Diabetes Risk: The Prediction of Diabetes Risk Through Unsupervised Learning

By: Katherine Nguyen

Data Science
UNIVERSITY OF COLORADO BOULDER

Agenda

- Outline/Agenda
- Introduction
- Approach
  - Data
  - Rundown
  - Data Cleaning
  - Models
  - Model Analysis
- Results
- Conclusion

Introduction

- **Problem**
  - Diabetes is prevalent and being able to predict whether one has diabetes is useful
  - How can we predict diabetes based on potential risk factors of diabetes?
- **Purpose**
  - To understand the diabetes risk factors that can help predict whether an individual will acquire diabetes based on those risk factors in the future
- **Why Is it Important?**
  - To raise awareness and potential address prevention strategies for earlier stages before diabetes

Approach: Data

- **Dataset:** "Diabetes Risk Prediction"
  - Diabetes Database: https://www.kaggle.com/datasets/rcratos/diabetes-risk-prediction
  - Categorical Data; Binary

Approach: Data

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes

Approach: Data

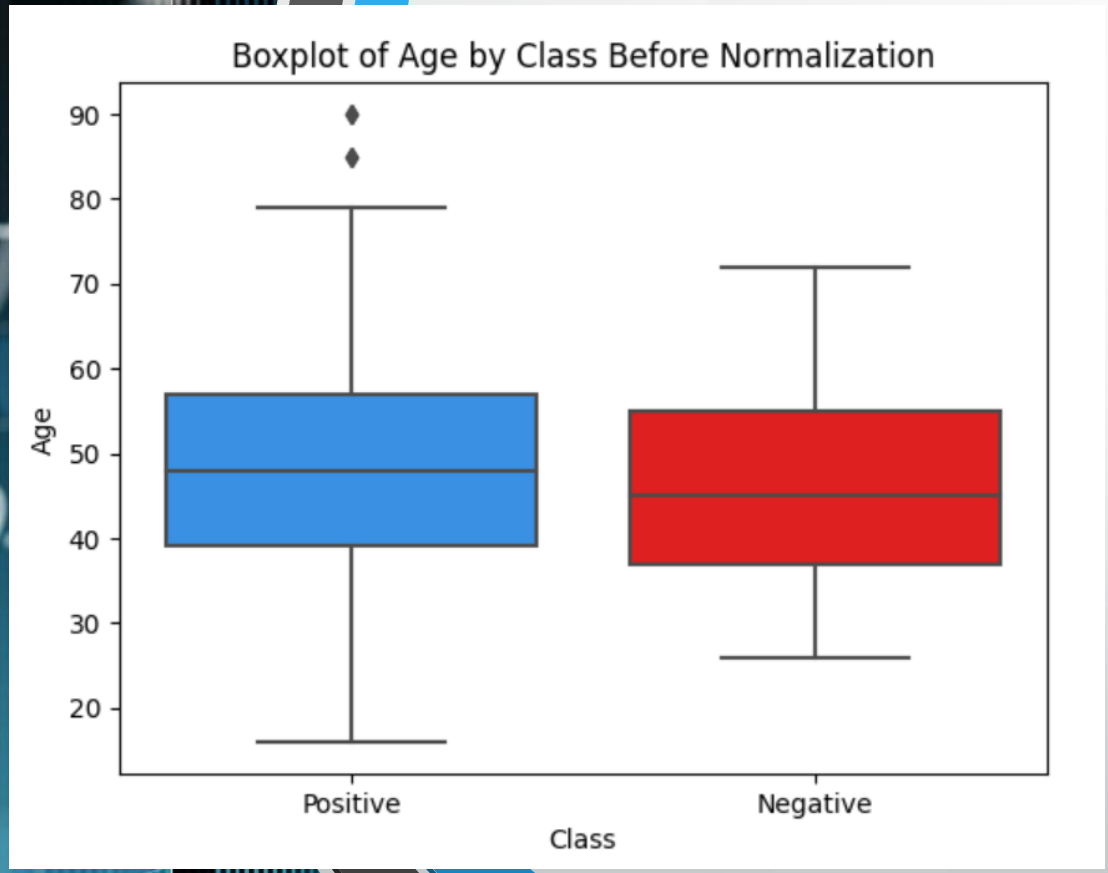
	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	1	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
1	58	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1
2	41	1	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1
3	45	1	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1
4	60	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	39	0	1	1	1	0	1	0	0	1	0	1	1	0	0	0	1
516	48	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	1
517	58	0	1	1	1	1	1	0	1	0	0	0	1	1	0	1	1
518	32	0	0	0	0	1	0	0	1	1	0	1	0	0	1	0	0
519	42	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

