Prepared by Nikshita Ranganathan, Archit Barua
Shyamala Venkatakrishnan, Heejae Roh
Professor Venkata Duvvuri

June 8th, 2023

3DHEALS®
CLUSTERING MODEL

Northeastern University

3DHEALS®

01 **Business Problem**
Building Customer email lookalikes

02 **Clustering Model**
Targeting with K-mode clustering

03 **Clustering Visualization**
Exploratory Data Analysis

04 **Predictive Model**
For Customer Behavior

05 **Recommendation & Future Plan**

06 **Q&A**

Business Problem

01

Building Customer email lookalikes

Business Problem

- **Attendee Segmentation:** Cluster attendees based on 'Industry', 'Organization', 'Job Title', 'Country/Region', and 'Source Name' for enhanced understanding of attendee demographics.

- **Event Attendance Prediction:** Utilize features like 'Industry', 'Organization', 'Job Title', 'Country/Region', etc., to predict likely attendance for future events.

- **Webinar/Session Optimization:** Analyze 'Time in Session (minutes)', 'Join Time', 'Leave Time', and 'Questions & Comments' to optimize the timing and content of sessions.

- **Consent Management:** Observe fields related to consent to understand attendee comfort levels with data sharing and recording, thereby improving data usage planning and addressing privacy concerns.

- **Sponsorship Analysis:** If 'Source Name' indicates sponsors, analyze the success of different sponsors in attracting attendees.

- **Attendee Origin Analysis:** Examine 'Country/Region', 'City', 'State/Province', and 'Zip/Postal Code' to gain insights into the geographical distribution of attendees for targeted marketing efforts.

- **Marketing Channel Effectiveness:** Analyze 'Email' domain, 'Linkedin Link', and 'Source Name' to determine which marketing channels drive the highest event attendance.

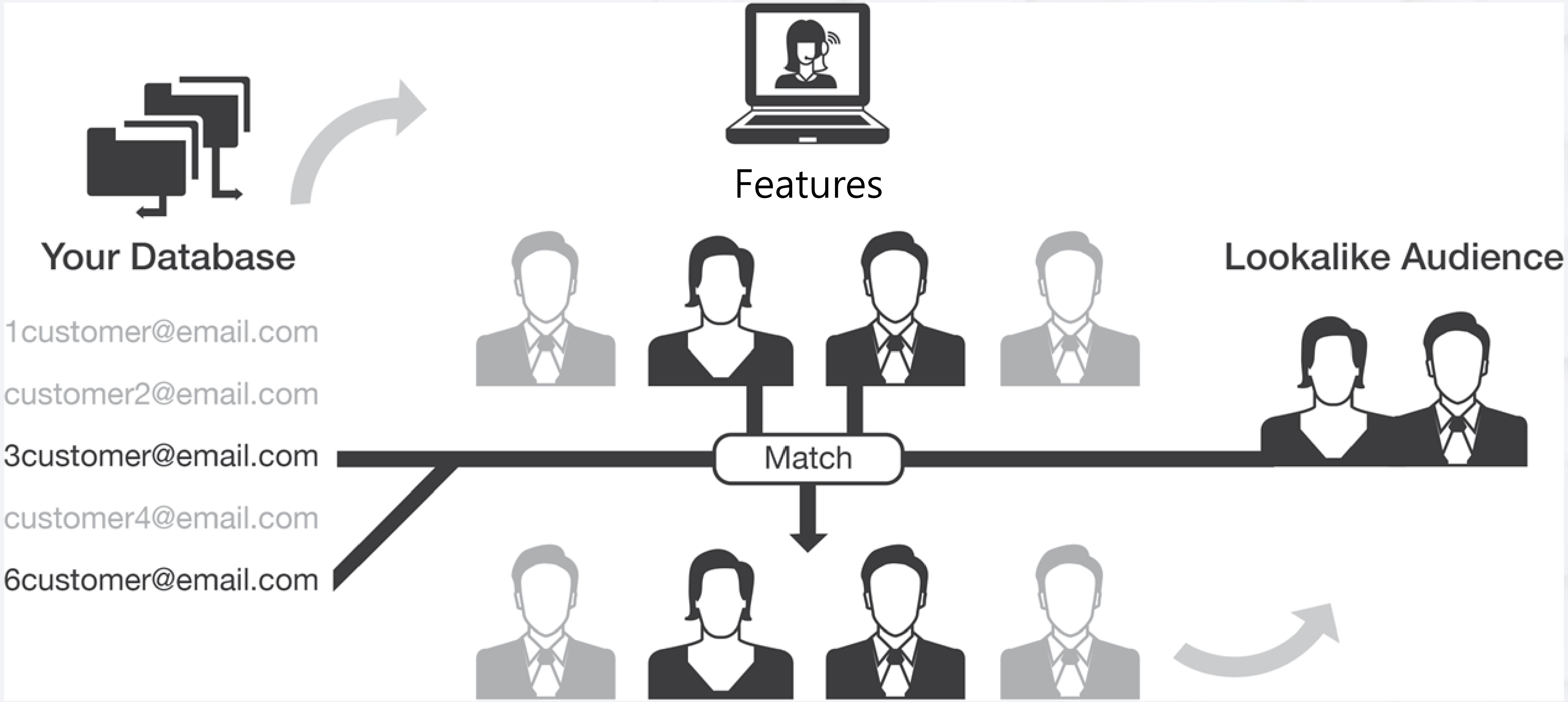
- **Session Feedback Analysis:** Mine 'Questions & Comments' for insights about the sessions' strengths and areas for improvement.

Clustering Model

02

Targeting with K-mode clustering

Model Target: Building customer email lookalike



Features

Your Database

1customer@email.com

customer2@email.com

3customer@email.com

customer4@email.com

6customer@email.com

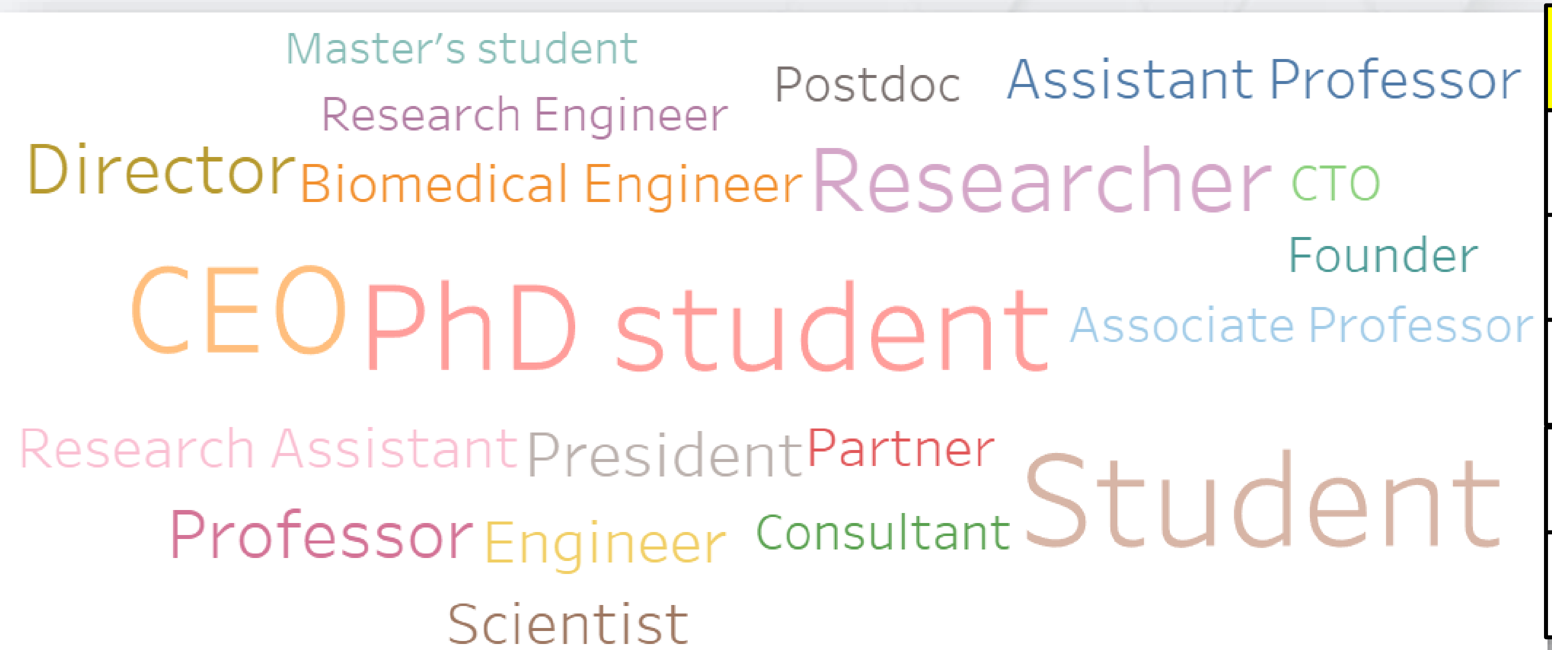
Match

Lookalike Audience

K-Mode Clustering

What if the data is

Categorical



Job Title

Country

Industry

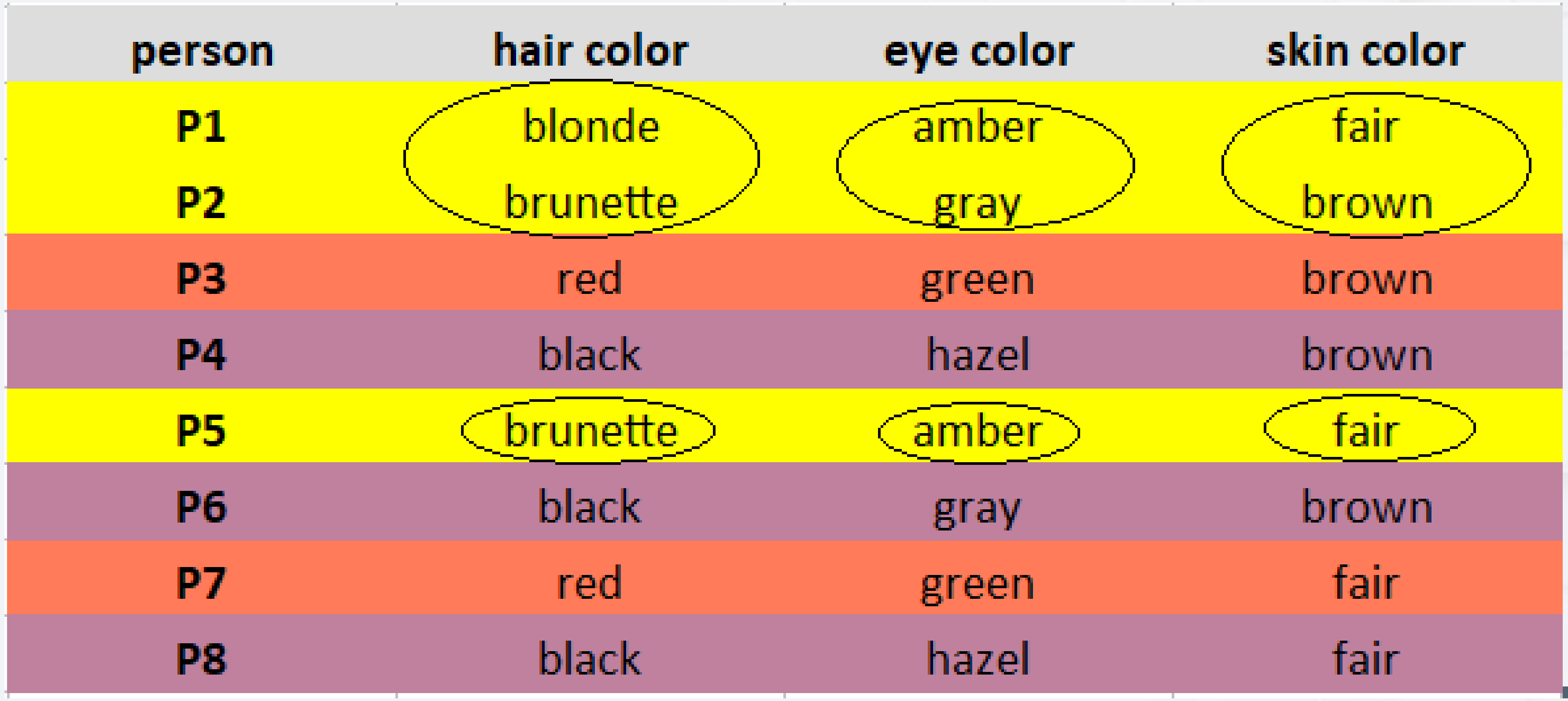
Organization

Source Name

Domain

K-Mode Clustering

Select Modes instead of average



person	hair color	eye color	skin color
P1	blonde	amber	fair
P2	brunette	gray	brown
P3	red	green	brown
P4	black	hazel	brown
P5	brunette	amber	fair
P6	black	gray	brown
P7	red	green	fair
P8	black	hazel	fair

Data Cleaning

□ .com
□ .net

Removing unwanted variables

Removed: 'User Name (Original Name)','First Name','Last Name','Is Guest’, ...

