Software Technologies Project, 2024, pp. 1–2

doi: DOI HERE
Advance Access Publication Date: 17th May 2024
Paper

PAPER

Data Analysis App for Thyroid Cancer

Panagiotis Tzounis,[1,*] Iosif Mourikis[2] and Konstantinos Selionis[3]

[1]Ionian University, Department of Informatics, Corfu, 49100, Ionian Islands, Greece, [2]Ionian University, Department of Informatics, Corfu, 49100, Ionian Islands, Greece and [3]Ionian University, Department of Informatics, Corfu, 49100, Ionian Islands, Greece
[*]Corresponding author. email-id.com
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Key words: Data, Application, Informatics, Thyroid Cancer

Introduction

Data analysis is an important part of modern research in Medicine and Health. In this paper, we analyze data related to thyroid cancer using various machine learning algorithms.

Dataset

The data used were obtained from the UCI Machine Learning Repository and are data from patients with thyroid cancer.More specifically, this dataset contains 13 clinicopathological features designed to predict recurrence of well-differentiated thyroid cancer. The dataset was collected over 15 years and each patient was followed for at least 10 years.[1] .

Machine Learning Algorithms

For the analysis of the data, machine learning algorithms were used to compare the data and present accurate results.These algorithms include both clustering and categorization methods.

K-Means Clustering

Description:

The K-Means algorithm is an unsupervised clustering method that attempts to divide a data set into K distinct groups (clusters). Each sample belongs to the cluster with the closest centroid.

Procedure:

1. Define the number of clusters (K).
2. Number of clusters (number of clusters).
3. Assignment of each sample to the nearest centroid.
4. Recalculation of centroids based on current assignments.
5. Repeat steps 3-4 until the centroids converge or a predefined maximum number of iterations is reached.

Application to Work:

To cluster the thyroid cancer data, we used the K-Means algorithm, trying different values for K to find the optimal clustering.

Spectrul Clustering

Description:

The Spectral Clustering algorithm is based on graph theory and uses the eigenvalues of the Laplacian matrix of the data graph for clustering.

Procedure:

1. Construct a graph from the data, where the nodes are the samples and the edges represent the similarities between them.
2. Calculate the Laplacian matrix of the graph.
3. Calculate the eigenvalues and eigenvectors of the Laplacian.
4. Using the first k eigenvectors to reduce the dimensions and applying a simple clustering algorithm (such as K-Means) to the new dimensions.

Application to Work:

Spectral Clustering was used to improve clustering, especially in cases where the cluster structures are nonlinear.

Results

The results obtained from the processing and analysis of the data from the medical clinical research on thyroid cancer are shown in PCA and t-SNE in the 2D Visualization tab.

Software Release Lifecycle

To develop our application, we followed the Agile model, which includes the following stages:

1. Initial Design: defining the requirements and creating the development plan.
2. Data Collection and Preparation: loading and cleaning of data.

[1] `https://archive.ics.uci.edu/`

3. Algorithm development: implementation and optimization of machine learning and dimension reduction algorithms.
4. Interface development: Creating the user interface and integrating the algorithms.
5. Testing and Validation: testing the correctness and performance of the algorithms.
6. Repeat and Improve: repeat the development cycle for improvements based on feedback and test results.
7. Final Delivery and Presentation: final presentation and delivery of the application.

Conclusion

In summary, we want to believe that the creation of our application on the analysis of medical data for Thyroid Cancer, with the help of Machine Learning algorithms, to contribute effectively to the field of both Computer Science and Medicine.More specifically, the application was created using Python and Streamlit, two important tools, but also the medical research we used as without it we would not have had any results.

Github Repository Link

https://github.com/kselionis/Data-Analysis-App-for-Thyroid-Cancer

Author contributions statement

- Selionis Konstantinos: Development of the machine learning algorithm integration and creation of 2D imaging features.
- Mourikis Iosif: Designed the user interface and manipulated the EDA diagrams.
- Tzounis Panagiotis: The data upload and review function was implemented.

Acknowledgments

The authors of this article would like to thank Ms. Shiva Borzooei and Aidin Tarokhian, from Hamadan University of medical Sciences for providing the medical research to the UCI Machine Learning Repository.

References

1. Shiva Borzooei, Giovanni Briganti, Mitra Golparian, Jerome R. Lechien, Aidin Tarokhian. Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study 2023.

