techniques for Chatbot Development: An Exhaustive Survey”, International Conference on Cyber and IT Service Management (CITSM),2022

[4]. Jash Doshi, “Chatbot User Interface for Customer Relationship Management using NLP models”,International Conference on Artificial Intelligence and Machine Vision (AIMV), 2021

[5]. Berti-Equille Laure, Bonifati Angela, Milo Tova, “Machine Learning to Data Management: A Round Trip”, IEEE 34th International Conference on Data Engineering (ICDE), 2018

[6]. Jiafu Liu, Jiangtao Huang, Ying Xie, “Educational visualization application based on machine learning algorithm to predict student learning”, IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA),2021

[7]. Qianwen Wang, Zhutian Chen, Yong Wang, Huamin Qu, “A Survey on ML4VIS: Applying Machine Learning Advances to Data Visualization”, IEEE Transactions on Visualization and Computer Graphics

[8]. Huang Li, Shiaofen Fang, Snehasis Mukhopadhyay, Andrew J. Saykin, Li Shen, “Interactive Machine Learning by Visualization: A Small Data Solution”, IEEE International Conference on Big Data (Big Data),2018

[9]. Jinhua Chen, Qin Jiang, Yuxin Wang, Jing Tang, “Study of data analysis model based on big data technology”, IEEE International Conference on Big Data Analysis (ICBDA),2016

[10]. Fabian Nagel, Giuliano Castiglia, Gemza Ademaj, Juri Buchmller, Udo Schlegel, Daniel A. Keim “cpmViz: A Web- Based Visualization Tool for Uncertain Spatiotemporal Data”, IEEE Conference on Visual Analytics Science and Technology (VAST),2019

[11]. Hangu Yeo, Elahe Khorasani, Vadim Sheinin, Irene Manotas, Ngoc Phuoc An Vo, Octavian Popescu, “Natural Language Interface for Process Mining Queries in Healthcare”, IEEE International Conference on Big Data (Big Data), 2022

[12]. Mengfei Guo, Yufeng Chen, Jinan Xu, Yujie Zhang, “Dynamic Knowledge Integration for Natural Language Inference”, 4th International Conference on Natural Language Processing (ICNLP),2020