01

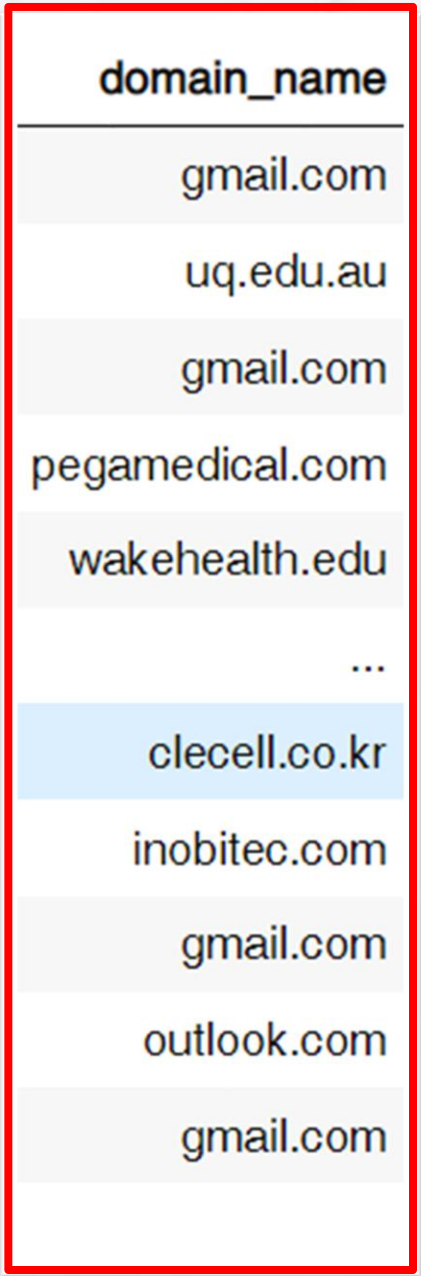
02

03

Extracting Domain Name

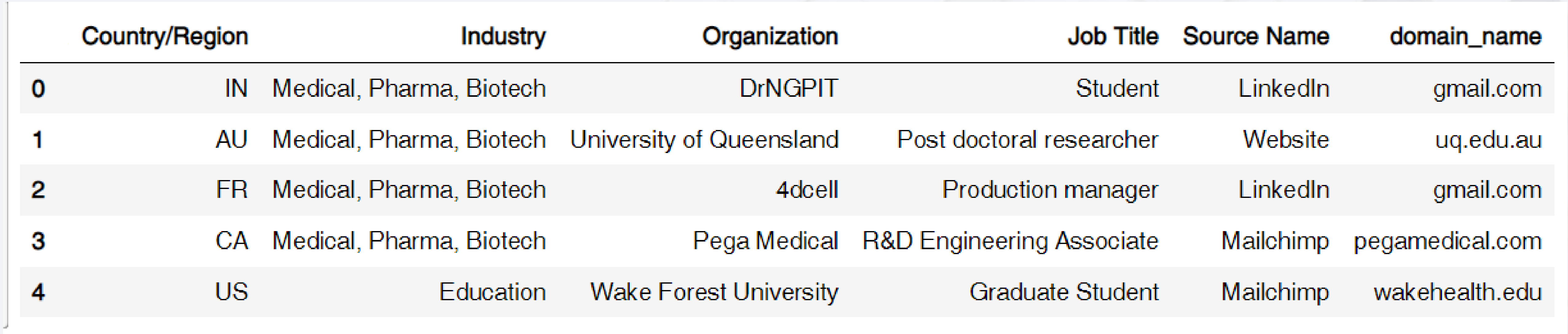
From email

Removing empty/missing records



domain_name
gmail.com
uq.edu.au
gmail.com
pegamedical.com
wakehealth.edu
...
clecell.co.kr
inobitec.com
gmail.com
outlook.com
gmail.com

After Data Cleaning



	Country/Region	Industry	Organization	Job Title	Source Name	domain_name
0	IN	Medical, Pharma, Biotech	DrNGPIT	Student	LinkedIn	gmail.com
1	AU	Medical, Pharma, Biotech	University of Queensland	Post doctoral researcher	Website	uq.edu.au
2	FR	Medical, Pharma, Biotech	4dcell	Production manager	LinkedIn	gmail.com
3	CA	Medical, Pharma, Biotech	Pega Medical	R&D Engineering Associate	Mailchimp	pegamedical.com
4	US	Education	Wake Forest University	Graduate Student	Mailchimp	wakehealth.edu

6 Variables
3,078 Observations

Finding optimal k with elbow method

Column	Unique
Country	87
Industry	30
Organization	1,876
Job Title	1,291
Source Name	9
Domain	745

Select Cluster variables

Decide n_Cluster = 2

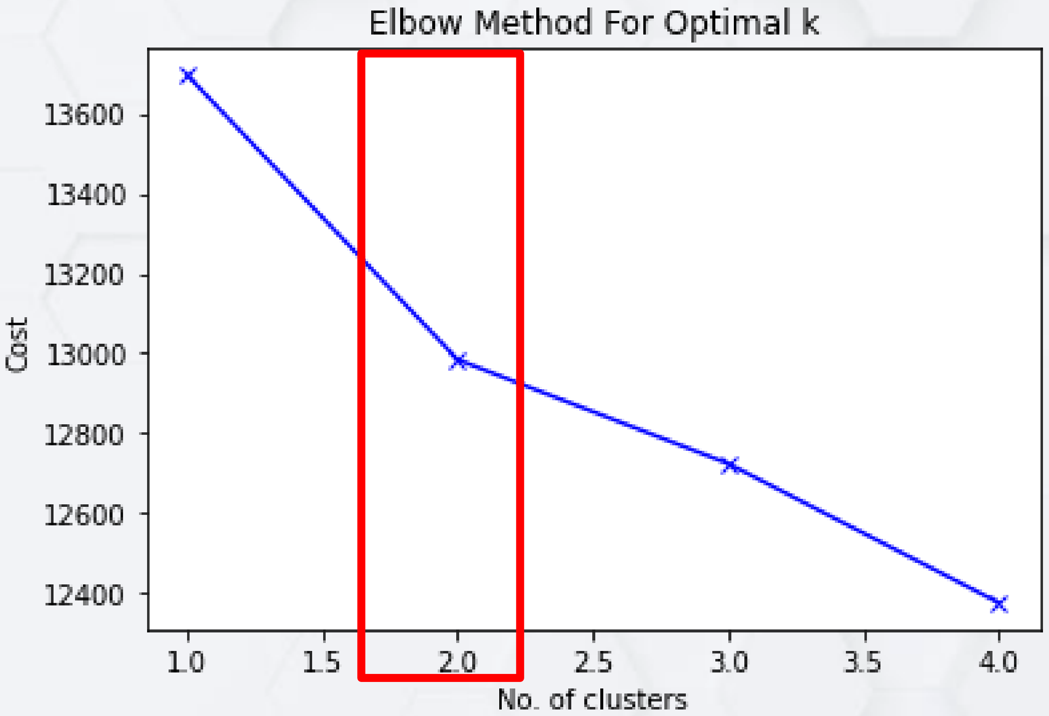
01

02

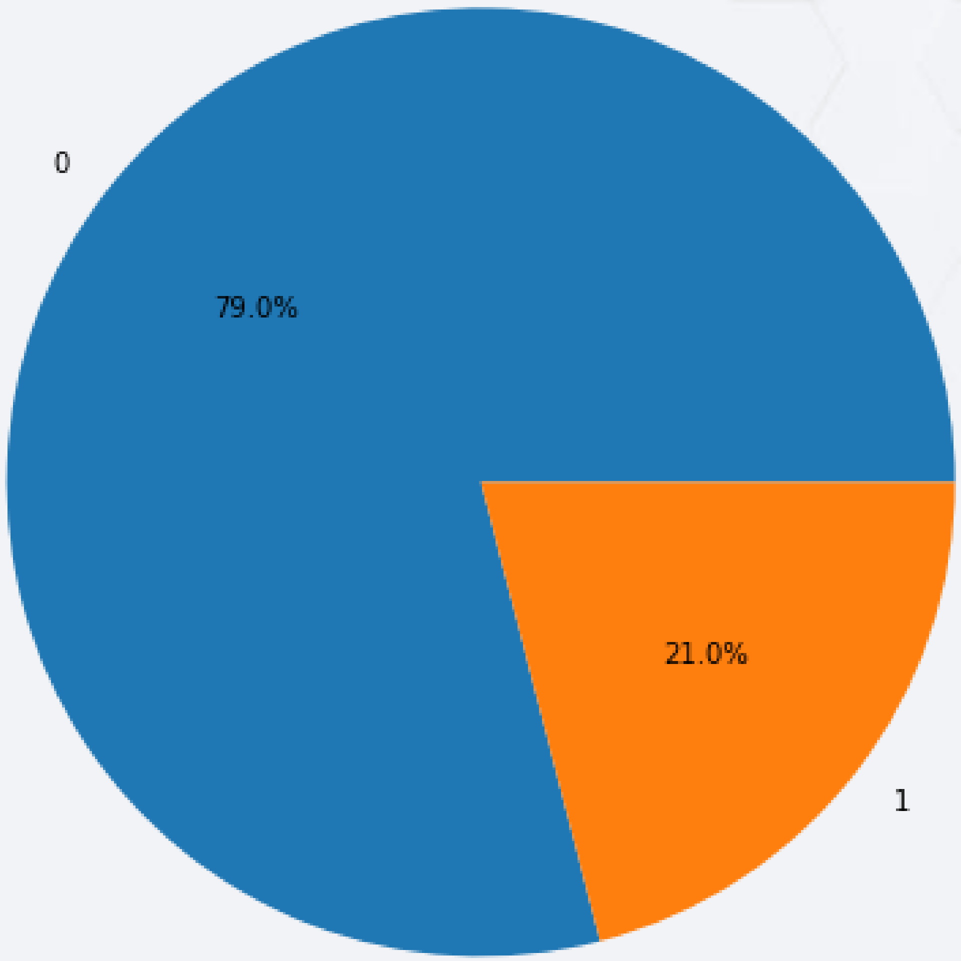
03

Elbow curve

Finding optimal ‘K’



Cluster Ratio



Cluster	Counts	Percent
0	2,505	79.0%
1	573	21.0%

Cluster ‘0’ and Centroids

US
United States
Country/Region

Medical, Pharma, Biotech
Industry

VoxCell BioInnovation
Organization

CEO
Job Title

Mailchimp
Source Name

Gmail.com
Domain_name

Cluster ‘1’ and Centroids

IN
India
Country/Region

Education
Industry

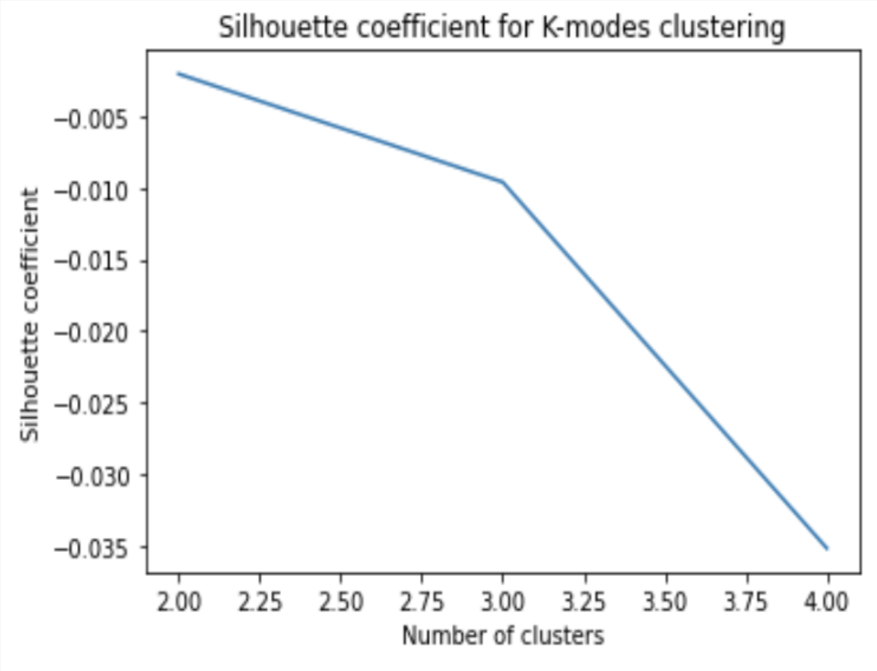
The Hong Kong Polytechnic University
Organization

