Defining students groups with clustering

Natalie Novosad

UNIVERSITY OF TORONTO
Technology UpSkilling
VECTOR INSTITUTE

Toronto, 2022

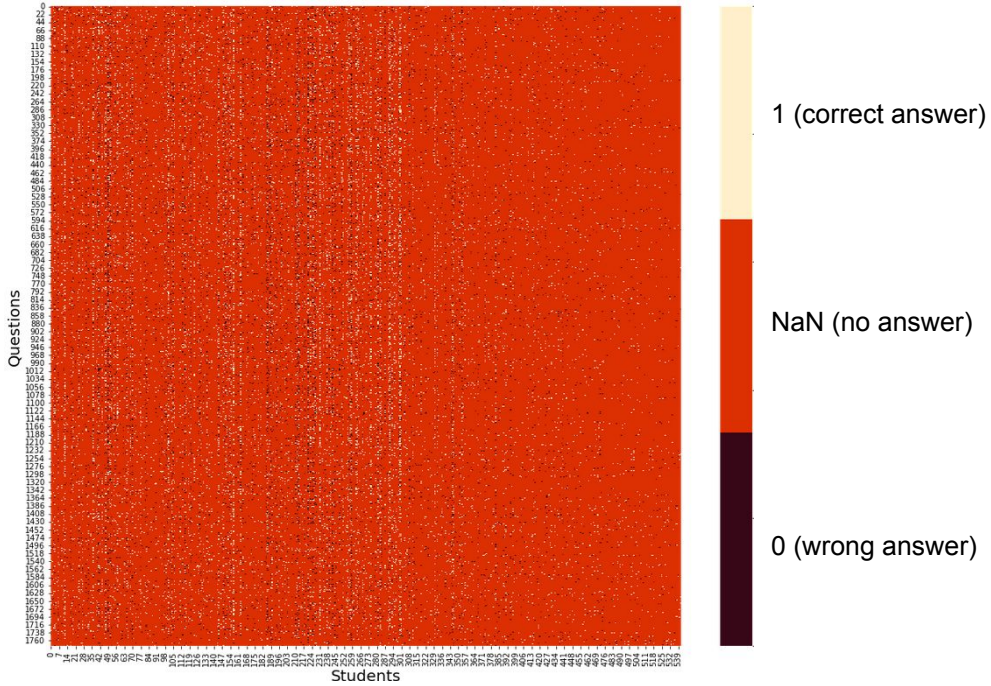
Task statement

542 students (~2004, max 18 years)

1774 questions (Math)

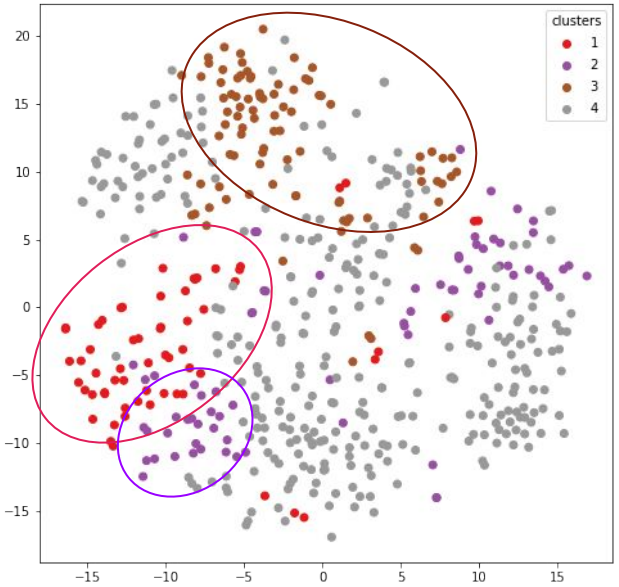
Goal:

1. split the students into the groups based on how they answering the questions, describe them.
2. predict the answers to the questions



Final groups

1. **struggling** - students who does lots of questions, but most of them are incorrect
2. **random/inactive** - students who did the least questions, and the success rate is almost random
3. **good students** - good success rate and lots of questions are done
4. **best performers** - the highest success rate



* visualization was built with T-SNE

Step 1: NaN imputation and distance calculation

A. Matrix imputation:
   a. 1 for correct answer
   b. 0 for no answer
   c. -1 for wrong answer
B. Calculate distance matrix (cosine, Euclidean) only on known answers.
   Problem: some students' questions don't overlap:
   a. just fill with 0
   b. graph representation and calculate the distance between those students through others.