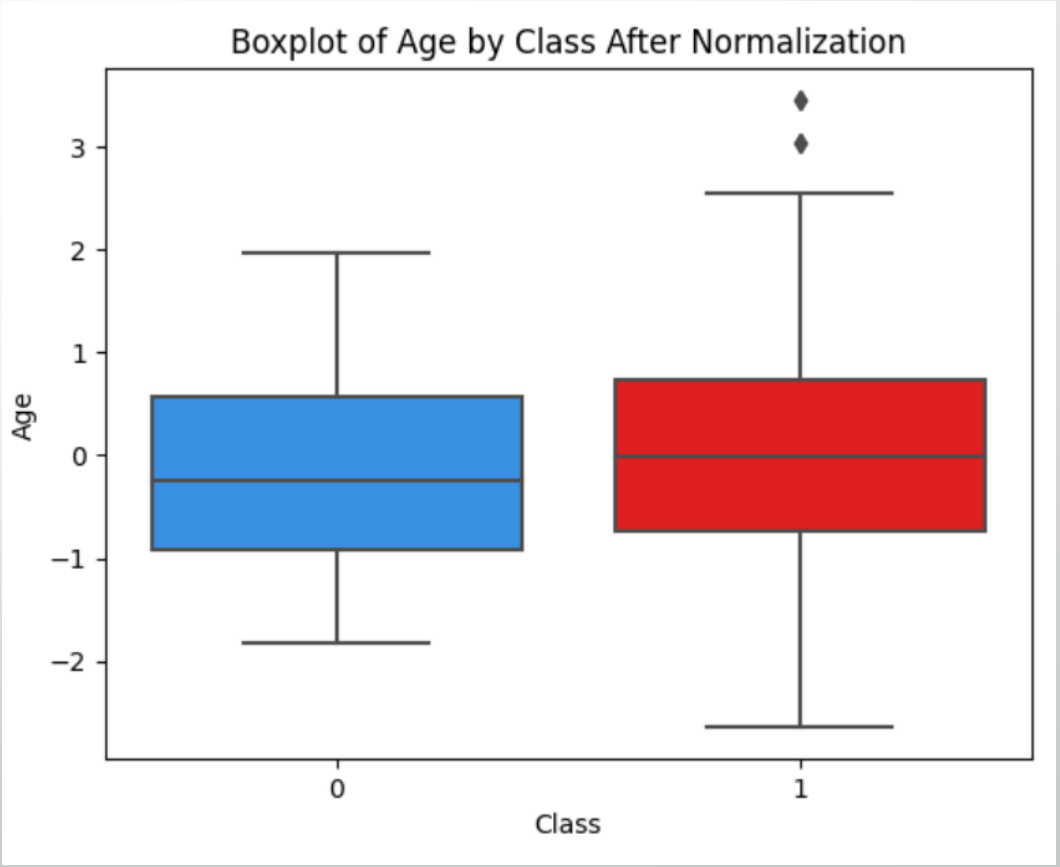
520 rows × 17 columns

Approach: Rundown

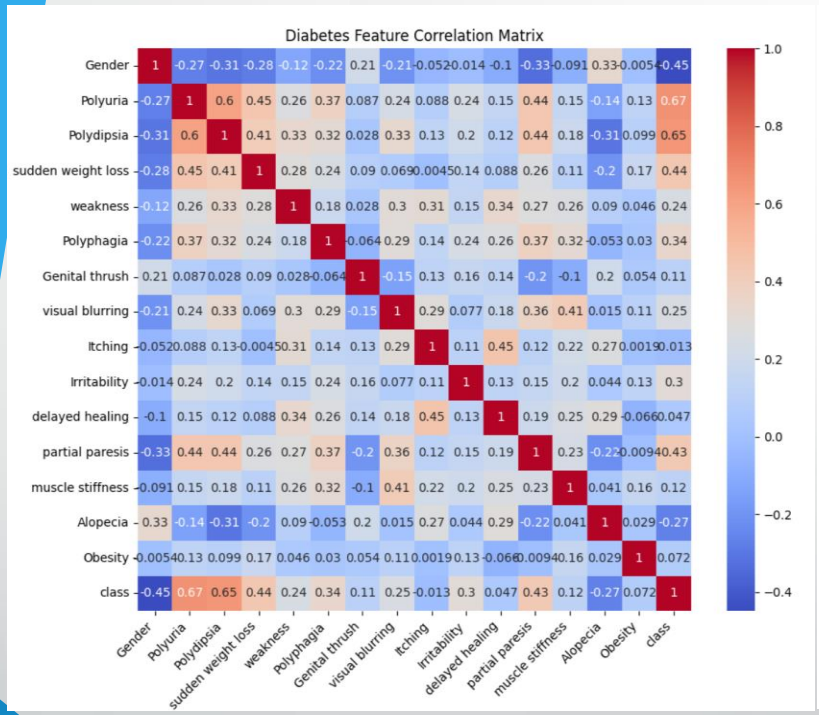
- **Data Cleaning :**
  - Simplified the Data
    - Convert labels into binary values (e.g. Yes = 1, No = 0)
    - Removed unnecessary features for prediction
    - Use PCA to Normalize Data, and Remove Unimportant Features
    - Use Non-Negative Matrix to reduce dimensionality, makes easier for clustering.
- **Exploratory Data Analysis :**
  - Used Bar Plots to show classifications of diabetes in respect to other features