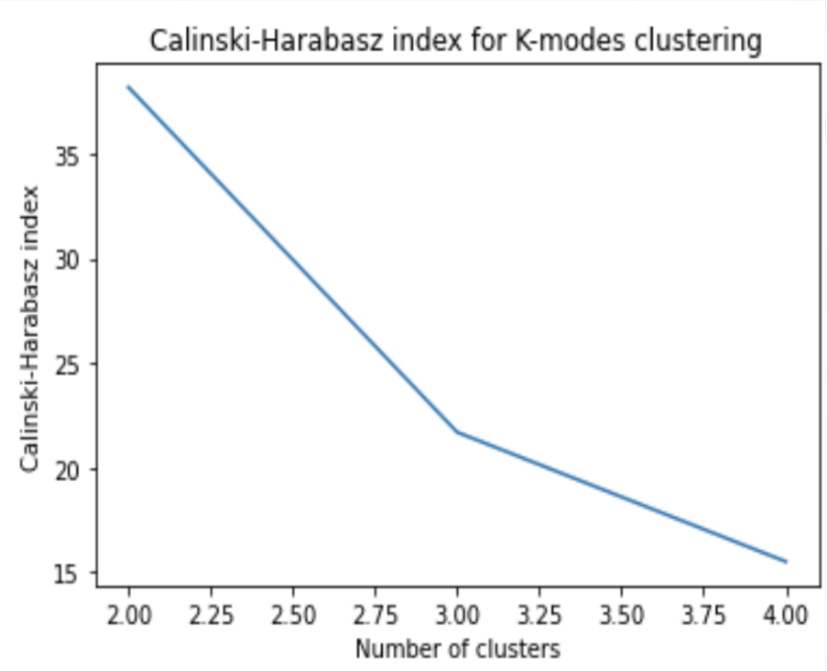
Student
Job Title

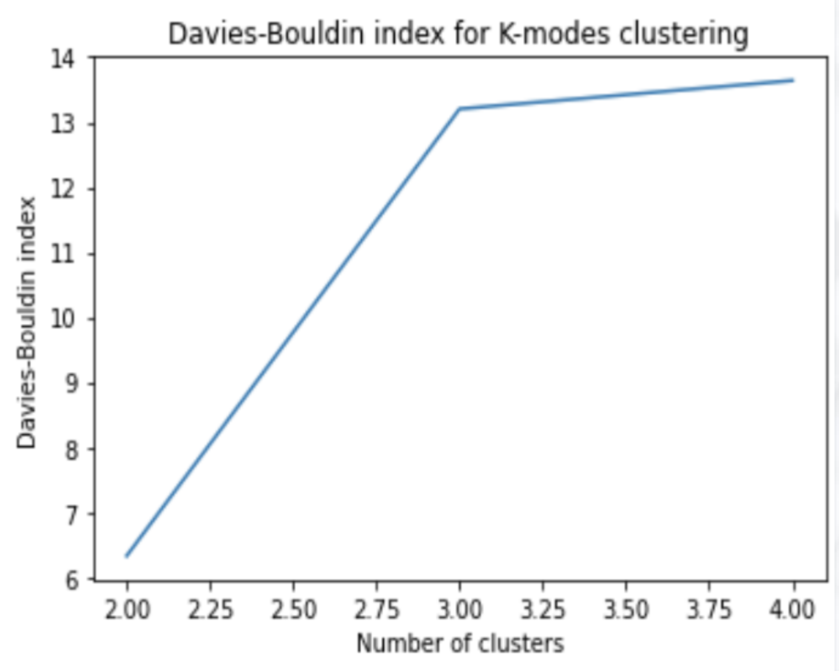
Mailchimp
Source Name

Poly.edu.hk
Domain_name

Validating the effectiveness of clusters







Silhouette coefficient

- A measure of how well each data point is assigned to its cluster.
- A high value indicates that the data point is well-assigned to its cluster.

Calinski - Harabasz index

- A measure of the separation between clusters.
- A high Calinski - Harabasz index indicates that the clusters are well-separated.

Davies-Bouldin index

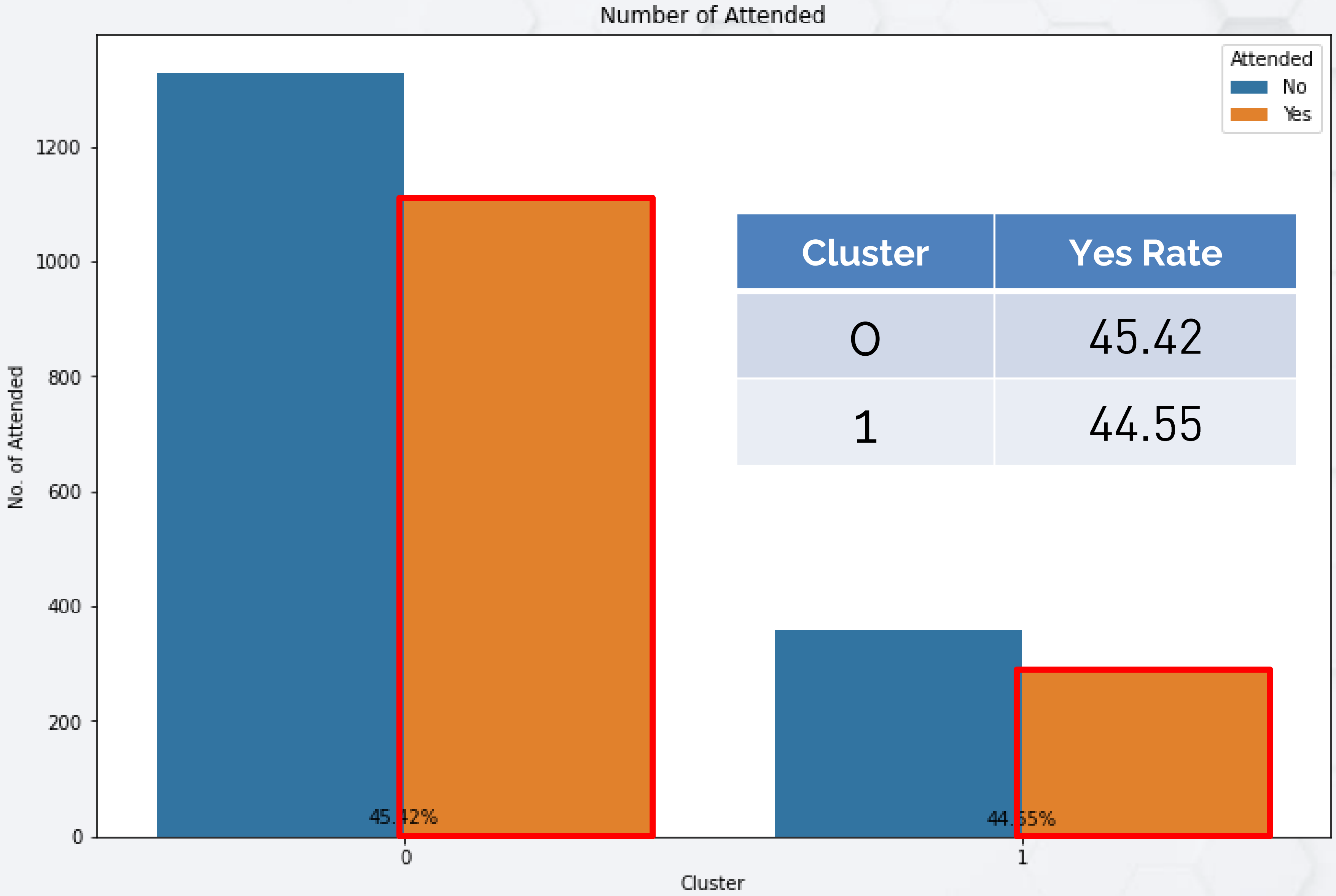
- A measure of the compactness and separation of clusters.
- A low value indicates that the clusters are compact and well-separated.