Step 2: clustering algorithm and parameter selection

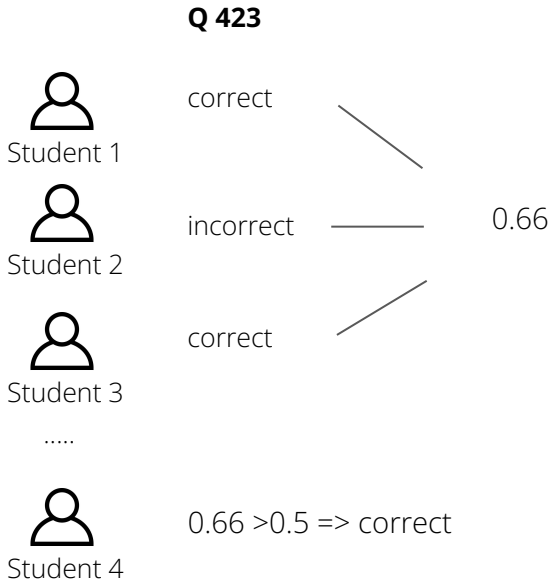
K-means	Hierarchical	Spectral	Gaussian Mixture
- too random	- slower	+ non-linear	+ different sizes of cluster
- equal cluster sizes (in distance context)	- does not give exact number of clusters	+ flexible	- assumes normality of the data
+ fast	+ uses interpretative distance metric		

Step 2: clustering algorithm and parameter selection

K-means	Hierarchical	Spectral	Gaussian Mixture
- too random	- slower	+ non-linear	+ different sizes of cluster
- equal cluster sizes (in distance context)	- does not give exact number of clusters	+ flexible	- assumes normality of the data
+ fast	+ uses interpretative distance metric Cosine distance [-1,0,1] imputation		

Step 3: majority classifier (MC)

Prediction is the rate of correct answers among all answers.
If rate > 0.5, we expect correct answer.

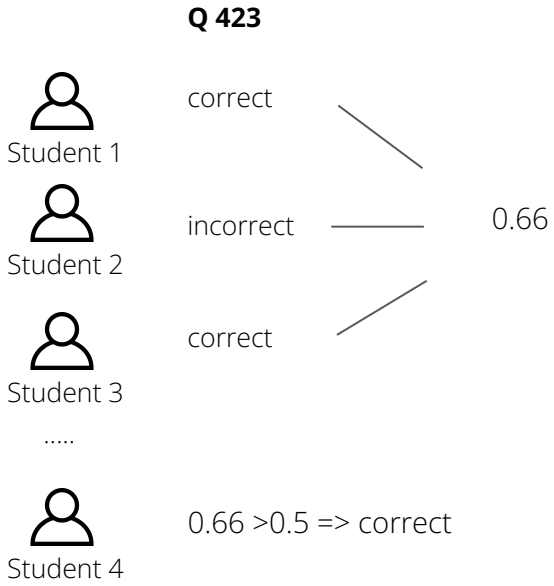


Step 3: majority classifier (MC)

Prediction is the rate of correct answers among all answers.
If rate > 0.5, we expect correct answer.