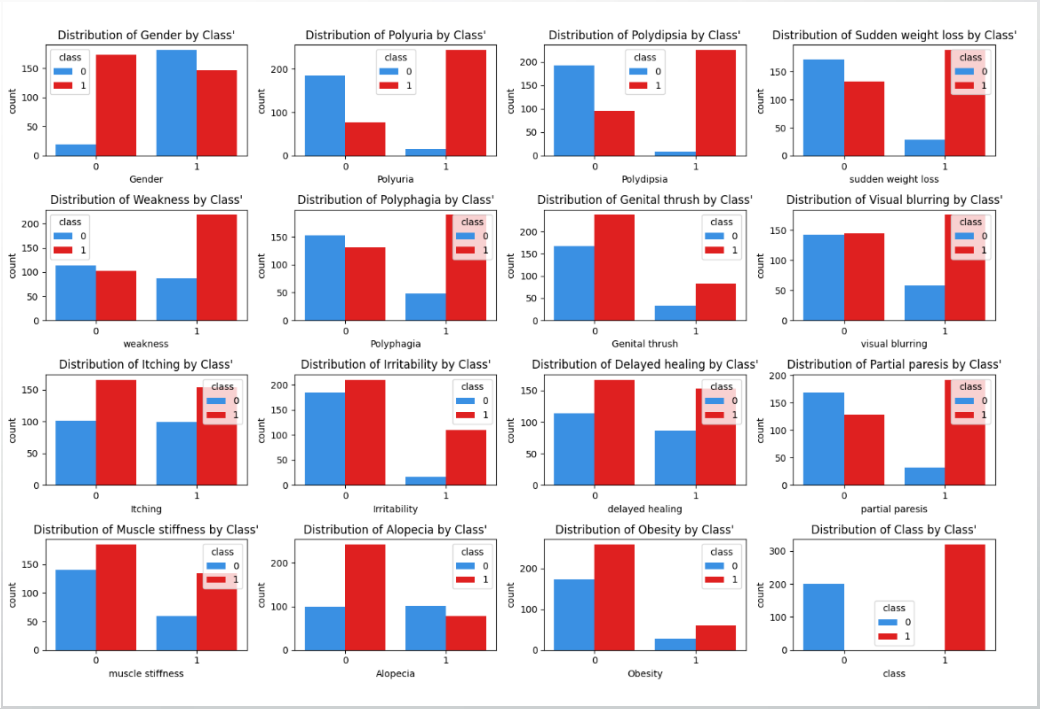
Approach: Data Cleaning





Approach: Data Cleaning





Approach: Data Cleaning

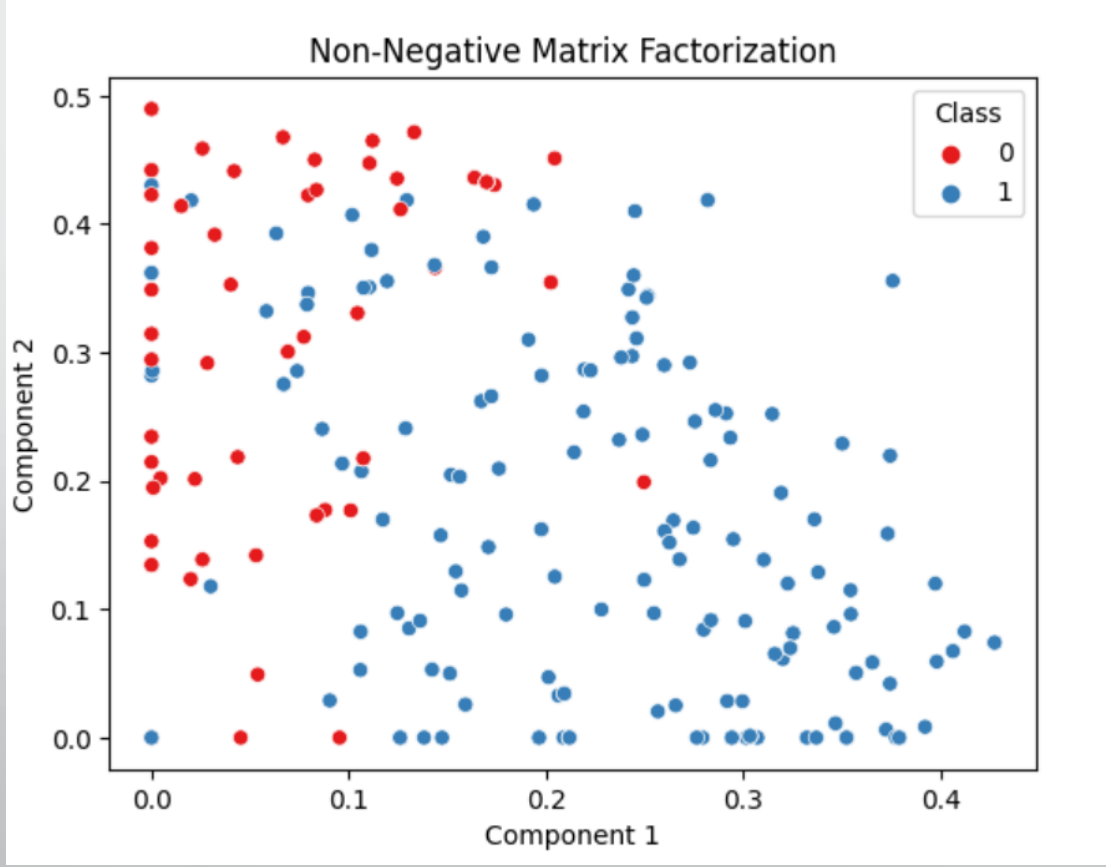
	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	p p
0	1	0	1	0	1	0	0	0	1	0	1	0
1	1	0	0	0	1	0	0	1	0	0	0	1
2	1	1	0	0	1	1	0	0	1	0	1	0
3	1	0	0	1	1	1	1	0	1	0	1	0
4	1	1	1	1	1	1	0	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
515	0	1	1	1	0	1	0	0	1	0	1	1
516	0	1	1	1	1	1	0	0	1	1	1	1
517	0	1	1	1	1	1	0	1	0	0	0	1
518	0	0	0	0	1	0	0	1	1	0	1	0
519	1	0	0	0	0	0	0	0	0	0	0	0