Clustering Visualization

03

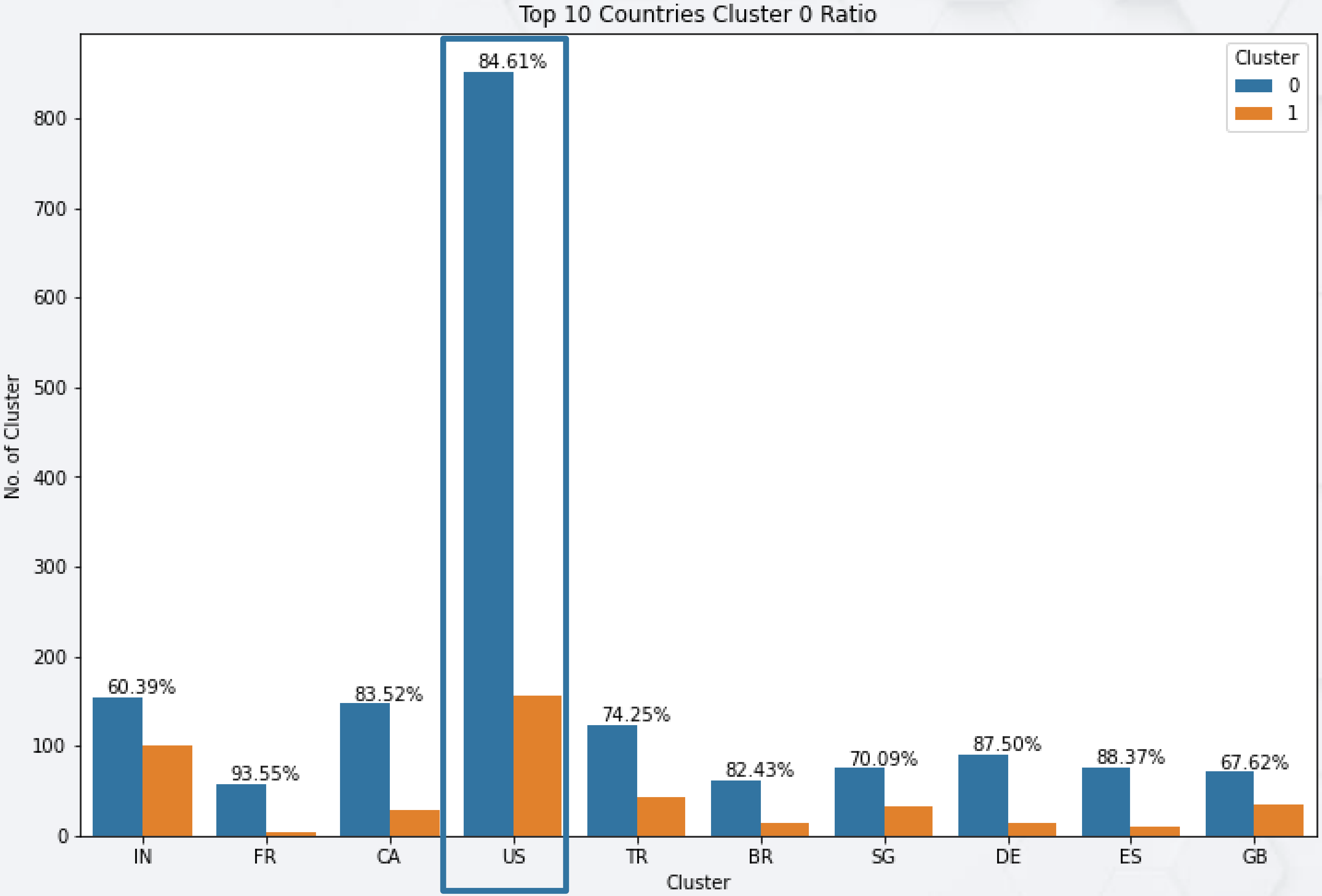
Exploratory Data Analysis

Attended Yes or No

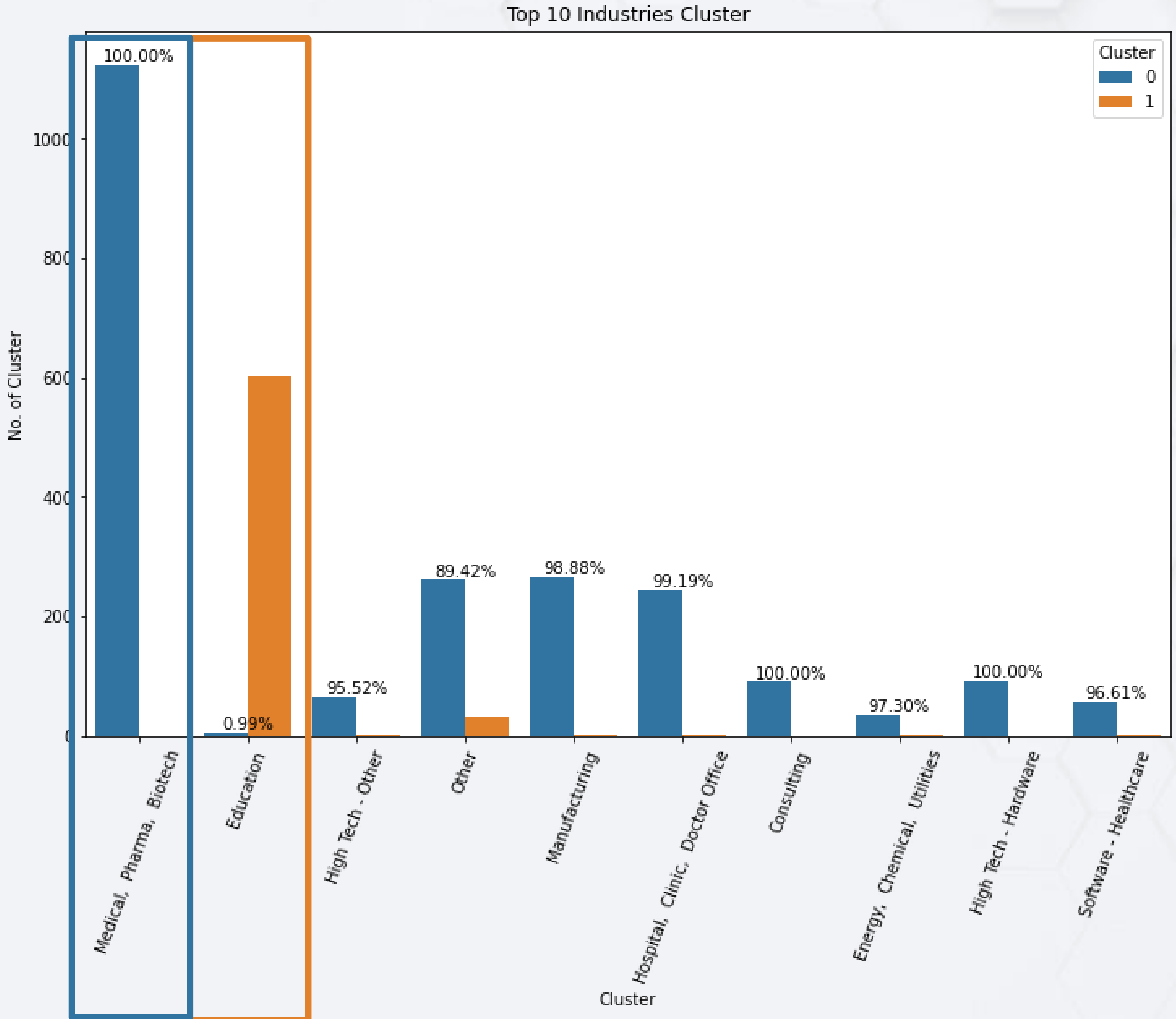


Cluster	Yes Rate
0	45.42
1	44.55

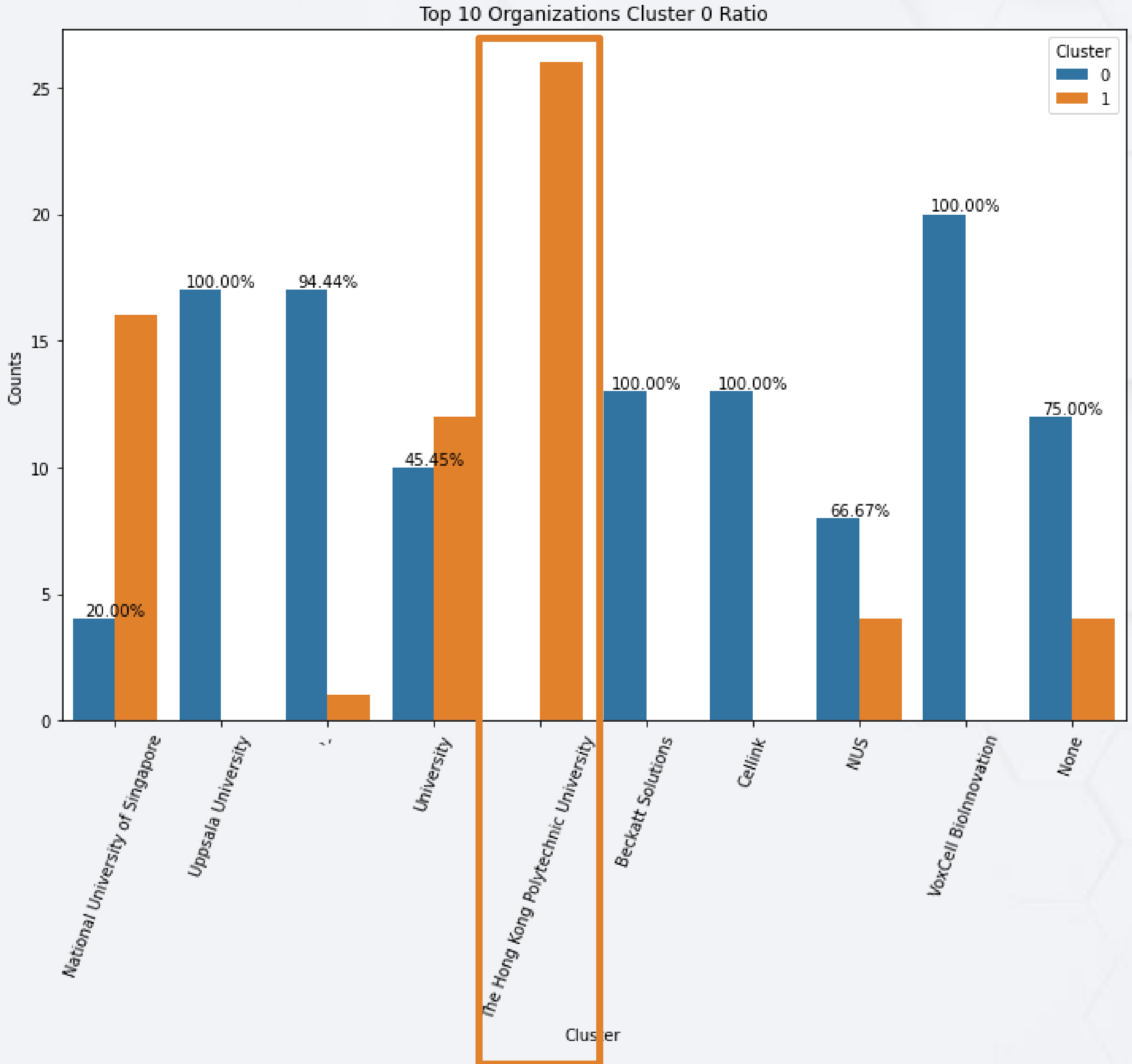
By Countries



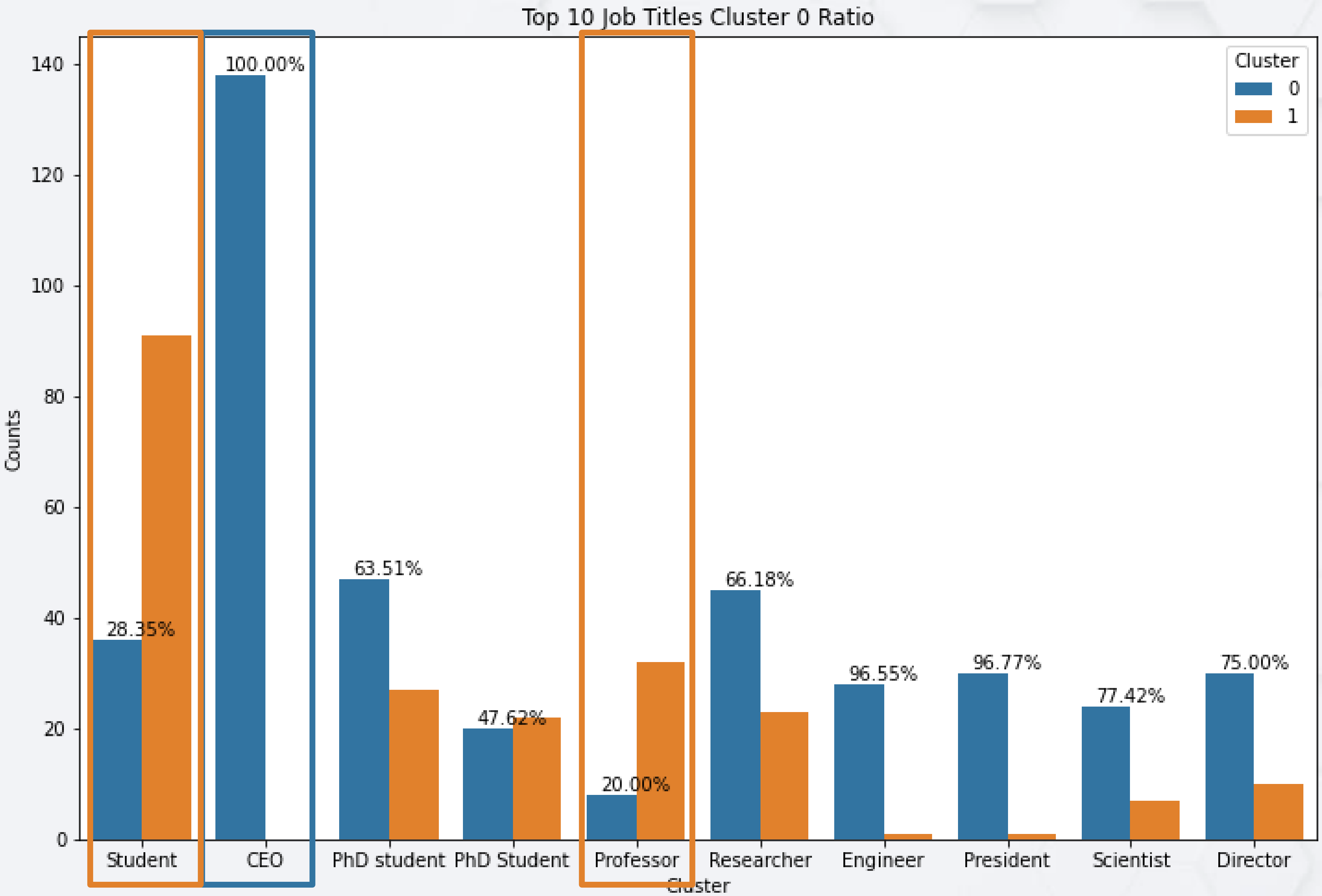
By Industries



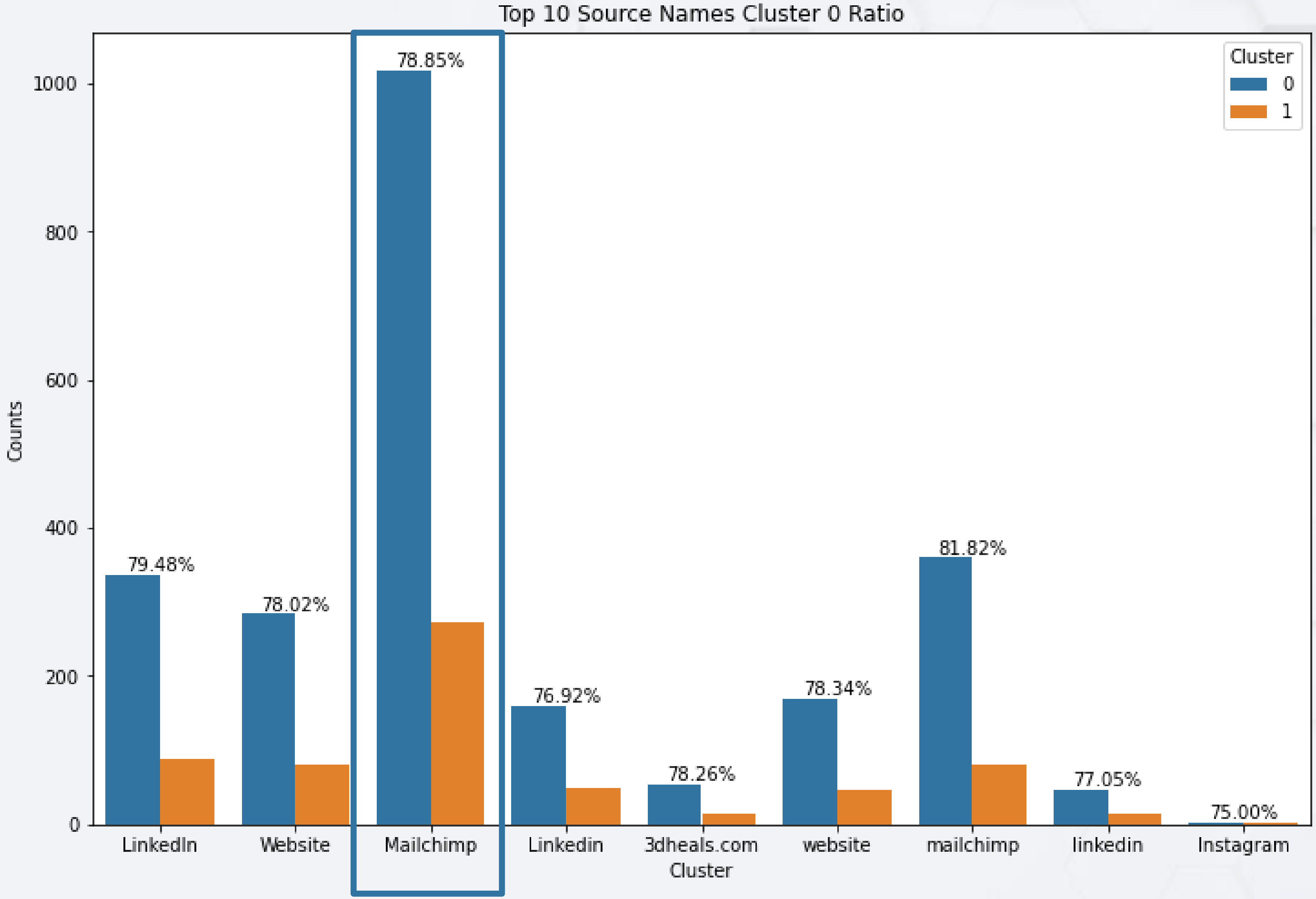
By Organizations



By Job Titles

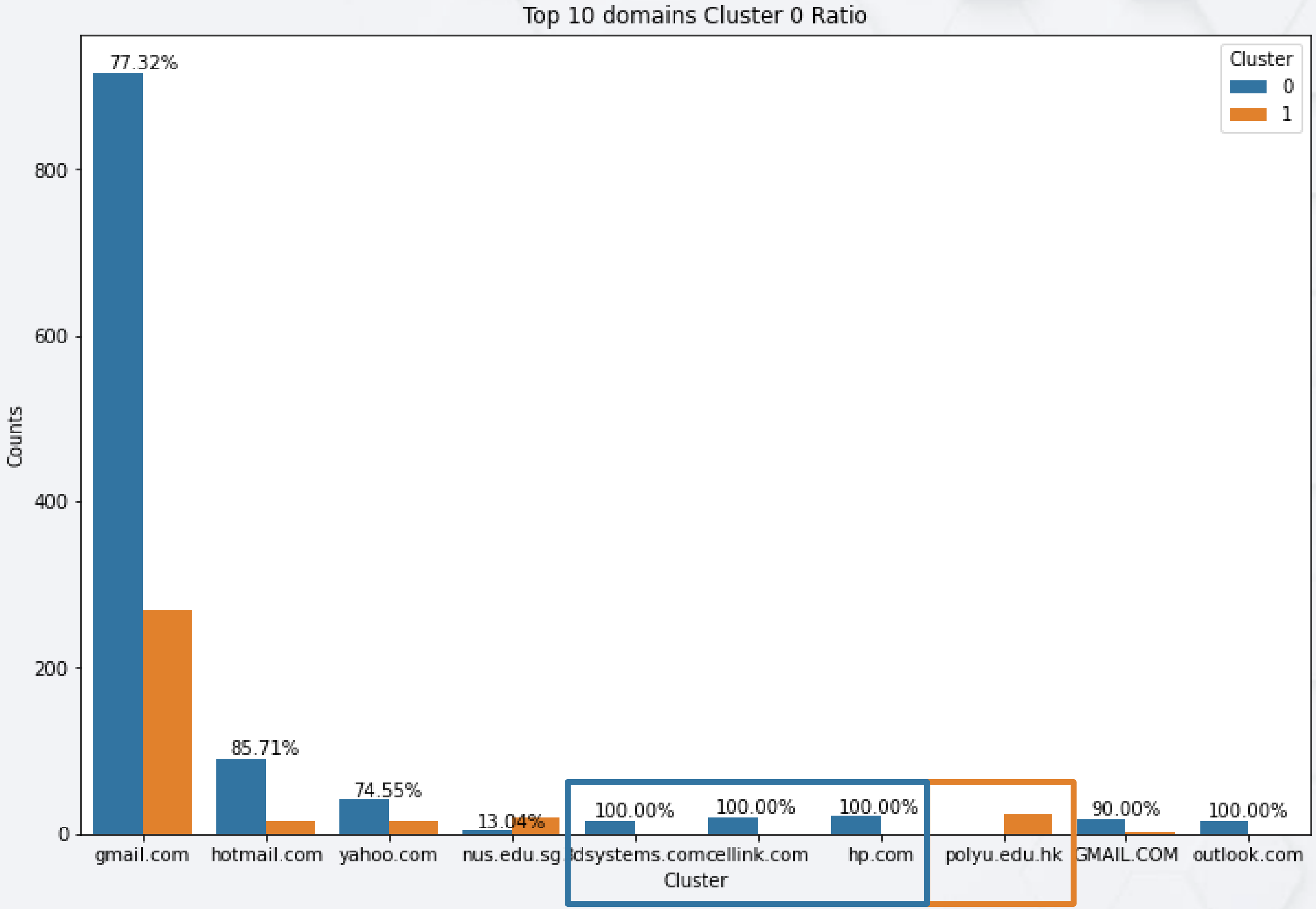