Accuracy of general MC = 62.48%

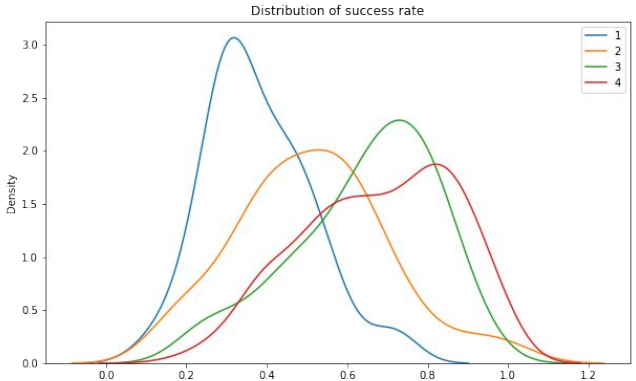
Accuracy of MC with clustering = 66.05%

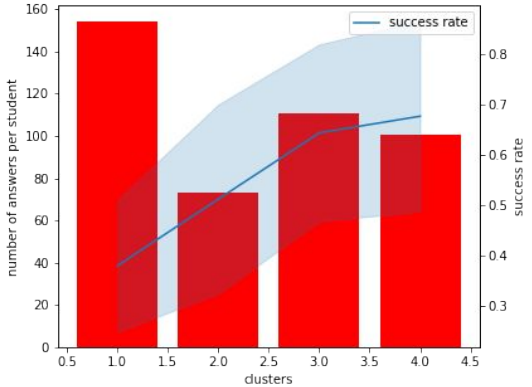


More information about the clusters

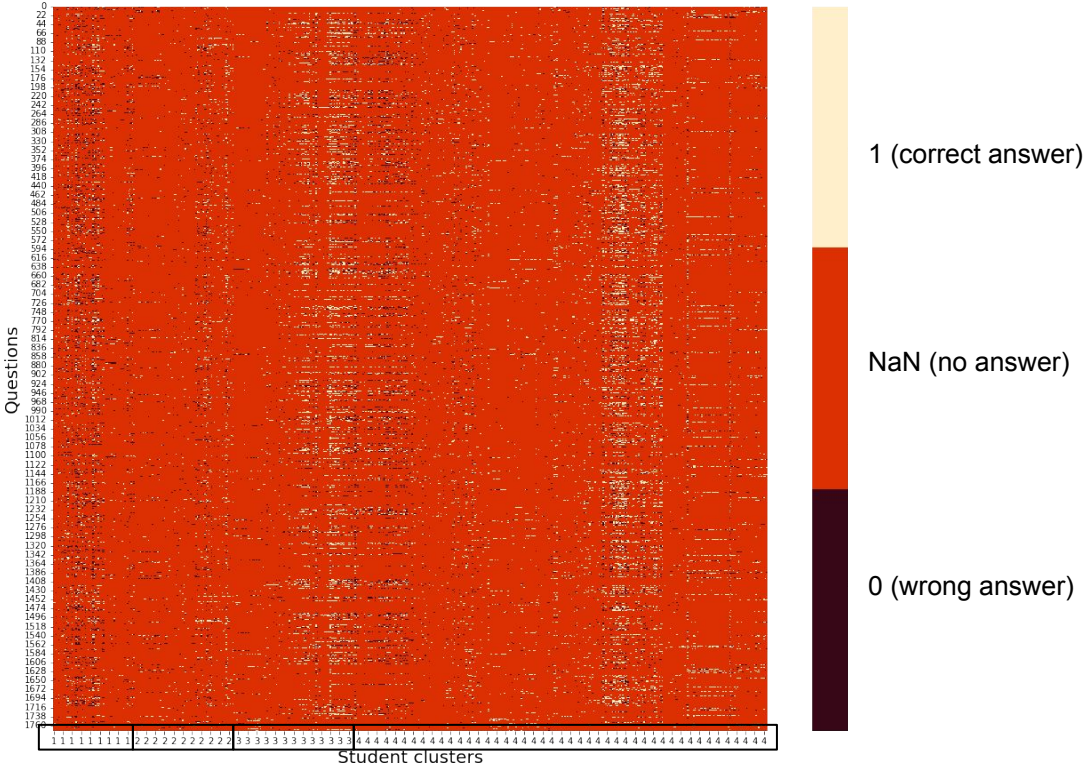
1. **struggling** (62 students)
2. **random/inactive** (75 students)
3. **good students** (90 students)
4. **best performers** (315 students)