Diagram UML

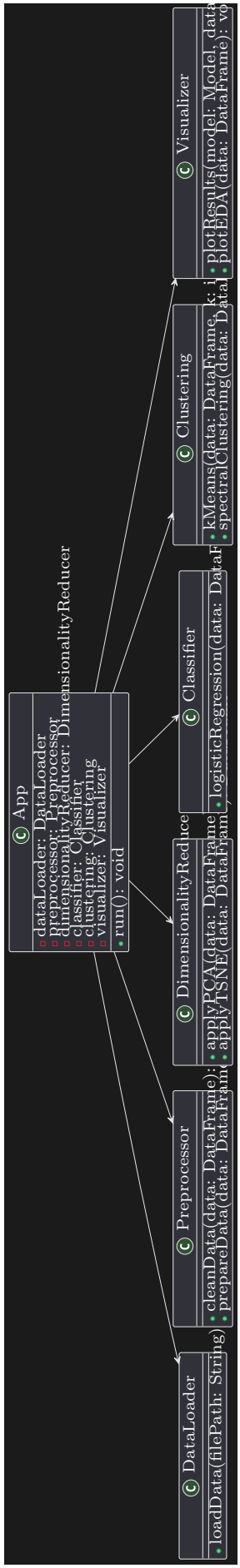


Fig. 1. UML Image of App

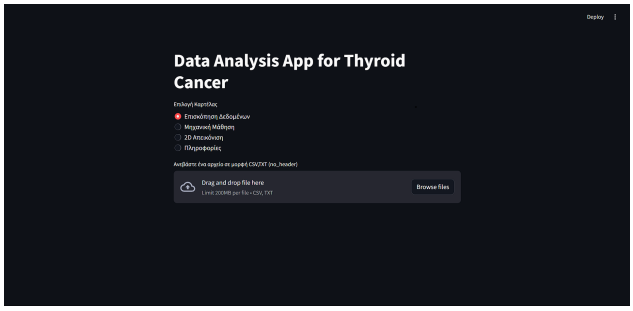
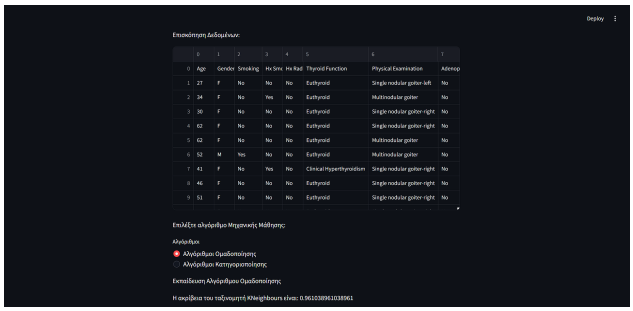


Fig. 2. App



	0	1	2	3	4	5	6	7
0	Age	Gender	Smoking	Hx Smo	Hx Rad	Thyroid Function	Physical Examination	Adenop
1	27	F	No	No	No	Euthyroid	Single nodular goiter-left	No
2	34	F	No	Yes	No	Euthyroid	Multinodular goiter	No
3	30	F	No	No	No	Euthyroid	Single nodular goiter-right	No
4	62	F	No	No	No	Euthyroid	Single nodular goiter-right	No
5	62	F	No	No	No	Euthyroid	Multinodular goiter	No
6	52	M	Yes	No	No	Euthyroid	Multinodular goiter	No
7	41	F	No	Yes	No	Clinical Hyperthyroidism	Single nodular goiter-right	No
8	46	F	No	No	No	Euthyroid	Single nodular goiter-right	No
9	51	F	No	No	No	Euthyroid	Single nodular goiter-right	No

Fig. 3. App Data

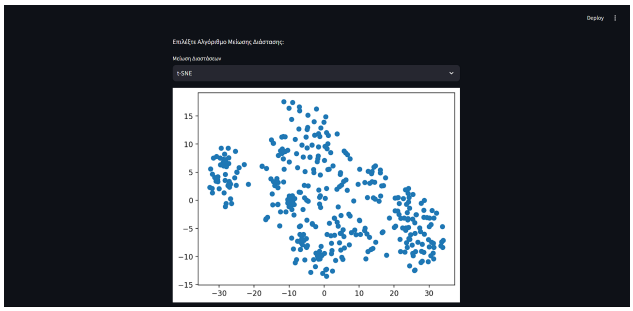


Fig. 4. App 2D Visualization