CSV File Analyzer

A3_10 OOPL Mini Project

Authors:

Somin Shah (EXCP)
somin.shah@somaiya.edu

Swajeet Salvi
swajeet.s@somaiya.edu

KJ Somaiya School of Engineering

November 5, 2024

***Abstract*—CSV File Analyzer allows for easy viewing and analysis of data**
Technology used: Java and Apache Netbeans
***Index Terms*—Java, JSwing, JFreeChart, Data Analysis, Regression**

I. Introduction

A lot of data analysts spend hours on writing multiple lines of code repeatedly to find the best form of representation for a particular piece of data. This is where the problem arrives, TIME. In order to reduce time and to help users understand their data better, we had come up with an idea to design an application that allows for easy viewing of the data. Something that allows the user to view the data in way they want with the click of a few buttons. Till now we have developed a prototype to such an extent that the user can import any CSV File and view the data with ease while also being able to perform Linear or Logistic regression on that data.

II. Brief introduction on Regression concepts used

A. Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It estimates the best-fitting line through data points, allowing predictions and insights about trends, correlations, and the strength of relationships in the data.

The formula for Linear Regression can be written as:

$$y = intercept + slope \times x \tag{1}$$

$$slope = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \tag{2}$$

$$intercept = \bar{y} - \beta_1 \bar{x} \tag{3}$$

B. Logistic Regression

Logistic regression is a statistical method for modeling binary outcomes based on one or more independent variables. It estimates the probability that a given input belongs to a particular category, using a logistic function to map predicted values to a range between 0 and 1, facilitating classification tasks.

The formula for Logistic Regression can be written as:

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{4}$$

III. System Architecture

In order for a seamless user experience and rich productivity, the architecture of this system has to be split into 3 major parts.

A. Viewing the data

The ability to view the data plays a huge role in the analysis of it. It is important that when accessing data, regardless of it's size and complexity it should be easily read and understood by the user and that there should be no missing data. This was possible using:

- JFileChooser: This allows for the data to be easily imported from any part of your system allowing for seamless access
- JTable: This allows for that data to be shown in a tabular form with all the columns and rows.

B. Plotting the data

Just showing the data in tabular form does not help in the analysis of it. It also helps to plot that data which then helps in planning how to analyze it further. This was possible using:

- JFreeChart: This allowed for us to easily plot a scatter graph and view the data

C. Data Analysis

Data analysis is the last layer for the system which currently includes performing linear and logistic regression on the data. This was done using manually calculating the variables and hard-coding the regression models and plotting them into the graph.

D. System Architecture Flowchart

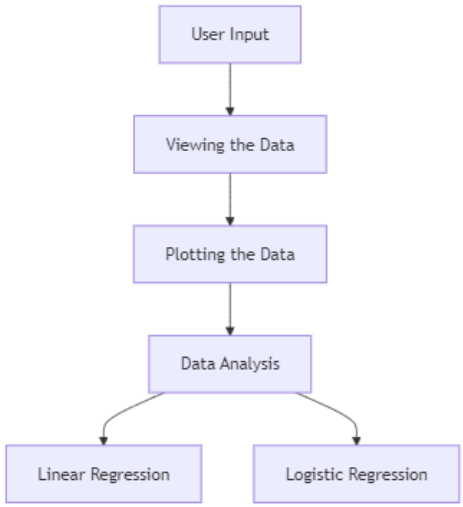

Fig. 1. System Architecture Flowchart

IV. Technology Stack

A. Language Used (Java)

Java is a versatile programming language used for developing applications across platforms. For GUI development, it offers frameworks like Swing for lightweight interfaces and JavaFX for modern, rich applications. Java's extensive libraries enable developers to create interactive desktop and web applications efficiently.

B. Software used (Apache Netbeans)

Apache NetBeans is an open-source integrated development environment (IDE) for Java and other languages. It simplifies GUI development through its drag-and-drop interface builder, particularly for Swing and JavaFX applications. NetBeans provides tools for designing, coding, and debugging, enhancing productivity for developers creating desktop applications.

C. Libraries

1) JavaX: JavaX is a set of extensions to the Java platform, providing additional libraries and tools for advanced application development. JSwing is a GUI toolkit within JavaX that allows developers to create rich, interactive desktop applications using customizable components. Together, they enhance Java's capabilities for building user-friendly interfaces.

2) JFreeChart: JFreeChart is an open-source Java library for creating a wide variety of charts and graphs. It supports various chart types, including line, bar, pie, and scatter plots, making it easy to visualize data in applications. With its flexible design, JFreeChart allows developers to customize chart appearance and behavior, integrating seamlessly into Java Swing and JavaFX applications, thereby enhancing data representation and analysis in software projects.