By Source Name

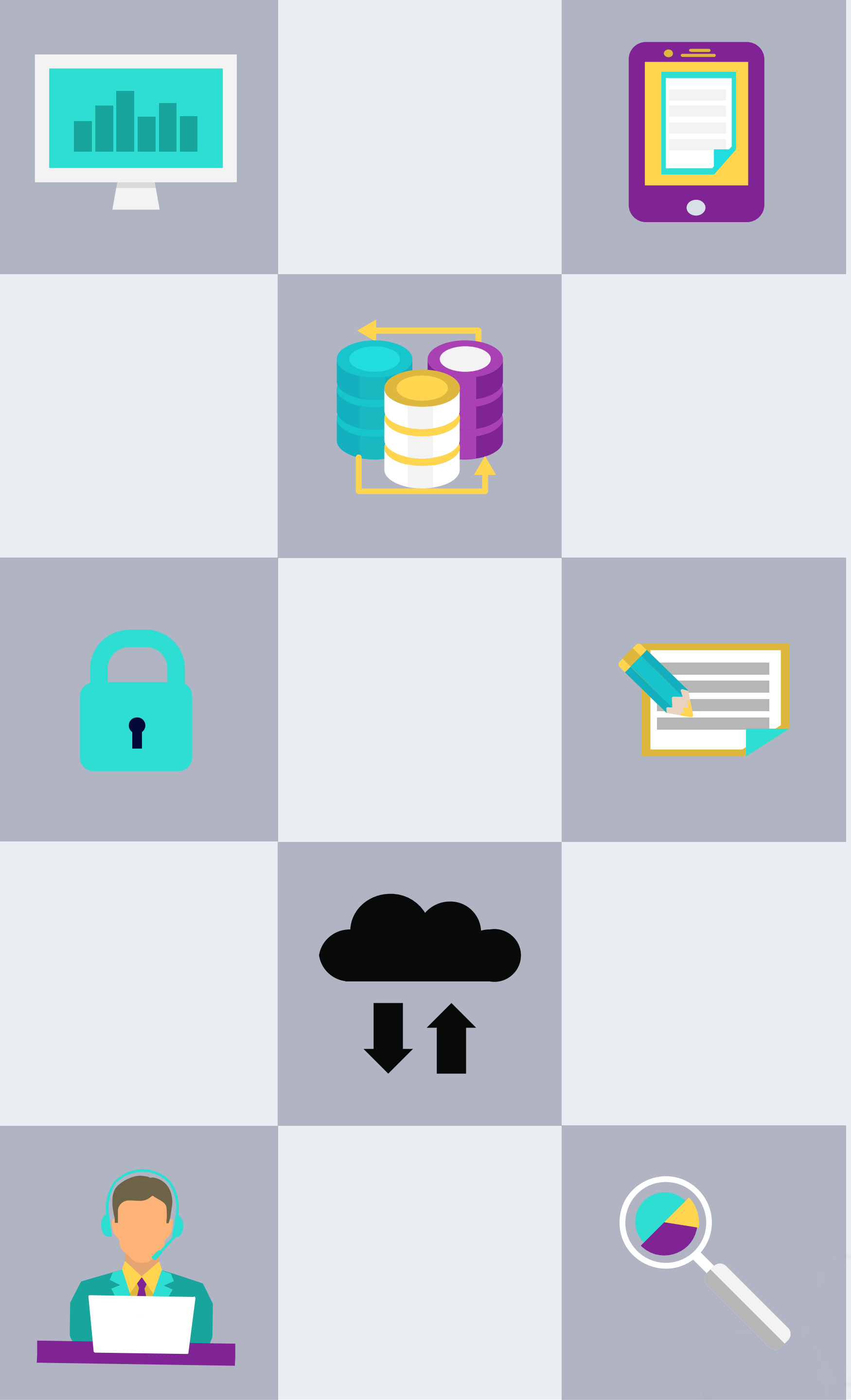


Northeastern University

By Domain_name



3DHEALS®

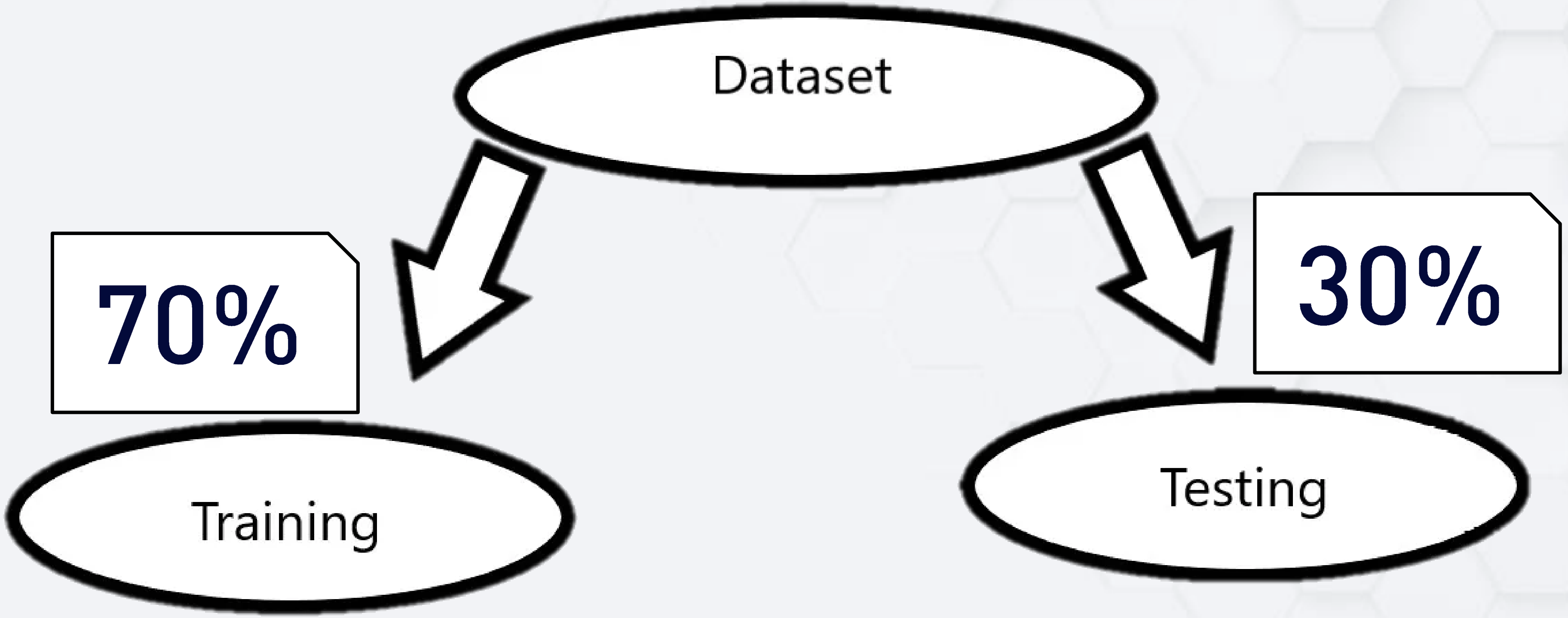


Predictive Modeling

04

For Customer Behavior

Split the train & test data



Dataset

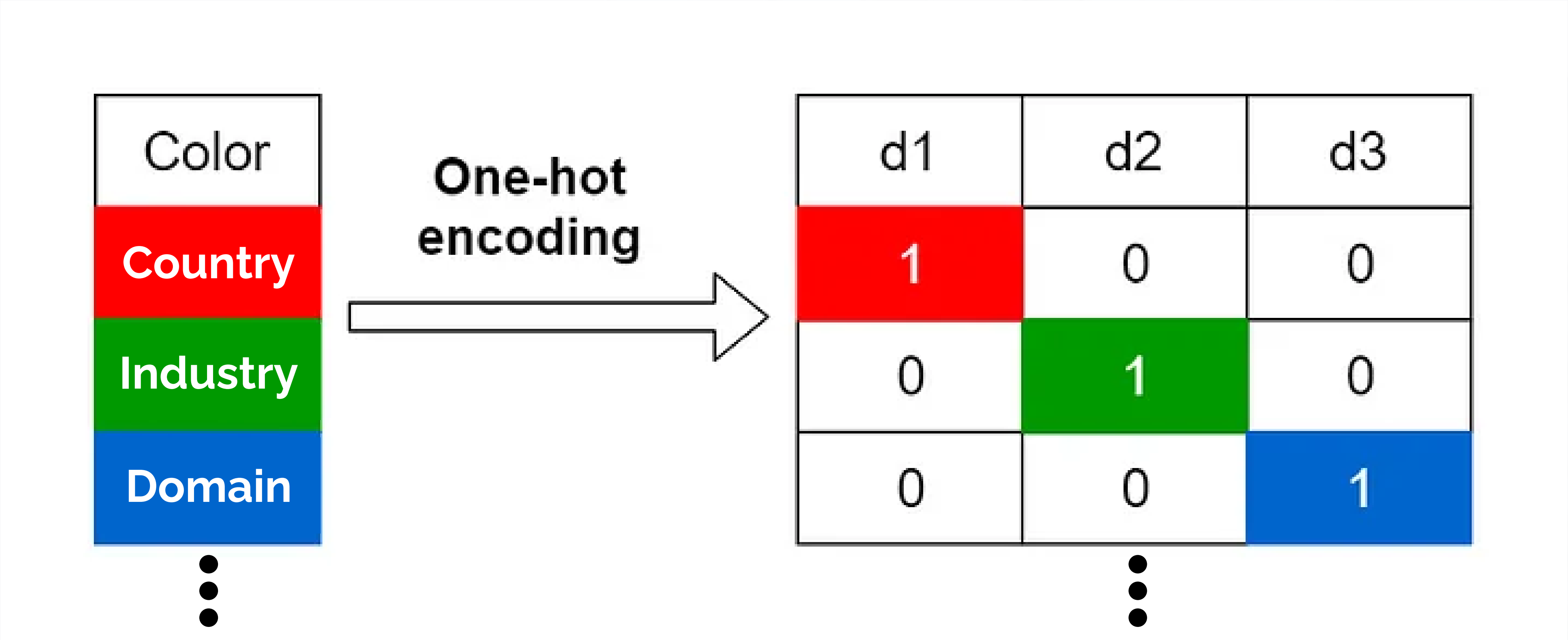
70%

Training

30%

Testing

One-Hot Encoding

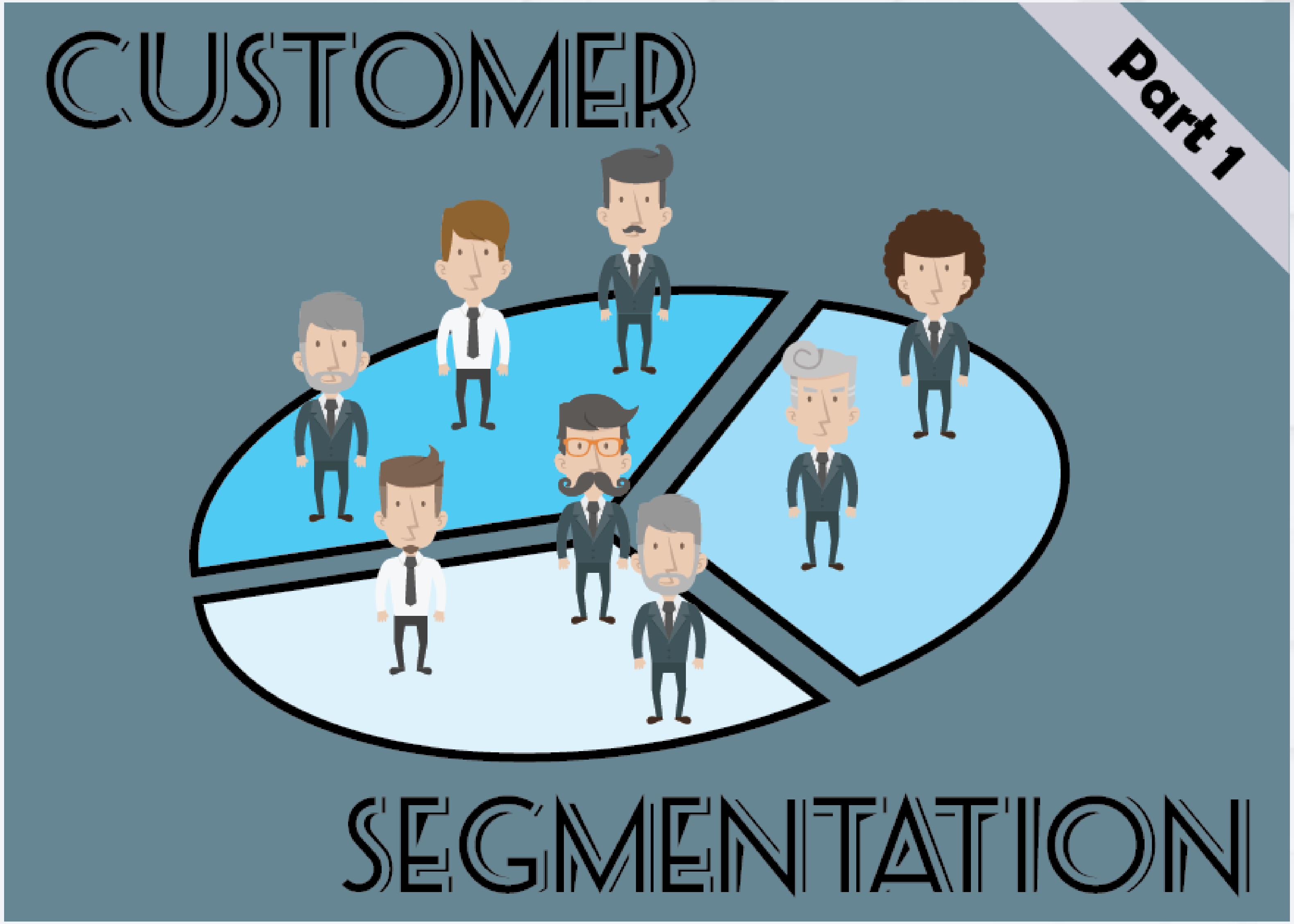


Color
Country
Industry
Domain
⋮

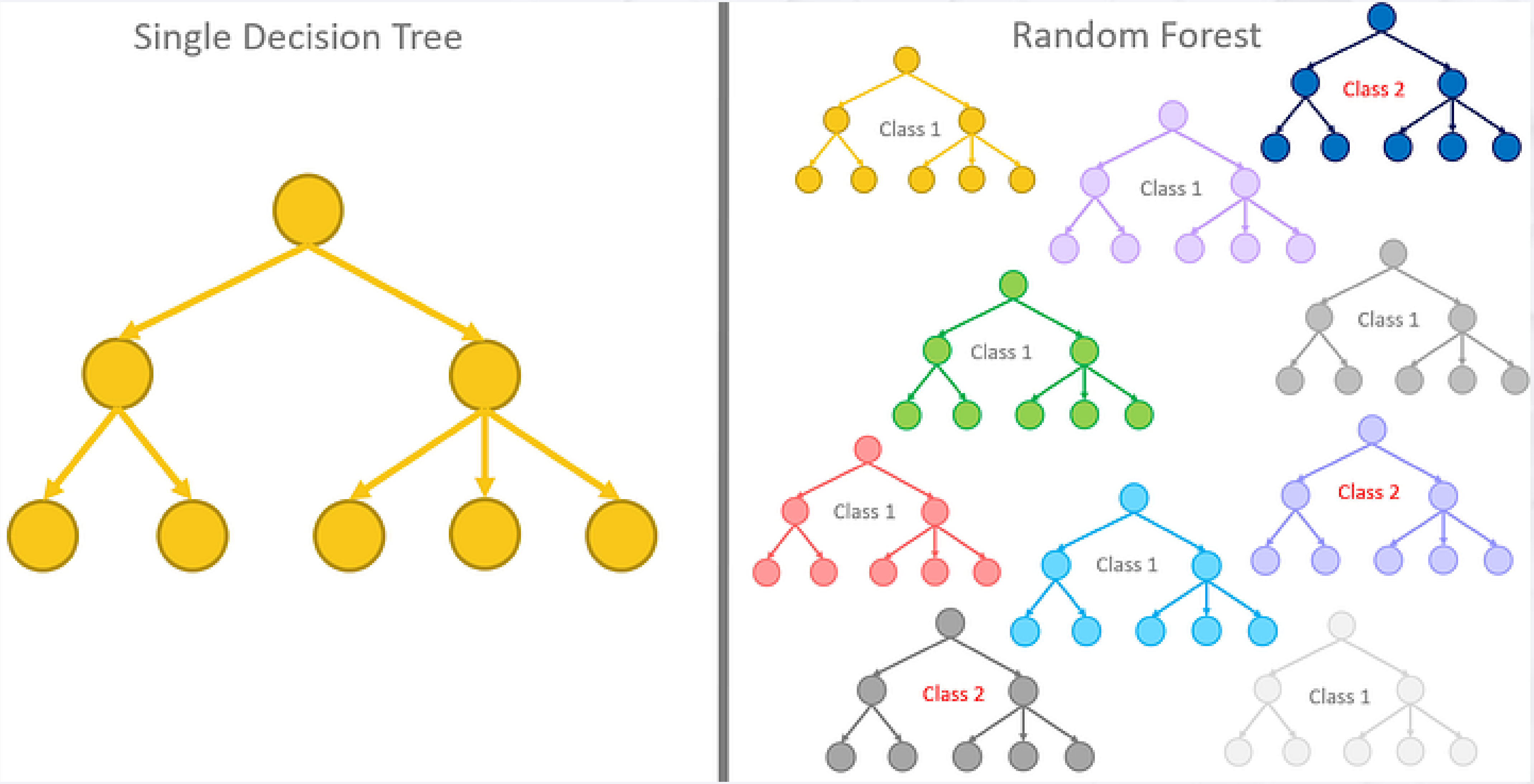
