Data Analysis Application - ”The Edgers”

Team The Edgers

13/05/2024

1 Introduction

For our Software Technology course, we developed a web-based data analysis application. It includes tools for data loading, 2D visualizations, data exploration, and comparing classification and clustering algorithms.

2 Development

We built the app using R and the Shiny framework. Key packages include `shiny`, `readr`, `ggplot2`, `plotly`, `caret`, `randomForest`, `cluster`, and `Rtsne`.

3 Functionality

Here’s what our app can do:

3.1 Data Loading

Users can load CSV or Excel files using `fileInput`, `read_csv`, and `read_excel`.

3.2 2D Visualizations

We support PCA and t-SNE for 2D data projections. Users choose the method with `radioButtons` and view interactive plots with `plotly`.

3.3 Exploratory Data Analysis (EDA)

Users can explore data with histograms and boxplots. They select the chart type with `radioButtons`, and `ggplot2` generates the plots for each data column.

3.4 Comparing Machine Learning Algorithms

Our app compares classification and clustering algorithms.

3.4.1 Classification Algorithm Comparison

Compare Logistic Regression and Random Forest classifiers. Users can adjust parameters like the C value for Logistic Regression and the number of trees for Random Forest.

3.4.2 Clustering Algorithm Comparison

Compare K-Means and Hierarchical Clustering. Users specify the number of clusters for each algorithm.

4 Team Collaboration

Our team, The Edgers, worked on this project together in person, collaborating on all aspects of the development. We also used a GitHub repository to manage our code and documents. You can check out our GitHub repo here: `https://github.com/The-Edgers/RShiny`.

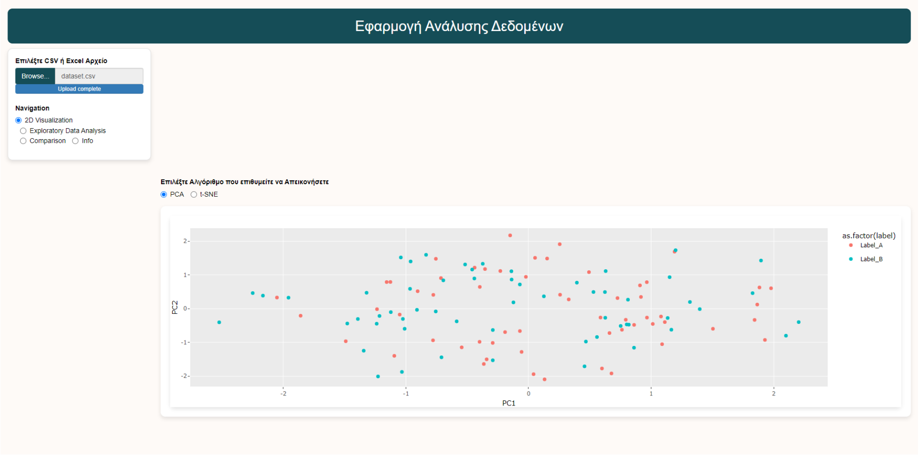
5 Using the App

5.1

Using our application is straightforward. Users start by uploading their data in CSV or Excel format. Once the data is loaded, they can navigate through the various tabs to perform different analyses.

5.2

For 2D visualizations, users select the desired projection method (PCA or t-SNE) and view the resulting plots interactively. We've attatcherd some screenshots below to showcase the results of a random dataset we made.