Approach: Models

- **Unsupervised Learning:**
  - K-Means Clustering
    - Predicts where assumed clusters will be in data
  - Hierarchical Clustering
    - Predicts clusters based on similarities and branches to next decision

Approach: Models

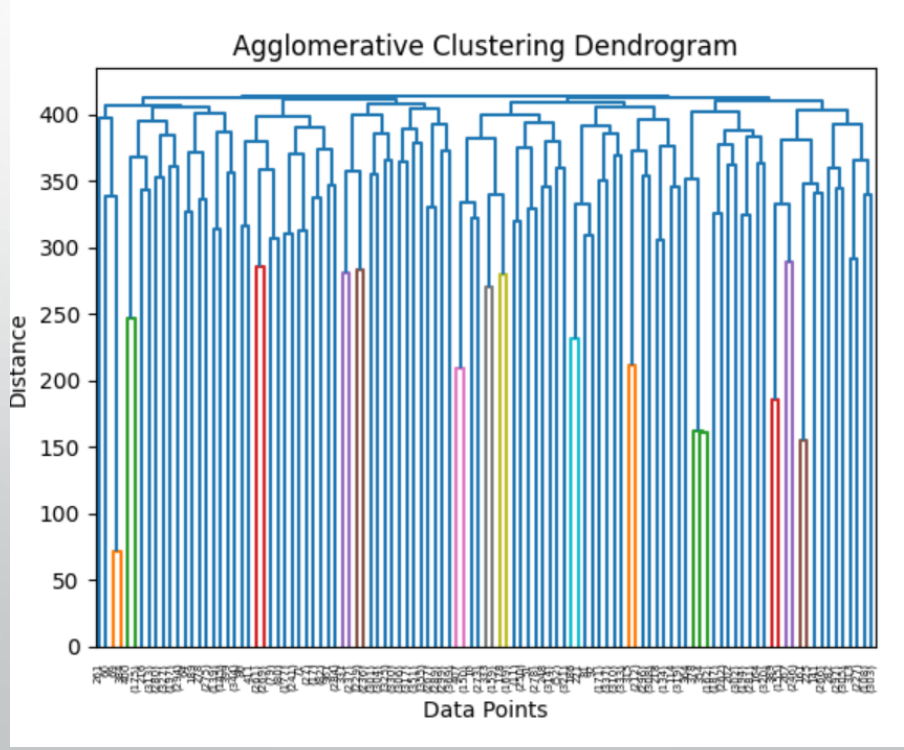
Dimensionality Reduction with Non Negative Matrix



