

Data Analysis Application Report

May 31, 2024

1 Introduction

This report documents the functionalities of the Data Analysis Application, designed for data mining and analysis. The application allows users to upload data, perform two-dimensional visualizations, conduct exploratory data analysis, and compare classification and clustering algorithms.

2 Data Upload

Users can upload CSV or Excel files. The data is then displayed in a table for initial inspection.

3 Two-Dimensional Visualization

The application provides two methods for 2D visualization:

- Principal Component Analysis (PCA)
- t-Distributed Stochastic Neighbor Embedding (t-SNE)

These methods reduce the dimensionality of the data to two dimensions for visualization purposes.

4 Exploratory Data Analysis (EDA)

Users can perform EDA using:

- Histograms
- Boxplots

These plots help in understanding the distribution and variance of numerical features in the dataset.

5 Classification Comparison

The application compares the performance of two classification algorithms:

- Logistic Regression
- Random Forest

Users can adjust the regularization parameter for Logistic Regression and the number of estimators for Random Forest. The accuracy of each classifier is displayed after running the comparison.

6 Clustering Comparison

The application supports clustering with:

- K-Means
- Hierarchical Clustering

Users can specify the number of clusters for each algorithm. The cluster labels are displayed for both methods after running the comparison.

7 Implementation Details

The application is implemented using:

- **Streamlit** for the web interface
- **Pandas** for data manipulation
- **Plotly** for visualizations
- **Scikit-learn** for machine learning algorithms

8 Team

Our team of two worked closely together to create the app and the code. We have uploaded the files to github so that there is access for the app to work. <https://github.com/TEKNOLOGIA-LOGISMIKOY>

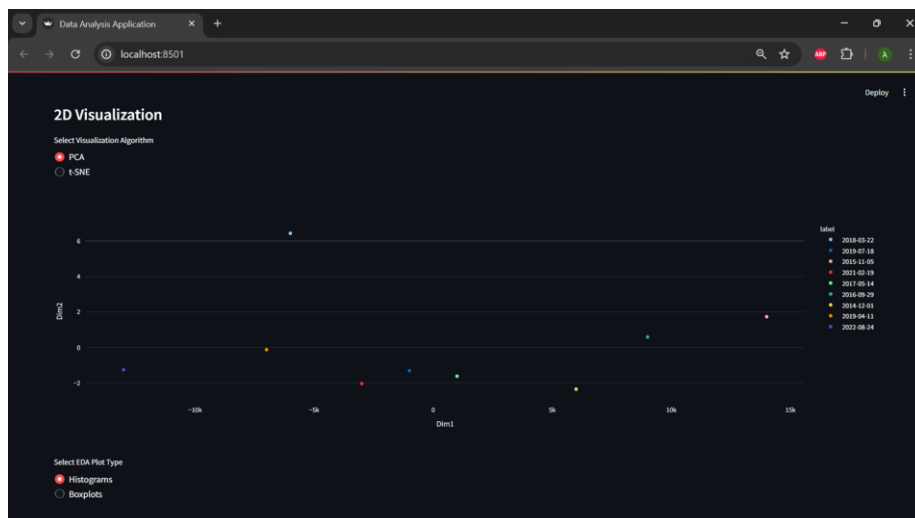
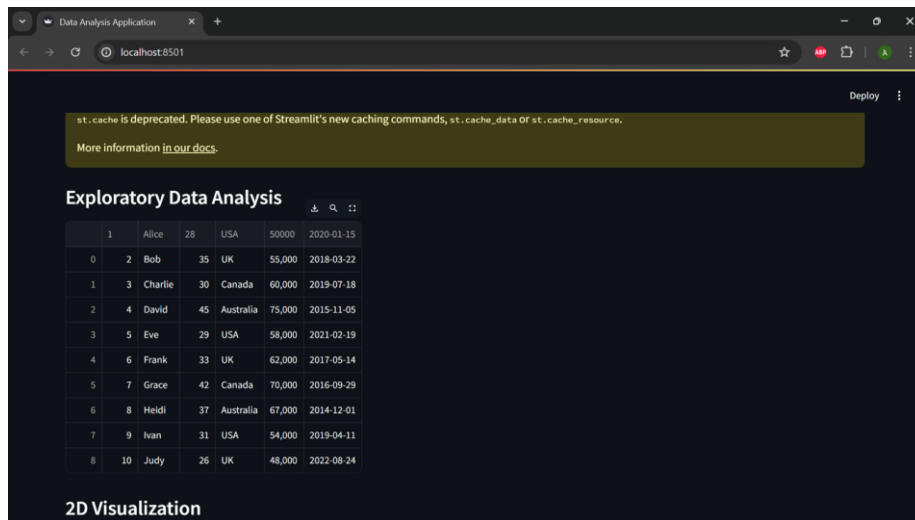
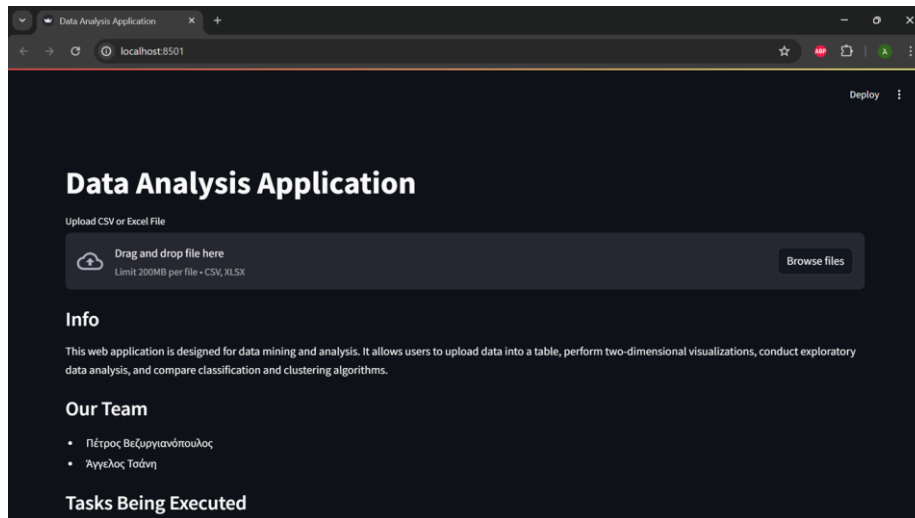
We used overleaf for our report with latex. The file is in a repository in github: <https://github.com/TEKNOLOGIA-LOGISMIKOY/Latex-anafora>