Εφαρμογή Ανάλυσης Δεδομένων

Επιλέξτε CSV ή Excel Αρχείο

Browse... dataset.csv

Upload complete

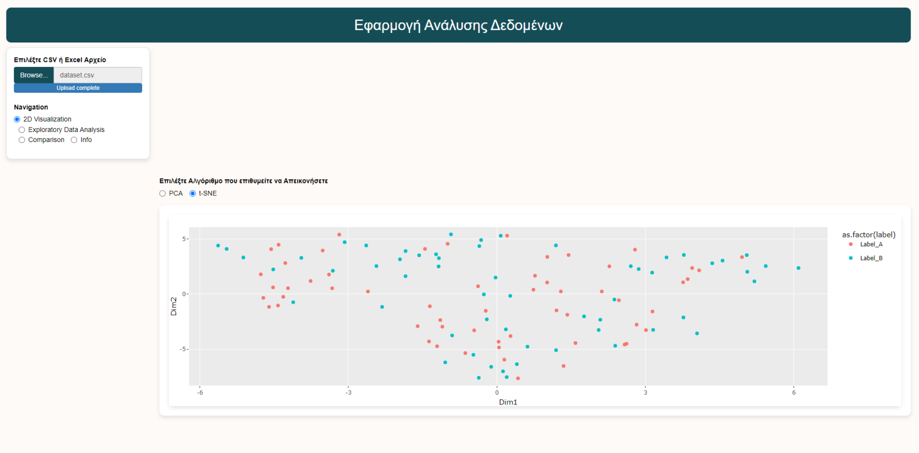
Navigation

● 2D Visualization
○ Exploratory Data Analysis
○ Comparison ○ Info

Επιλέξτε Αλγόριθμο που επιθυμείτε να Απεικονήσετε

● PCA ○ t-SNE

Figure 1: PCA



Εφαρμογή Ανάλυσης Δεδομένων

Επιλέξτε CSV ή Excel Αρχείο

Browse... dataset.csv

Upload complete

Navigation

● 2D Visualization
○ Exploratory Data Analysis
○ Comparison ○ Info

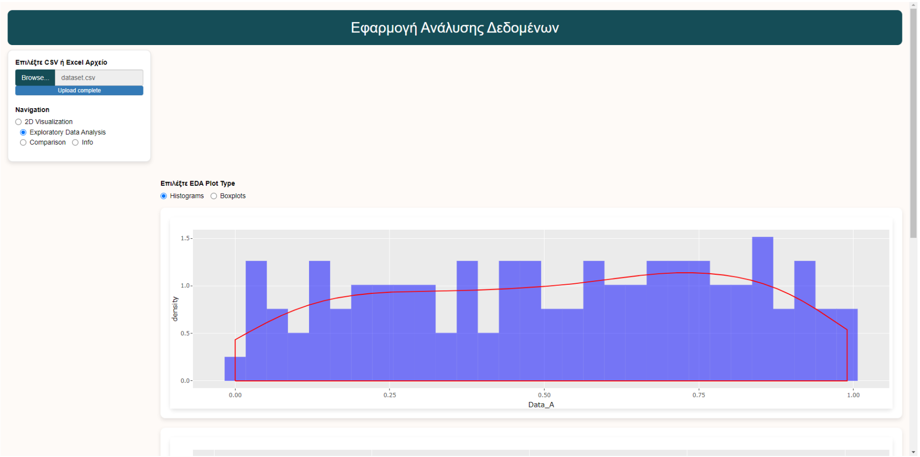
Επιλέξτε Αλγόριθμο που επιθυμείτε να Απεικονήσετε

○ PCA ● t-SNE

Figure 2: T-SNE

5.3

The EDA section allows users to choose between histograms and boxplots to explore the distributions and characteristics of their data columns.