* success rate = number of correct answers divided by the number of all answers

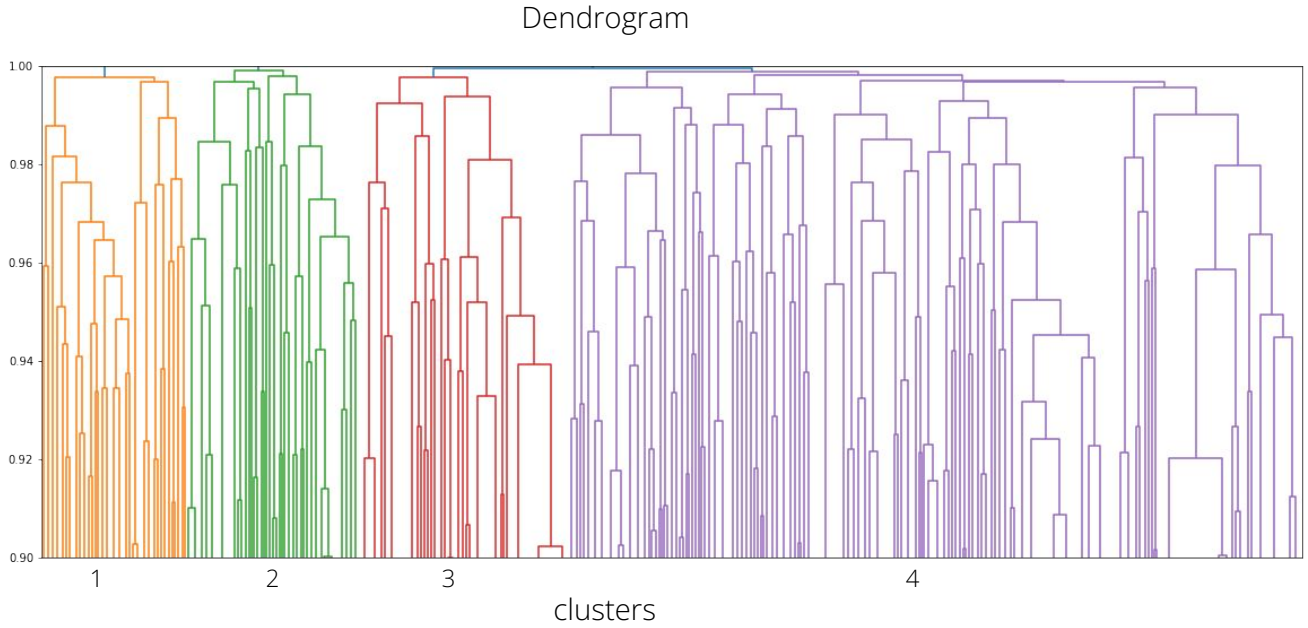




More information about the clusters

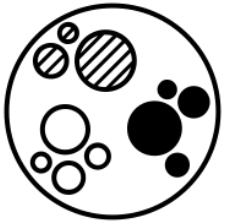


More information about the clusters



Business results

Individual approach to each student can be overwhelming



Clustering helps to unite similar users into groups

Students will be satisfied with more individual recommendations

Future work

1. Use cluster labels as a feature for ML model
2. Incorporate additional features for modeling and analysis
3. Create more interpretable clusters with rule-base hypotheses

Thank you!
See you on Dec 20!

Natalia Novosad
natalinovosad27@gmail.com

UNIVERSITY OF TORONTO Technology UpSkilling VECTOR INSTITUTE

Toronto, 2022

Thank you!
Q&A