V. Design and Implementation

A. Blueprint

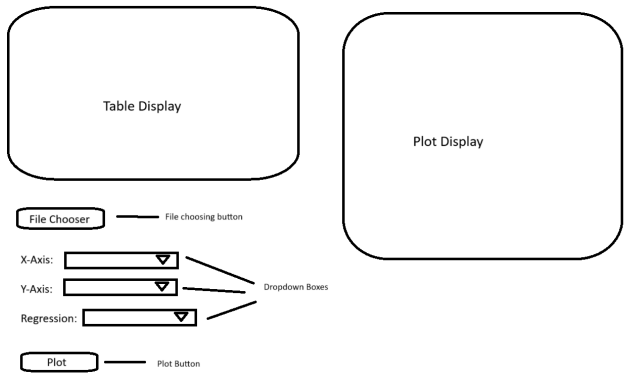

Fig. 2. System Blueprint

B. Module breakdown

1) UI Design(mainpage.form): this page allowed us to use the drag and drop features of Apache Netbeans that allowed us to seamlessly create the GUI of the application.

2) UX Experience(mainpage.java): This page had all the back-end. The functions that are to be executed when a button is pressed is written here. This page also consists of all the hard-coded regression models that will be executed when needed.

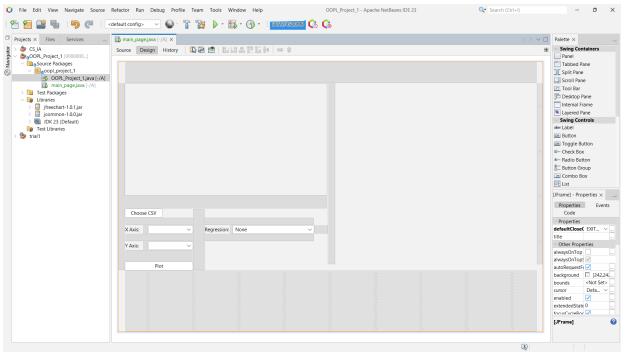

Fig. 3. mainpage.form

3) Graph Plotting: This was done using JFreeChart that consists of all the functions that help in plotting a scatter graph or a line graph.

C. Key Feature: Data Analysis

Both Linear regression and Logistic regression models were manually coded where all functions and variables were manually calculated.

D. Challenges

- Plotting the graph: There were multiple issues in plotting the graph as sometimes the data was not full converted to an array or some other technical issue.
- Data analysis: Since the functions for the regression model were manually written, there were some issues in making sure whether the model were correct.

VI. Testing

A. Test Cases

1) File Choosing: It was important to make sure that JFileChooser was working properly and was able to import a CSV file correctly.
2) Plotting: After importing the CSV file, it is important to ensure that a correct and accurate graph is plotted and also based on the columns chosen by the user.
3) Multiple plotting: It was also important to make sure that a different plot can be generated without having to re-run the project, whether it is selecting a new column or a completely new CSV File.
4) Data Analysis: It was important to make sure that both the model were working properly and that there graphs are also shown properly.

B. Result and Screenshots

Using these photos, it is understood that all current test cases have been passed. Graph plotting plays a huge role in this application and have been met as seen by these screenshots.