Εφαρμογή Ανάλυσης Δεδομένων

Επιλέξτε CSV ή Excel Αρχείο

Browse... dataset.csv

Upload complete

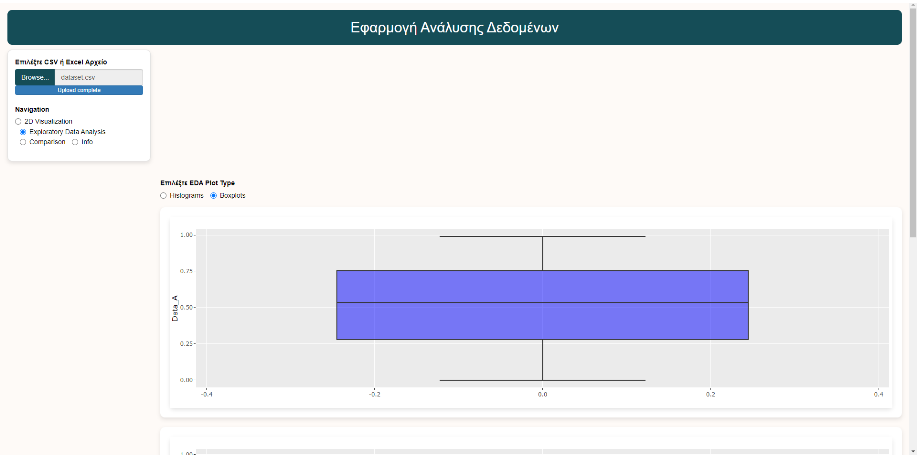
Navigation

○ 2D Visualization
● Exploratory Data Analysis
○ Comparison ○ Info

Επιλέξτε EDA Plot Type

● Histograms ○ Boxplots

Figure 3: Histograms



Εφαρμογή Ανάλυσης Δεδομένων

Επιλέξτε CSV ή Excel Αρχείο

Browse... dataset.csv

Upload complete

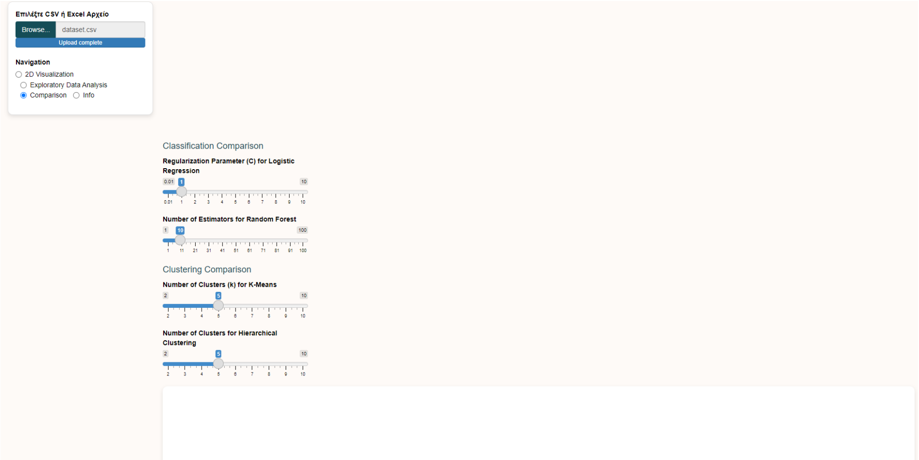
Navigation

○ 2D Visualization
● Exploratory Data Analysis
○ Comparison ○ Info

Επιλέξτε EDA Plot Type

○ Histograms ● Boxplots

Figure 4: Boxplots



Classification Comparison

Regularization Parameter (C) for Logistic Regression: 1

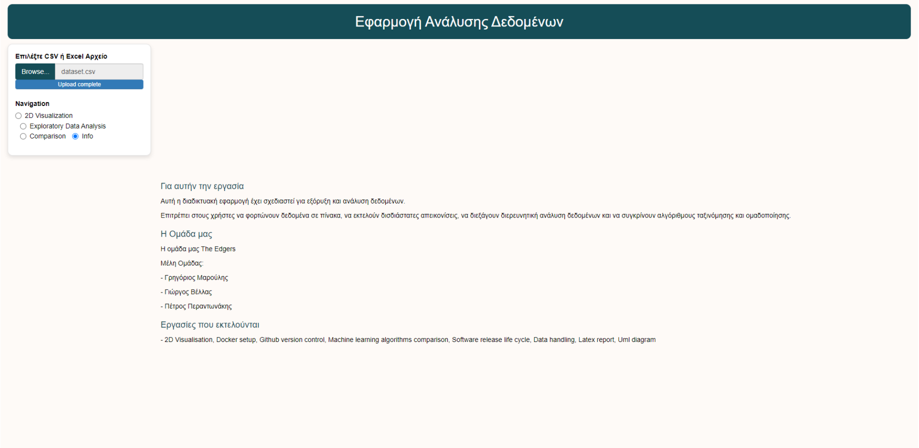
Number of Estimators for Random Forest: 10