One-hot encoding

d1	d2	d3
1	0	0
0	1	0
0	0	1
	⋮	

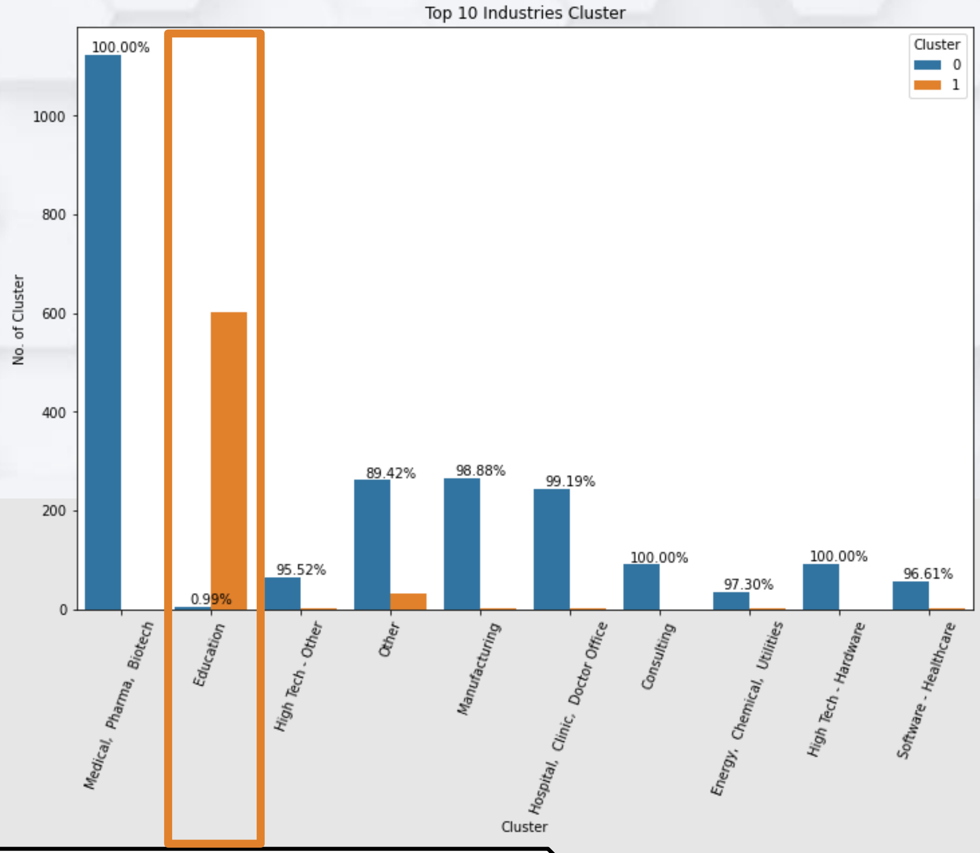
New Customer Clustering



Decision Tree & Random Forest



Decision Tree



Industry: Education

encoder__Industry_8 <= 0.5
gini = 0.308
samples = 2154
value = [1745, 409]

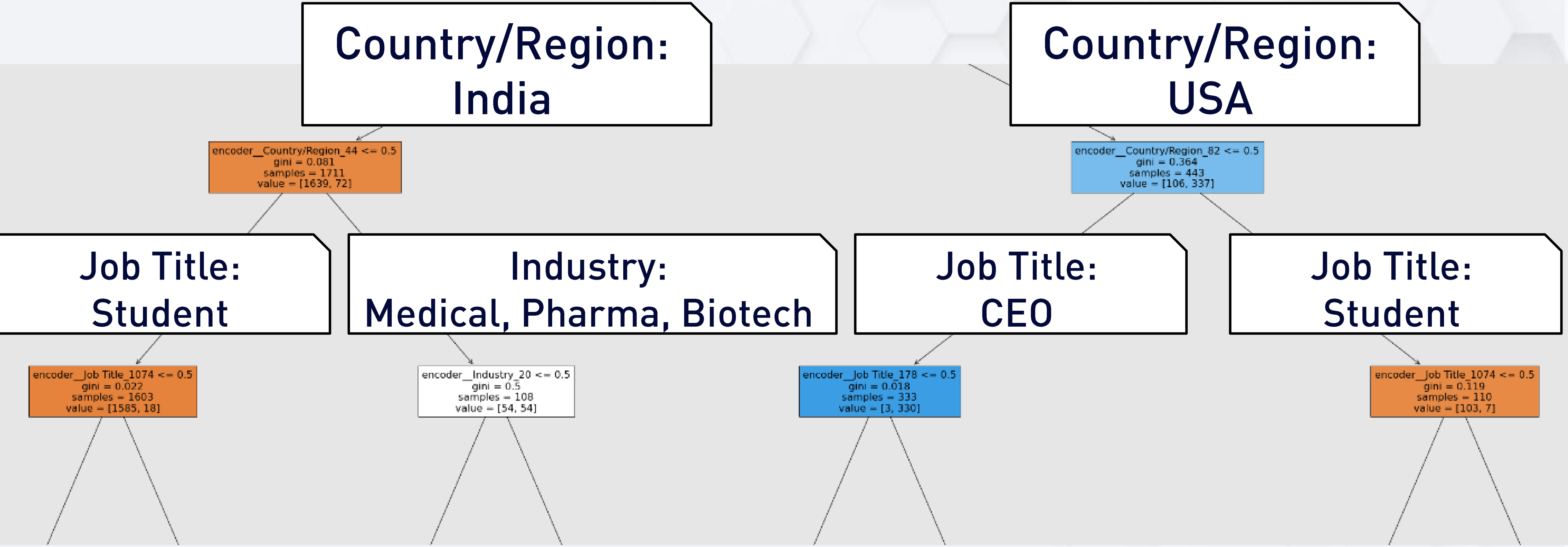
Country/Region: India

encoder__Country/Region_44 <= 0.5
gini = 0.081
samples = 1711
value = [1639, 72]