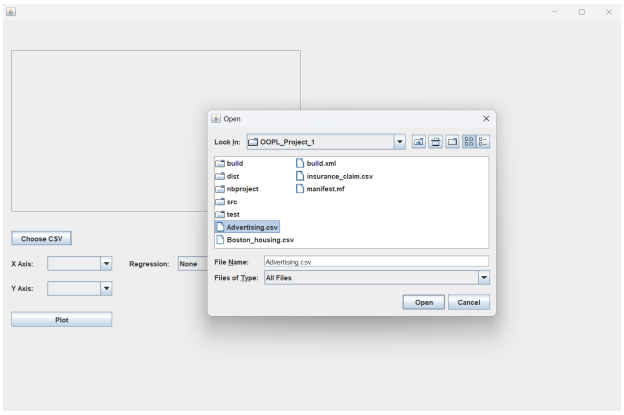

Fig. 4. Test 1 part 1

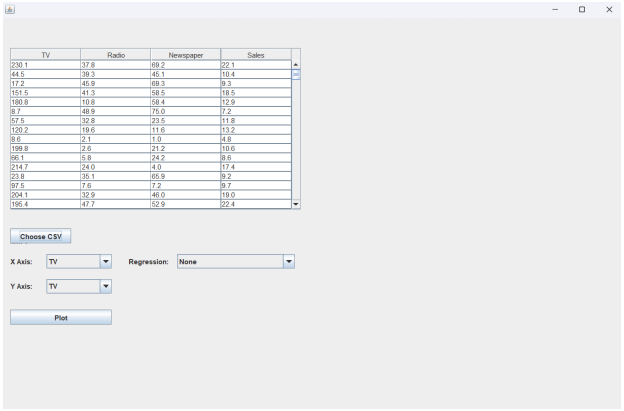

Fig. 5. Test 1 part 2

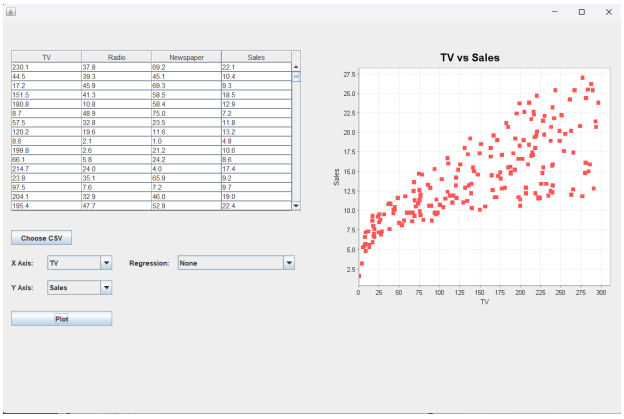

Fig. 6. Test 2

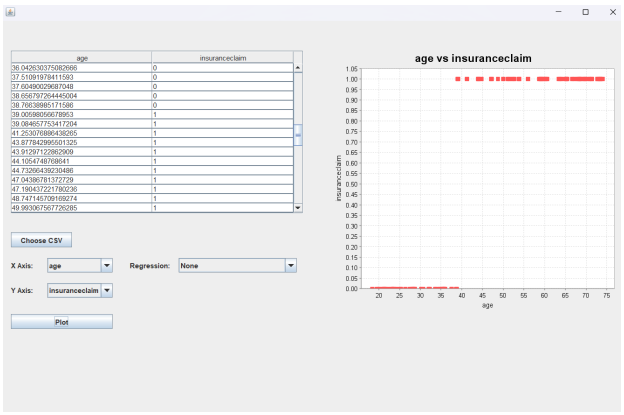

Fig. 7. Test 3

Fig. 8. Test 4 part 1

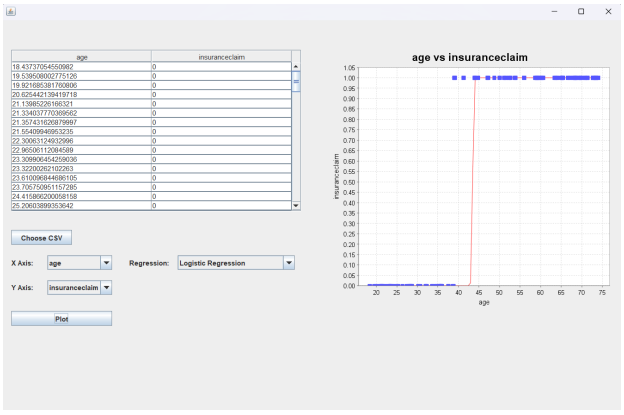

Fig. 9. Test 4 part 2

VII. Conclusion and Future work

A. Summary

This prototype is able to meet the expectations and standards that were set before developing it. It is now able to be a productive tool for users allowing them to efficiently analyze data. It can prove to be an incredible time saver, preventing repetitive programming. Linear regression and logistic regression are two of the most common models used by Data analysts, especially beginners. This allows them to learn better and apply their skills at a more efficient scale.

B. Future work

There are limitations to this application. Such as only being able to import a CSV file, or only being to analyze strictly numerical data. There are also only two regression models being used out of the numerous. It is evident that this application can still be developed at a larger scale, where there can be the use of more Analysis methods, prediction models, and also the ability to extract and analyze different types of data from different types of files.

References

[1] JFreeChart. (n.d.). https://www.jfree.org/jfreechart/

[2] GeeksforGeeks. (2024, September 9). Linear regression: Definition, formula derivation and examples. https://www.geeksforgeeks.org/linear-regression-formula/

[3] GeeksforGeeks. (2024, June 20). Logistic regression in machine learning. GeeksforGeeks. https://www.geeksforgeeks.org/understanding-logistic-regression/

[4] GeeksforGeeks. (2024a, July 30). Introduction to java swing. https://www.geeksforgeeks.org/introduction-to-java-swing/