Clustering Comparison

Number of Clusters (k) for K-Means: 5

Number of Clusters for Hierarchical Clustering: 5

Figure 5: Comparison


Εφαρμογή Ανάλυσης Δεδομένων
Επιλέξτε CSV ή Excel Αρχείο
Browse...
dataset.csv
Upload complete
Navigation
○ 2D Visualization
○ Exploratory Data Analysis
○ Comparison
● Info

Για αυτήν την εργασία

Αυτή η διαδικτυακή εφαρμογή έχει σχεδιαστεί για εξόρυξη και ανάλυση δεδομένων.

Επιτρέπει στους χρήστες να φορτώνουν δεδομένα σε πίνακα, να εκτελούν δισδιάστατες απεικονίσεις, να διεξάγουν διερευνητική ανάλυση δεδομένων και να συγκρίνουν αλγόριθμους ταξινόμησης και ομαδοποίησης.

Η Ομάδα μας

Η ομάδα μας The Edgers

Μέλη Ομάδας:

- Γρηγόριος Μαρούλης
- Γιώργος Βέλλας
- Πέτρος Περαντωνάκης

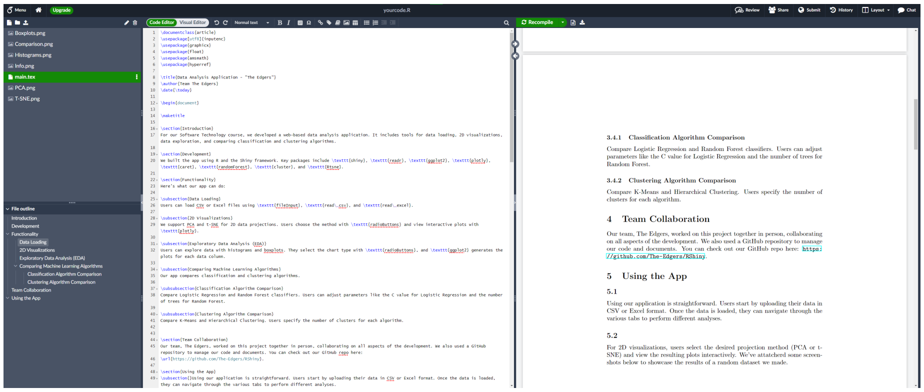
Εργασίες που εκτελούνται

- 2D Visualisation, Docker setup, Github version control, Machine learning algorithms comparison, Software release life cycle, Data handling, Latex report, Uml diagram

