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/TEXNOLOGIA-LOGISMIKOY

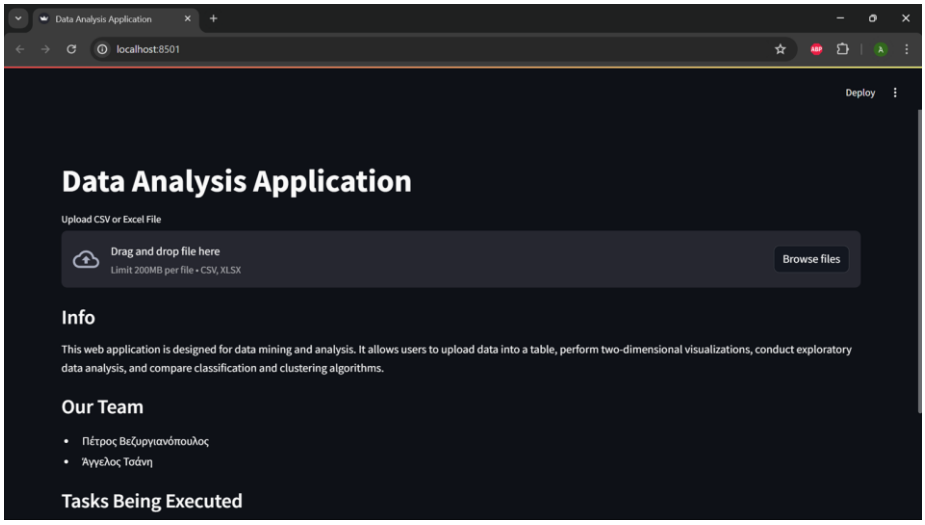
We used overleaf for our report with latex. The file is in a repository in github: https://github.com/TEXNOLOGIA-LOGISMIKOY/Latex-anafora

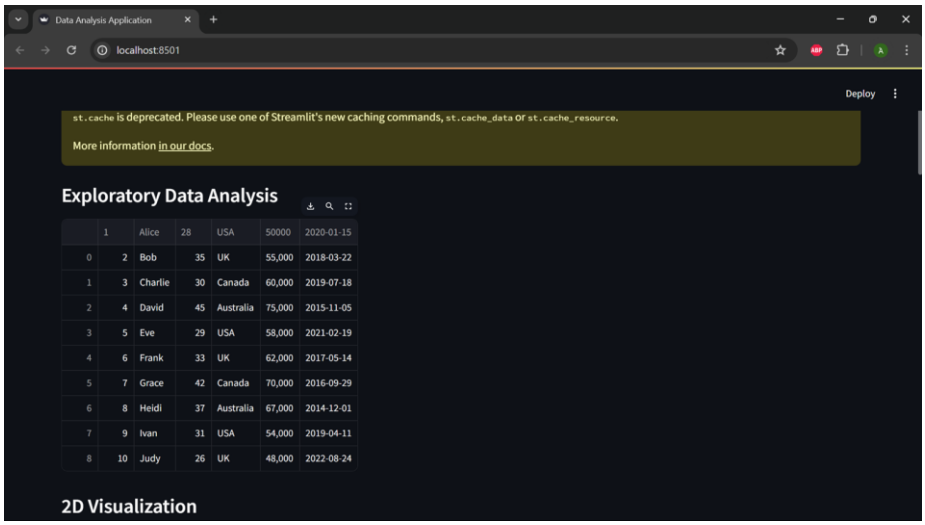
9 Tasks Executed

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

10 Conclusion

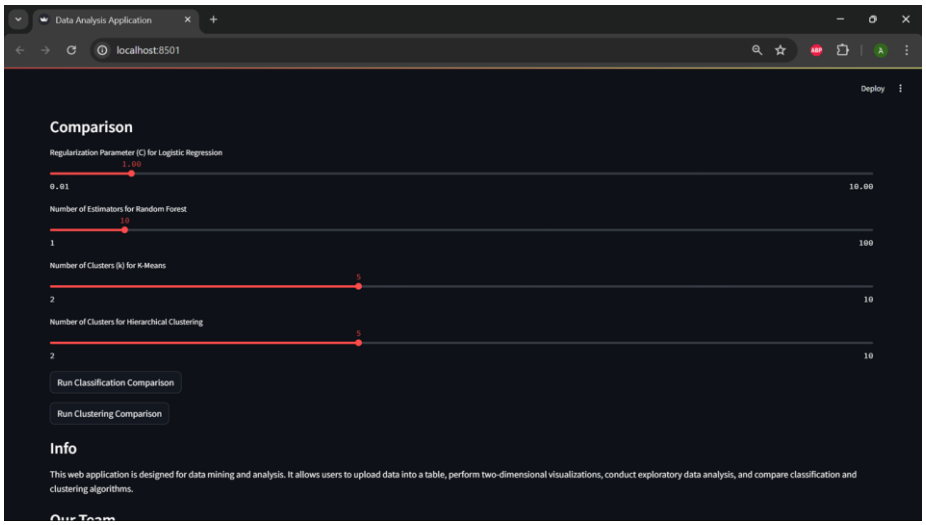
This application provides a comprehensive tool for data analysis, enabling users to visualize, analyze, and compare data using various methods. The integration of different machine learning algorithms and visualization techniques offers a robust platform for exploratory data analysis and model evaluation.