9 Tasks Executed

- 2D Visualization
- Docker setup
- Github version control
- Machine learning algorithms comparison
- Software release life cycle

10 Conclusion

This application provides a comprehensive tool for data analysis, enabling users to visualize, analyze, and compare data using various methods. The integration of different machine learning algorithms and visualization techniques offers a robust platform for exploratory data analysis and model evaluation.





The figure shows a "Comparison" interface in a web application. It features five horizontal sliders for configuring machine learning parameters:

- Regularization Parameter (C) for Logistic Regression:** Range 0.01 to 10.00, with a red dot at 1.00.
- Number of Estimators for Random Forest:** Range 1 to 100, with a red dot at 10.
- Number of Clusters (k) for K-Means:** Range 2 to 10, with a red dot at 5.
- Number of Clusters for Hierarchical Clustering:** Range 2 to 10, with a red dot at 5.

Below the sliders are two buttons: "Run Classification Comparison" and "Run Clustering Comparison".

Info

This web application is designed for data mining and analysis. It allows users to upload data into a table, perform two-dimensional visualizations, conduct exploratory data analysis, and compare classification and clustering algorithms.

Our Team

Info

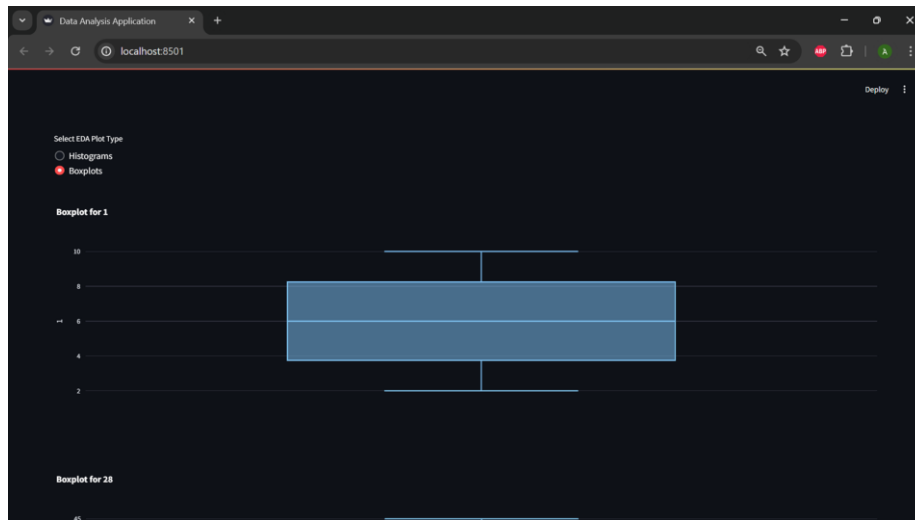
This web application is designed for data mining and analysis. It allows users to upload data into a table, perform two-dimensional visualizations, conduct exploratory data analysis, and compare classification and clustering algorithms.

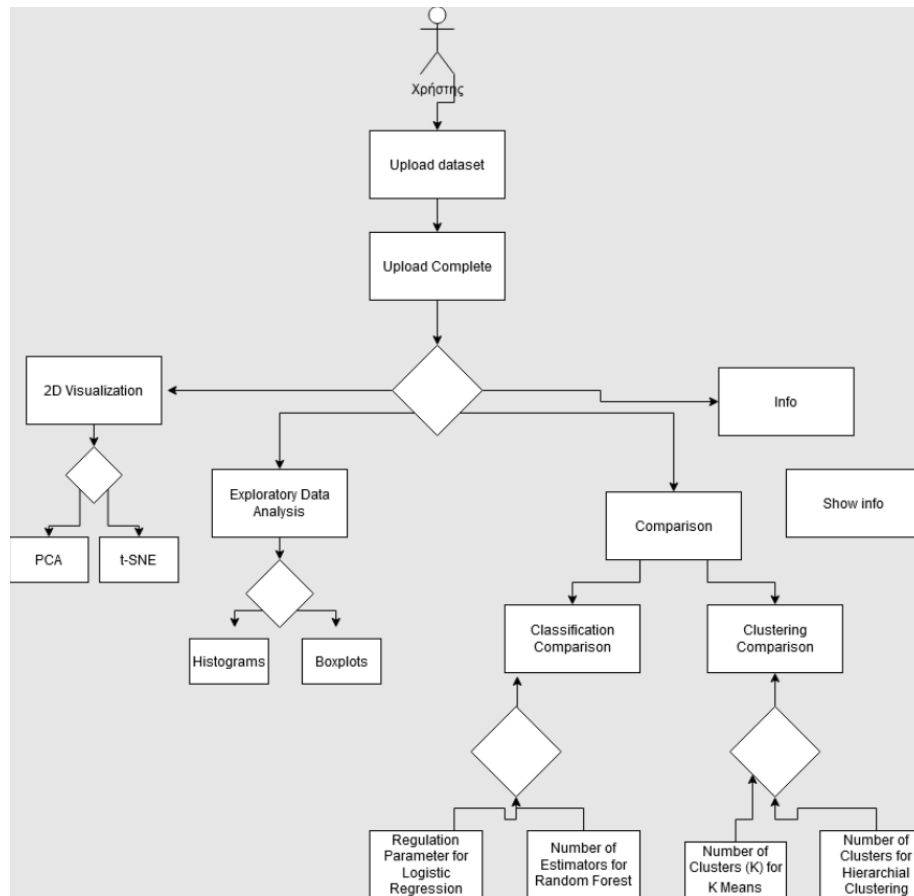
Our Team

- Πέτρος Θεοφανίδης
- Άγγελος Τσίμης

Tasks Being Executed

- 2D Visualization, Docker setup, Github version control, Machine learning algorithms comparison, Software release life cycle





ΜΕΛΗ ΟΜΑΔΑΣ

1. ΠΕΤΡΟΣ

ΒΕΖΥΡΓΙΑΝΟΠΟΥΛΟΣ(inf2021026)

2. ΑΓΓΕΛΟΣ ΤΣΑΝΗ(inf2021231)