Country/Region: USA

encoder__Country/Region_82 <= 0.5
gini = 0.364
samples = 443
value = [106, 337]

Decision Tree



Country/Region: India

Country/Region: USA

Job Title: Student

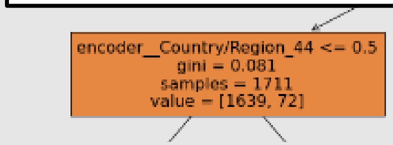
Industry: Medical, Pharma, Biotech

Job Title: CEO

Job Title: Student

Decision Tree

Country/Region: India

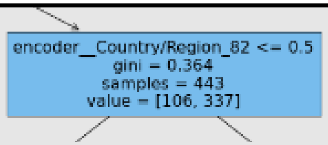


encoder__Country/Region_44 <= 0.5
gini = 0.081
samples = 1711
value = [1639, 72]

Cluster ‘1’ and Centroids

IN
India
Country/Region

Country/Region: USA



encoder__Country/Region_82 <= 0.5
gini = 0.364
samples = 443
value = [106, 337]

Cluster ‘0’ and Centroids

US
United States
Country/Region

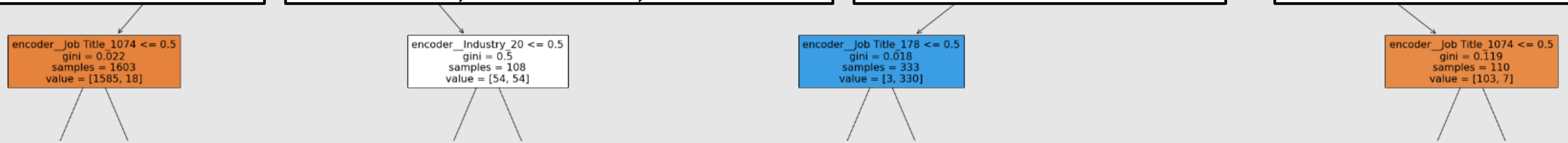
Decision Tree

Job Title: Student

Industry: Medical, Pharma, Biotech

Job Title: CEO

Job Title: Student

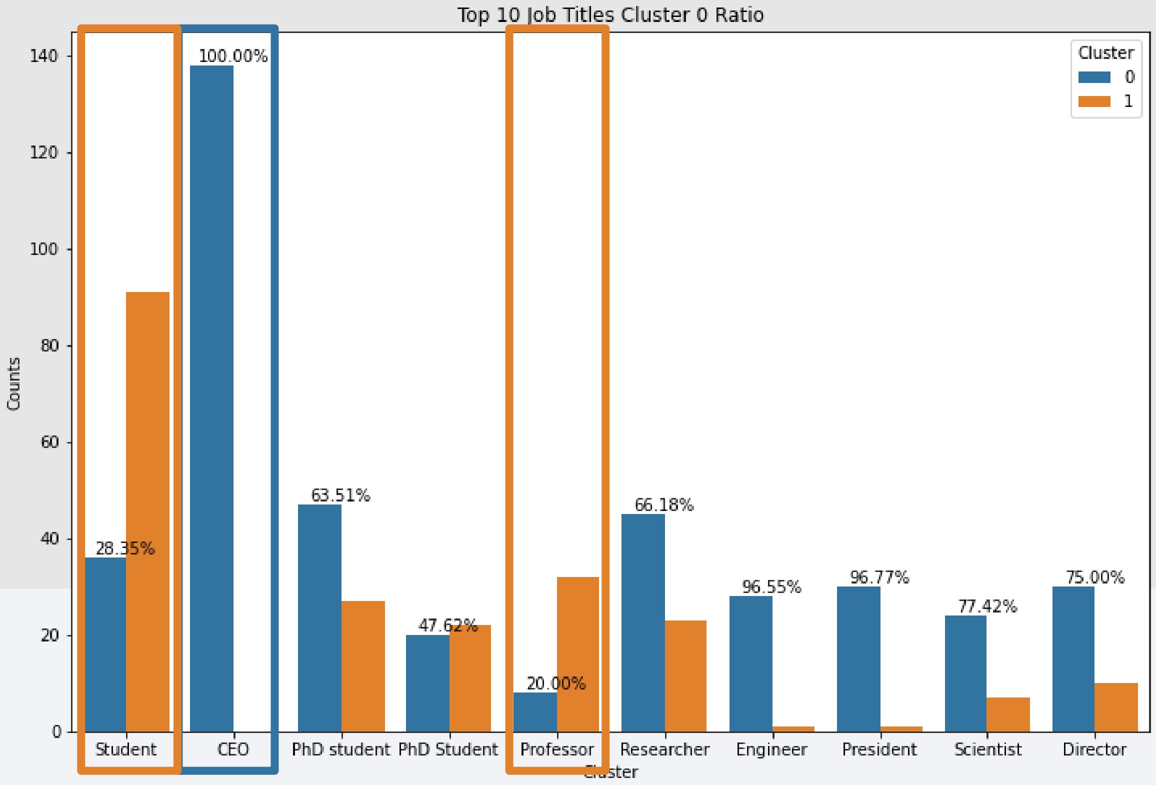


encoder__Job Title_1074 <= 0.5
gini = 0.022
samples = 1603
value = [1585, 18]

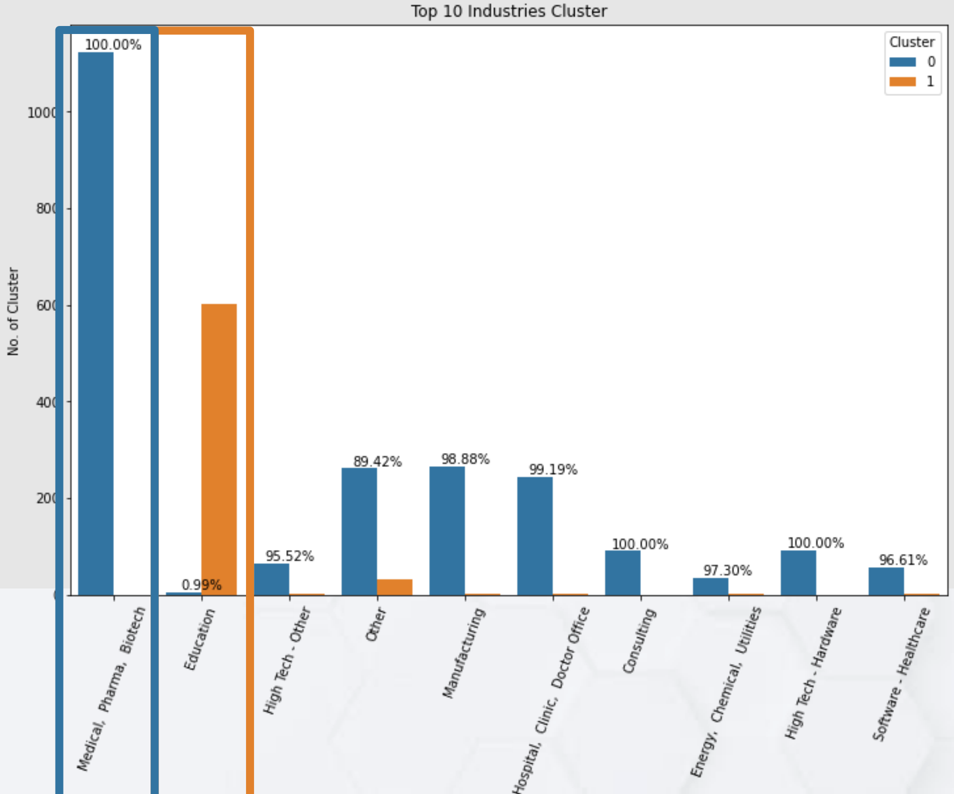
encoder__Industry_20 <= 0.5
gini = 0.5
samples = 108
value = [54, 54]

encoder__Job Title_178 <= 0.5
gini = 0.018
samples = 333
value = [3, 330]

encoder__Job Title_1074 <= 0.5
gini = 0.119
samples = 110
value = [103, 7]