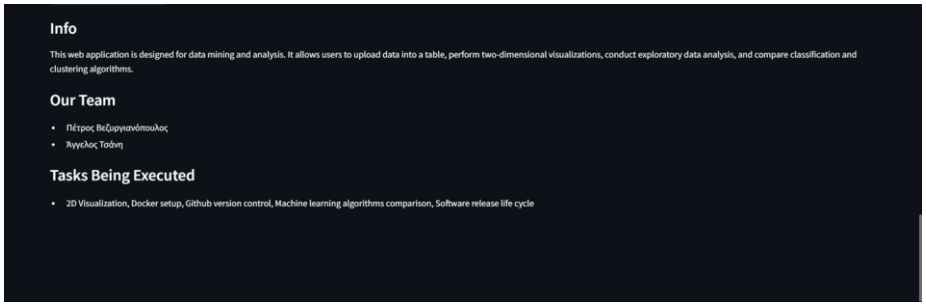

Data Analysis Application
localhost:8501
Deploy
Data Analysis Application
Upload CSV or Excel File
Drag and drop file here
Limit 200MB per file • CSV, XLSX
Browse files
Info
This web application is designed for data mining and analysis. It allows users to upload data into a table, perform two-dimensional visualizations, conduct exploratory data analysis, and compare classification and clustering algorithms.
Our Team
Πέτρος Βεζυργιανόπουλος
Άγγελος Τσάνη
Tasks Being Executed


Data Analysis Application
localhost:8501
Deploy
st.cache is deprecated. Please use one of Streamlit's new caching commands, st.cache_data or st.cache_resource.
More information in our docs.
Exploratory Data Analysis
1 Alice 28 USA 50000 2020-01-15
0 2 Bob 35 UK 55,000 2018-03-22
1 3 Charlie 30 Canada 60,000 2019-07-18
2 4 David 45 Australia 75,000 2015-11-05
3 5 Eve 29 USA 58,000 2021-02-19
4 6 Frank 33 UK 62,000 2017-05-14
5 7 Grace 42 Canada 70,000 2016-09-29
6 8 Heidi 37 Australia 67,000 2014-12-01
7 9 Ivan 31 USA 54,000 2019-04-11
8 10 Judy 26 UK 48,000 2022-08-24
2D Visualization

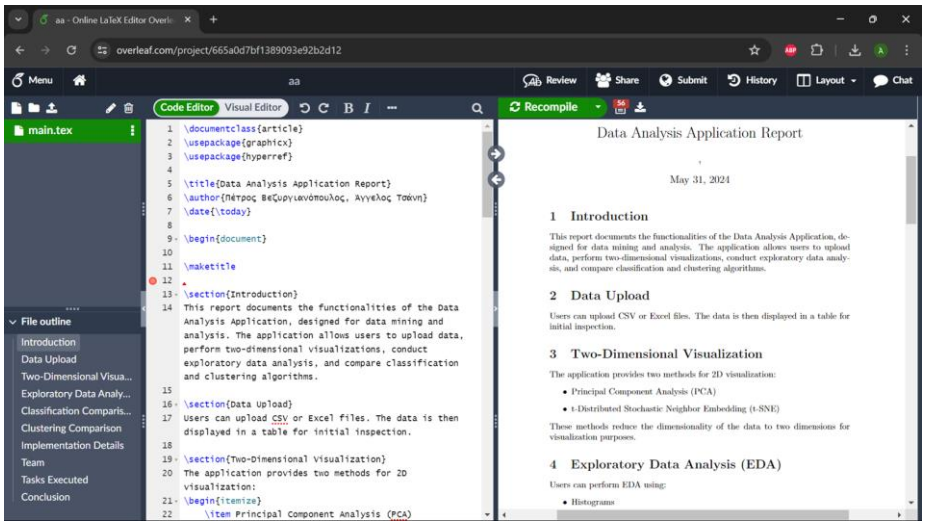

Data Analysis Application
localhost:8501
Deploy
2D Visualization
Select Visualization Algorithm
PCA
t-SNE
label
2018-03-22
2019-07-18
2015-11-05
2021-02-19
2017-05-14
2016-09-29
2014-12-01
2019-04-11
2022-08-24
Dim2
Dim1
Select EDA Plot Type
Histograms
Boxplots