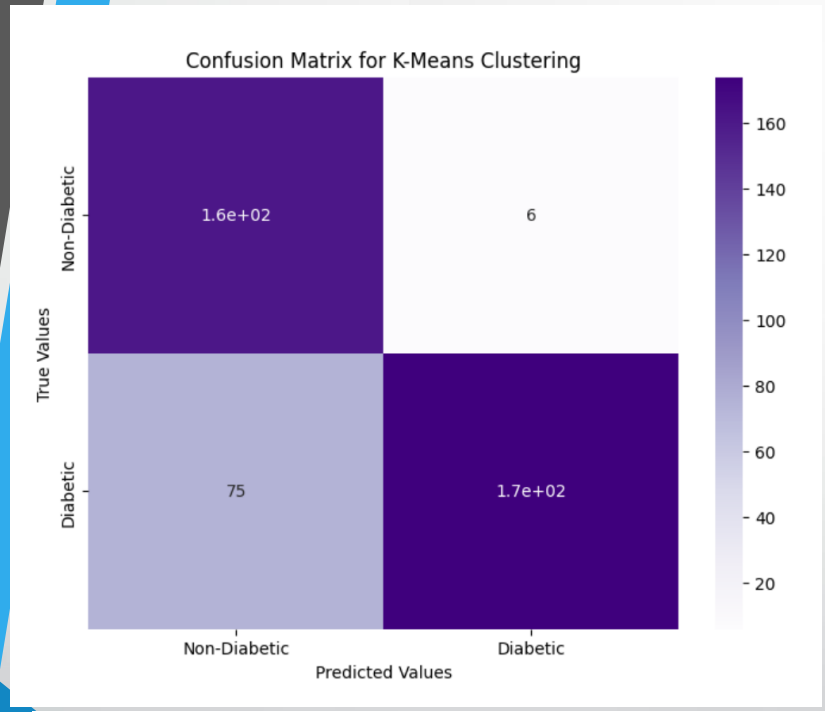
Approach: Models

Logistic Regression

Hierarchical Clustering

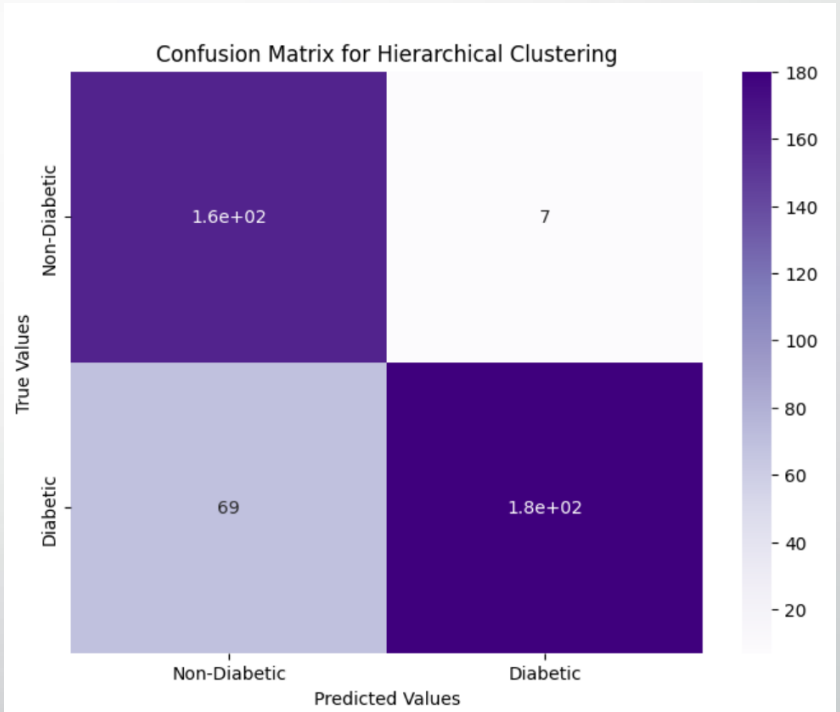


Results: Model Analysis



K-Means Clustering Label Ordering: (0, 1)
K-Means Clustering Accuracy: 0.8052884615384616
K-Means Clustering Precision: 0.9666666666666667

Silhouette Score: 0.44072574501920675



Hierarchical Clustering Label Ordering: (0, 1)
Hierarchical Clustering Accuracy: 0.8173076923076923
Hierarchical Clustering Precision: 0.9625668449197861

Silhouette Score: 0.37396613190254996

Results

- **Evaluation:**
  - K-Means Cluster
    - Somewhat accurate and precise; Decent predictor model
    - Not ideal based on silhouette score; lacks similarity in clusters
    - Has potential with different parameters and less features
  - Decision Tree classifier
    - Somewhat accurate and precise; Decent predictor model
    - Not ideal based on silhouette score; lacks similarity in clusters
      - Worse than K-Means Cluster
    - Has potential with different parameters and less features

Conclusion

- **K-Means Clustering** is the better model and has potential, but it fails to cluster by similarity, probably due to too many features
- **Hierarchical Clustering** has potential, but it fails to cluster by similarity, probably due to too many features and hyperparameters