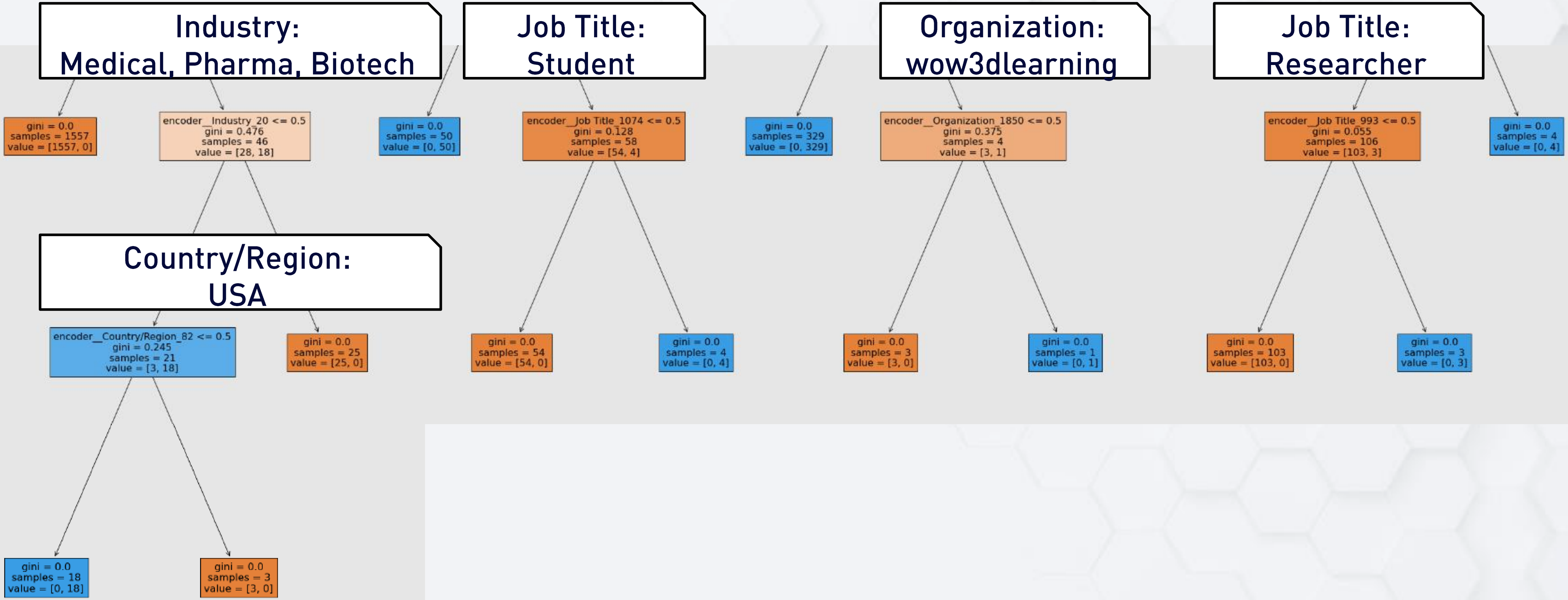


Top 10 Job Titles Cluster 0 Ratio



Top 10 Industries Cluster

Decision Tree



Industry: Medical, Pharma, Biotech

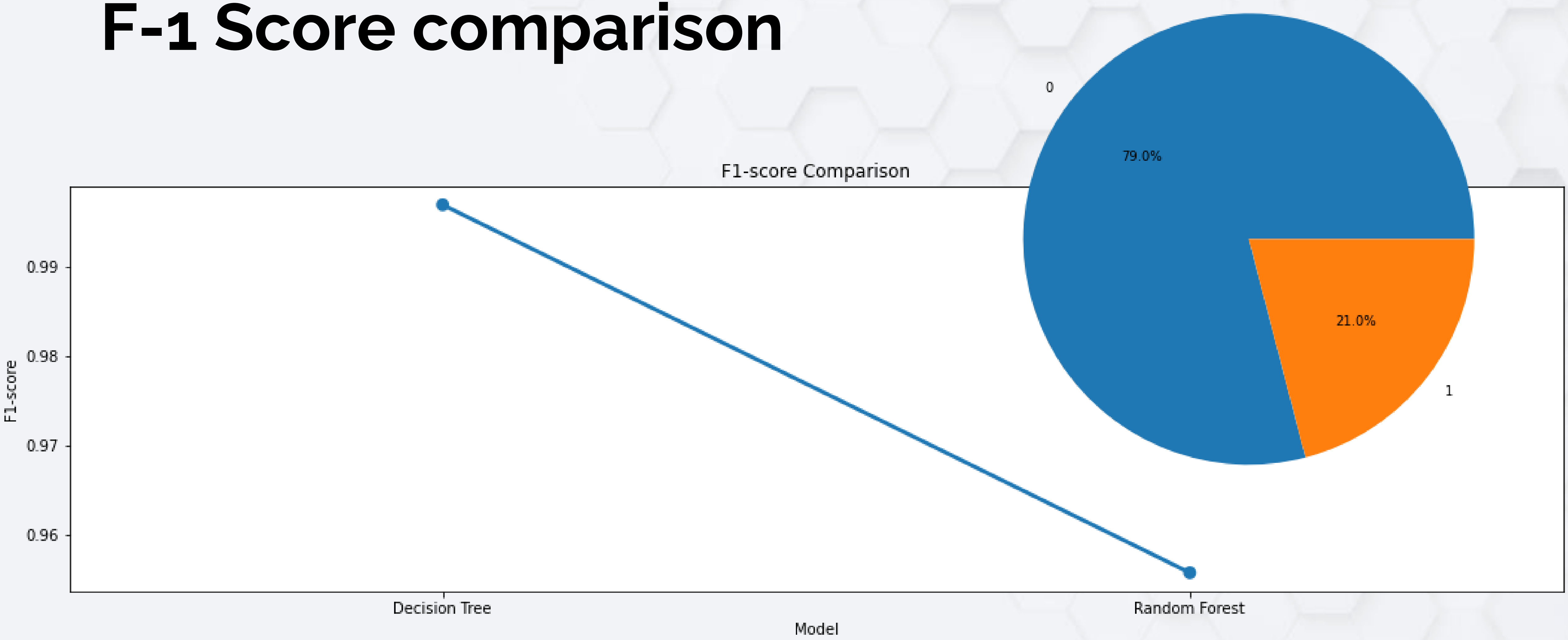
Job Title: Student

Organization: wow3dlearning

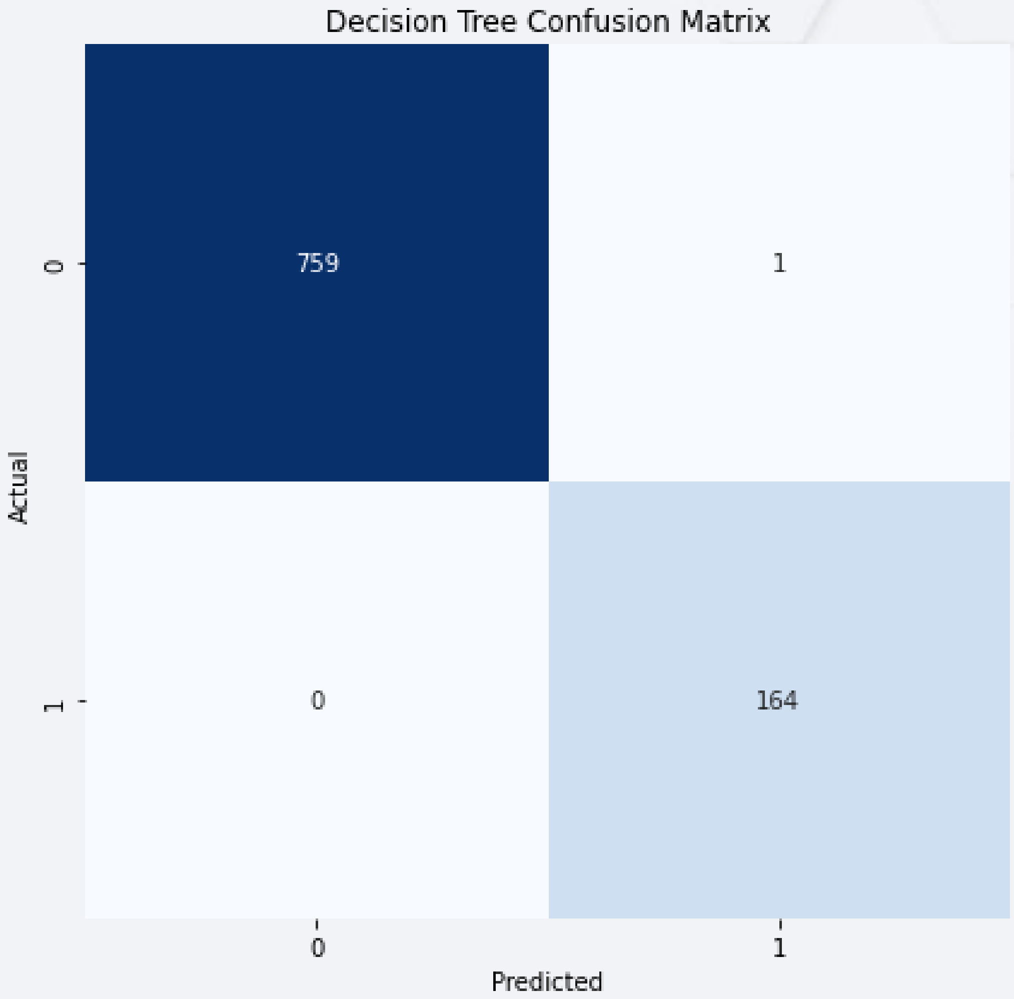
Job Title: Researcher

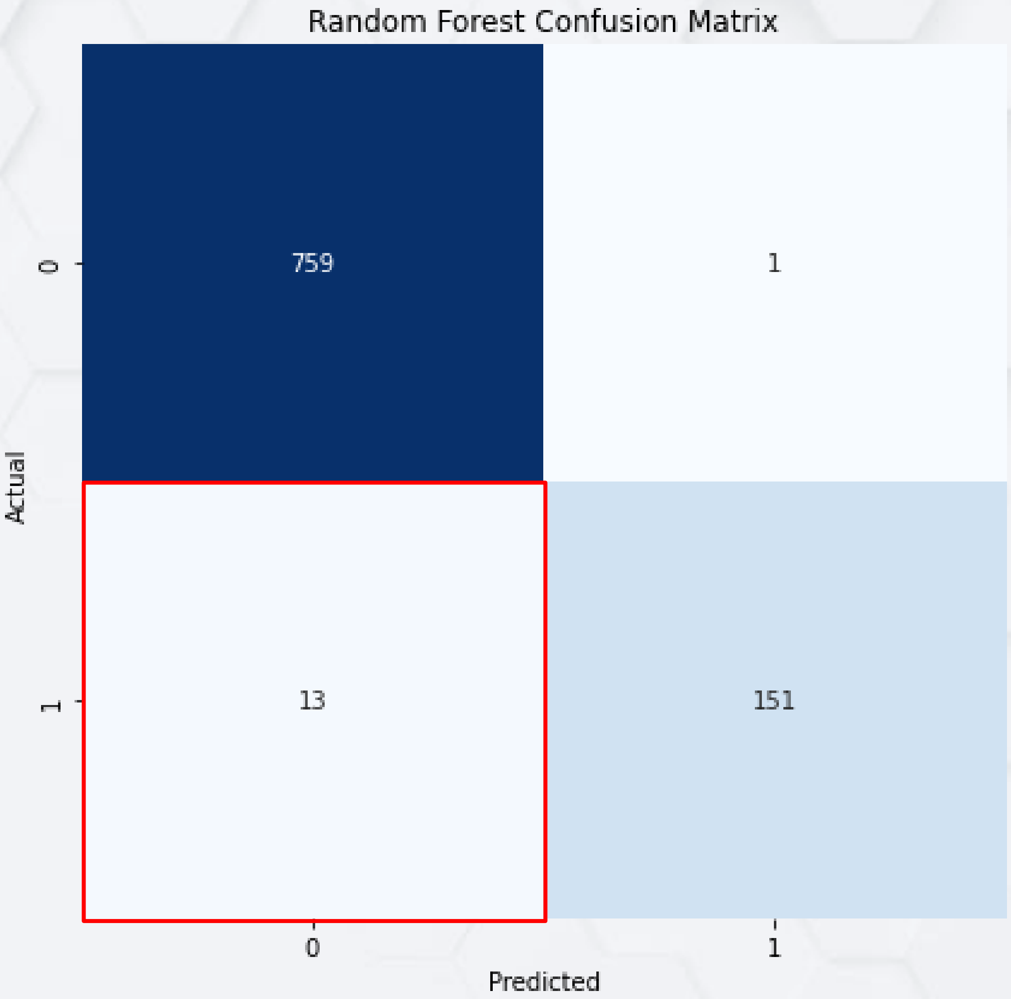
Country/Region: USA

F-1 Score comparison



Confusion Matrix





Recommendations & Future Plan

05

Next Steps

Recommendations

- **Customized Marketing:** Leverage attendee segmentation for designing personalized marketing strategies to increase engagement and event attendance.
- **Predictive Modelling:** Employ predictive models to forecast event attendance for efficient resource planning and management.
- **Event Timing and Content:** Evaluate session/webinar engagement to identify areas for content enhancement or timing adjustment.
- **Data Privacy Considerations:** Maintain respect for data privacy, exploring alternatives if attendees express discomfort with recording or info rmation sharing.
- **Sponsorship Collaboration:** Analyze sponsor influence on event attendance for potential strategic collaborations.
- **Geographic Targeting:** Consider focusing on regions with higher attendance rates for event hosting or increased marketing efforts.
- **Marketing Channels:** Assess the effectiveness of various marketing channels to concentrate efforts on the most impactful ones.
- **Feedback Analysis:** Apply natural language processing techniques to 'Questions & Comments' for valuable attendee sentiment analysis and direct feedback.

Centroids

Cluster ‘0’		Cluster ‘1’
US United States	Country/Region	IN India
Medical, Pharma, Biotech	Industry	Education
VoxCell BioInnovation	Organization	The Hong Kong Polytechnic University
CEO	Organization	Student
Mailchimp	Source Name	Mailchimp
Gmail.com	Domain	Poly.edu.hk

Future Plans

Sentiment Analysis: Using Natural Language Processing (NLP) techniques in the 'Questions & Comments' field might reveal valuable insights about the attendees' opinions and attitudes. This can help in improving the event experience further.

Predictive Modeling for Attendance: Build a predictive model using machine learning techniques to forecast an individual's attendance for future events based on their past records and engagement.

Network Analysis: If data on the relationships between attendees is available (such as connections on LinkedIn), network analysis could be used to identify influencers and key clusters within the attendee community.

Time Series Analysis: This can be used to understand the patterns in attendee registrations and dropout rates over time. This can help in better planning and predicting future event attendance.

Churn Analysis: Perform a churn analysis to identify individuals who have stopped attending the events. Understanding these individuals' characteristics can help devise strategies to re-engage them.

Q&A

Thank you!

3DHEALS®

REFERENCE

- 3DHEALS. (n.d.). About us. Retrieved from https://3dheals.com/about-us

- Deloitte. (2020). 2020 global health care outlook: Laying a foundation for the future. Retrieved from https://www2.deloitte.com/us/en/pages/life-sciences-and-health-care/articles/global-health-care-sector-outlook.html

- Machado, L. (2018, August 20). Artificial intelligence and 3D printing. Medium. https://medium.com/healthcare-3d-printing-stories/artificial-intelligence-and-3d-printing-94f45f3e45dd

- PwC. (2019). What's next for the pharmaceuticals industry, amid digital disruption and rapid technological advances? Retrieved from https://www.pwc.com/gx/en/pharma-life-sciences/pdf/pwc-pharma-2020.pdf

- Statista. (2021). 3D printing - Statistics & Facts. Retrieved from https://www.statista.com/topics/1174/3d-printing

- 3D Systems. (n.d.). Healthcare. Retrieved from https://www.3dsystems.com/healthcare

- 3DHEALS. (n.d.). About us. Retrieved from https://3dheals.com/about

- Gartner. (2018). Gartner says 3D printing is changing the landscape of the medical device market. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2018-04-03-gartner-says-3d-printing-is-changing-the-landscape-of-the-medical-device-market

- Huang, S. H., Liu, P., Mokasdar, A., & Hou, L. (2020). Additive manufacturing and its societal impact: A literature review. The International Journal of Advanced Manufacturing Technology, 67(5-8), 1191-1203. doi: 10.1007/s00170-012-4558-5

REFERENCE

- Organovo. (n.d.). About Organovo. Retrieved from https://organovo.com/about/Rengier, F., Mehndiratta, A., von Tengg-Kobligk, H., Zechmann, C. M., Unterhinninghofen, R.,Kauczor, H. U., & Giesel, F. L. (2010). 3D printing based on imaging data: Review of medical applications. International Journal of Computer Assisted Radiology and Surgery, 5(4), 335-341.doi: 10.1007/s11548-010-0476-x

- Stratasys. (n.d.). Medical solutions. Retrieved from https://www.stratasys.com/medical

- Suganya Karunamurthy. (2022). K-Mode Clustering | solved example | implementation. YouTube. Retrieved from https://www.youtube.com/watch?v=EVl2ejcsTfg

- Harika Bonthu. (2021, June 13). KModes Clustering algorithm for categorical data. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/#h-2-scree-plot-or-elbow-curve-to-find-optimal-kvalue

- PRASHANT BANERJEE. (2022). K-Means clustering with Python. Kaggle. Retrieved from https://www.kaggle.com/code/prashant111/k-means-clustering-with-python

- ASHISH. (2020). Bank customer clustering (K-Modes Clustering). Kaggle. Retrieved from https://www.kaggle.com/code/ashydv/bank-customer-clustering-k-modes-clustering/notebook

- rakshithvasudev. (2017, August 2). What is one hot encoding? why and when do you have to use it?. Hackernoon. Retrieved from https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f