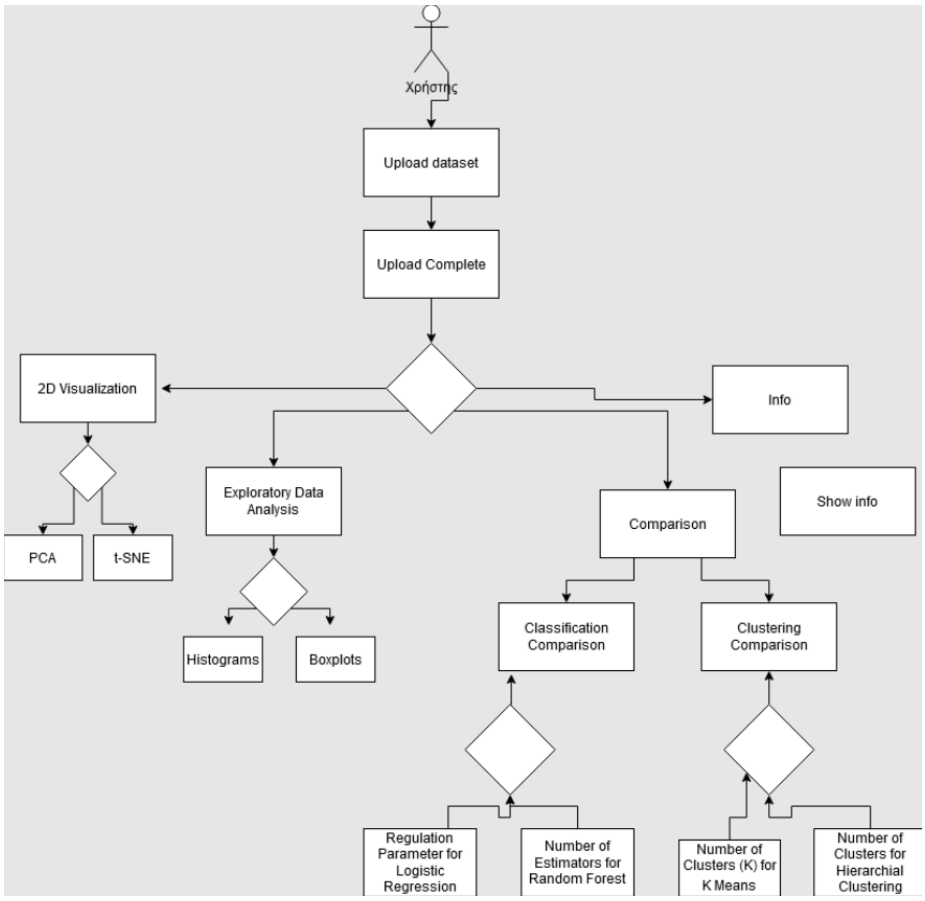

Data Analysis Application
localhost:8501
Deploy
Histogram for 1
count
Histogram for 28
count


Data Analysis Application
localhost:8501
Deploy
Comparison
Regularization Parameter (C) for Logistic Regression
1.00
0.01
10.00
Number of Estimators for Random Forest
10
1
100
Number of Clusters (k) for K-Means
5
2
10
Number of Clusters for Hierarchical Clustering
5
2
10
Run Classification Comparison
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.


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 Βεζυργιανόπουλος
Άγγελος Τσάνη
Tasks Being Executed
2D Visualization, Docker setup, Github version control, Machine learning algorithms comparison, Software release life cycle


Data Analysis Application
localhost:8501
Deploy
Select EDA Plot Type
Histograms
Boxplots
Boxplot for 1
Boxplot for 28


overleaf.com/project/665a0d7bf1388093e92b2d12
Menu
aa
Review
Share
Submit
History
Layout
Chat
Code Editor
Visual Editor
Recompile
main.tex
File outline
Introduction
Data Upload
Two-Dimensional Visua...
Exploratory Data Analy...
Classification Comparis...
Clustering Comparison
Implementation Details
Team
Tasks Executed
Conclusion
Data Analysis Application Report
May 31, 2024
1 Introduction
2 Data Upload
3 Two-Dimensional Visualization
4 Exploratory Data Analysis (EDA)



ΜΕΛΗ ΟΜΑΔΑΣ

1. ΠΕΤΡΟΣ ΒΕΖΥΡΓΙΑΝΟΠΟΥΛΟΣ

2. ΑΓΓΕΛΟΣ ΤΣΑΝΗ