Figure 6: Info

6 Latex

We used overleaf to create our final report as per the instructions along with latex. the latex file is available in a separate repository in the organization's GitHub: `https://github.com/The-Edgers/Overleaf_Latex_Anafora`



```latex
\documentclass{article}
\usepackage[utf8]{inputenc}
\usepackage{graphicx}
\usepackage{float}
\usepackage{amsmath}
\usepackage{hyperref}

\title{Data Analysis Application - "The Edgers"}
\author{Team The Edgers}
\date{\today}

\begin{document}

\maketitle

\section{Introduction}
For our Software Technology course, we developed a web-based data analysis application. It includes tools for data loading, 2D visualizations, data exploration, and comparing classification and clustering algorithms.

\section{Development}
We built the app using R and the Shiny framework. Key packages include \texttt{shiny}, \texttt{readr}, \texttt{ggplot2}, \texttt{plotly}, \texttt{caret}, \texttt{randomForest}, \texttt{cluster}, and \texttt{Rtsne}.

\section{Functionality}
Here's what our app can do:

\subsection{Data Loading}
Users can load CSV or Excel files using \texttt{fileInput}, \texttt{read\_csv}, and \texttt{read\_excel}.

\subsection{2D Visualizations}
We support PCA and t-SNE for 2D data projections. Users choose the method with \texttt{radioButtons} and view interactive plots with \texttt{plotly}.

\subsection{Exploratory Data Analysis (EDA)}
Users can explore data with histograms and boxplots. They select the chart type with \texttt{radioButtons}, and \texttt{ggplot2} generates the plots for each data column.

\subsection{Comparing Machine Learning Algorithms}
Our app compares classification and clustering algorithms.

\subsubsection{Classification Algorithm Comparison}
Compare Logistic Regression and Random Forest classifiers. Users can adjust parameters like the C value for Logistic Regression and the number of trees for Random Forest.

\subsubsection{Clustering Algorithm Comparison}
Compare K-Means and Hierarchical Clustering. Users specify the number of clusters for each algorithm.


\section{Team Collaboration}
Our team, The Edgers, worked on this project together in person, collaborating on all aspects of the development. We also used a GitHub repository to manage our code and documents. You can check out our GitHub repo here:
\url{https://github.com/The-Edgers/RShiny}.

\section{Using the App}
\subsection{}Using our application is straightforward. Users start by uploading their data in CSV or Excel format. Once the data is loaded, they can navigate through the various tabs to perform different analyses.
```

3.4.1 Classification Algorithm Comparison

Compare Logistic Regression and Random Forest classifiers. Users can adjust parameters like the C value for Logistic Regression and the number of trees for Random Forest.

3.4.2 Clustering Algorithm Comparison

Compare K-Means and Hierarchical Clustering. Users specify the number of clusters for each algorithm.

4 Team Collaboration

Our team, The Edgers, worked on this project together in person, collaborating on all aspects of the development. We also used a GitHub repository to manage our code and documents. You can check out our GitHub repo here: https://github.com/The-Edgers/RShiny.

5 Using the App

5.1

Using our application is straightforward. Users start by uploading their data in CSV or Excel format. Once the data is loaded, they can navigate through the various tabs to perform different analyses.

5.2

For 2D visualizations, users select the desired projection method (PCA or t-SNE) and view the resulting plots interactively. We've attatcherd some screenshots below to showcase the results of a random dataset we made.

Figure 7: Latex

7 Uml Diagram

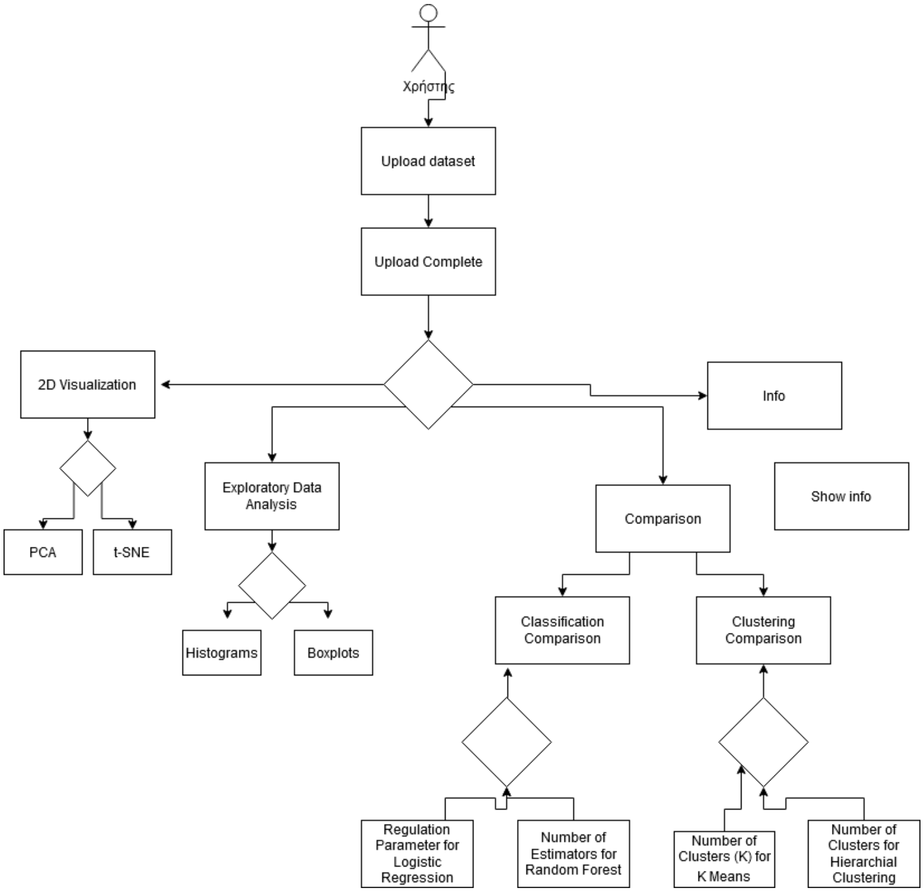

Χρήστης
Upload dataset
Upload Complete
2D Visualization
Info
Exploratory Data Analysis
Comparison
Show info
PCA
t-SNE
Histograms
Boxplots
Classification Comparison
Clustering Comparison
Regulation Parameter for Logistic Regression
Number of Estimators for Random Forest
Number of Clusters (K) for K Means
Number of Clusters for Hierarchial Clustering

Figure 8: Diagram

8 Conclusion

The results of the comparisons between the algorithms can be found in the screenshots we've provided above. In conclusion, the app can be used as a very powerful tool for data analysis because of its accurate and productive visualization produced with the charts. Our application successfully demonstrates the capabilities of R and Shiny in creating interactive and user-friendly data analysis tools.

9 Life Cycle Model

We have decided to use Agile for our Life Cycle model because it will allow us to adapt to changes quickly and iteratively improve the application based on user feedback.

9.1 Life Cycle Stages

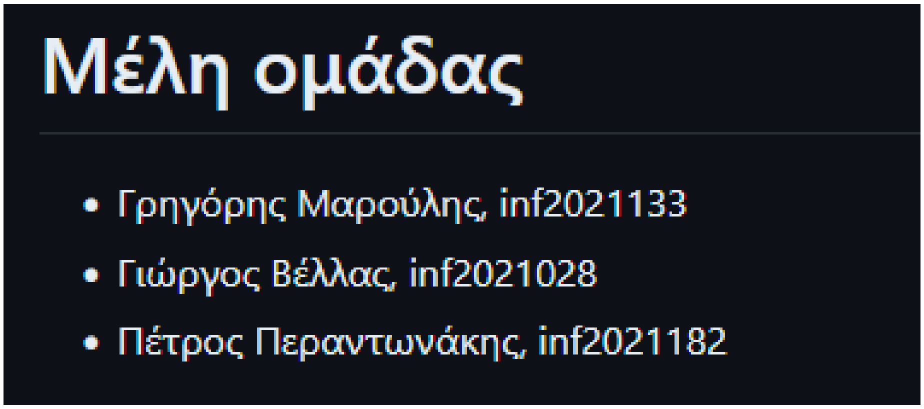
Identify key stakeholders

Create a backlog of features and prioritize features based on importance and user needs.

Implement the user interface (ui) and server logic (server) iteratively, start with basic features and gradually add more complex functionalities, continuously test and integrate new code.

Conduct integration tests to ensure components work together seamlessly.

Monitor the application's performance and address any issues promptly and gather continuous feedback and iterate on new features.



Μέλη ομάδας

- Γρηγόρης Μαρούλης, inf2021133
- Γιώργος Βέλλας, inf2021028
- Πέτρος Περαντωνάκης, inf2021182

Figure 9: names