ASHESI

Masters in Intelligent Computing Systems
Machine Learning and Data Science (September 2025)

Problem Situation 2 (PROSIT 2): Hidden Patterns in Student Journeys

The leadership team at Ashesi is grateful for your initial work in helping them see and understand the data more clearly. Your consultancy firm has made a strong impression, so much so they have decided to share semester-by-semester transcript data of all the students. You are now being asked to take the analysis one step further. Specifically, they want to discover any meaningful structure in the data.

The leadership believes that beneath the surface of the admissions and registry data, there may be hidden structures that could guide better decisions. For instance:

- Are there groups of students who share similar academic pathways in terms of course selections and GPA progression?
- Do certain clusters of students consistently hit bottlenecks in the curriculum?
- Are there outliers whose academic journeys look very different, perhaps signalling unique strengths or risk factors?
- Could visualizing student trajectories reveal emerging trends that aren't visible through raw numbers?

The challenge is that the dataset has no labels such as "at-risk" or "successful." The leadership team is used to hearing about supervised learning models that predict grades or retention, but in this case they want to explore unsupervised approaches that uncover patterns without predefined outcomes.

Your Deliverables

- Detailed notebook with model implementation, optimization, and evaluation
- Technical Report

Learning Outcomes

- Translate an open-ended business or research question from a stakeholder into a testable statistical/ML problem statement.
- Explain the role of unsupervised learning in the machine learning pipeline and contrast it with supervised learning.
- Distinguish among common unsupervised learning tasks (clustering, dimensionality reduction, anomaly detection, topic modeling).
- Define internal and external evaluation criteria for unsupervised learning, noting challenges in the absence of ground truth labels.
- Describe, implement, and compare clustering algorithms (k-means, hierarchical clustering, DBSCAN, Gaussian mixture models) on real-world datasets.
- Describe and apply dimensionality reduction methods (PCA, t-SNE, UMAP) to visualize high-dimensional data and support downstream tasks.
- Evaluate clustering performance using internal metrics (silhouette score, Davies–Bouldin index) and, where possible, external validation metrics (adjusted Rand index, normalized mutual information).
- Identify ethical issues in applying unsupervised learning.
- Communicate results to technical and non-technical stakeholders, emphasizing assumptions, limitations, uncertainty, and next-step recommendations

Resources

Textbooks

Anand, G., & Sharma, R. (2022). *Data science fundamentals and practical approaches.* BPB Publications.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

Gupta, P. (2022). *Practical data science with Jupyter.* BPB Publications.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning with applications in Python*. Springer.

Ozdemir, S. (2024). *Principles of data science: A beginner's guide to essential math and coding skills for data fluency and machine learning* (3rd ed.). Packt Publishing Ltd.

Jupyter Notebooks

Wine Clustering with Unsupervised Learning Models:
https://github.com/Ireanuoluwa/Wine-Clustering

Different Clustering Techniques and Algorithms:
https://www.kaggle.com/code/azminetoushikwasi/different-clustering-techniques-and-algorithms

Articles

Clustering with scikit-learn: A Tutorial on Unsupervised Learning:
https://www.kdnuggets.com/2023/05/clustering-scikitlearn-tutorial-unsupervised-learning.html

Unsupervised Clustering: Methods, Examples, and When to Use:
https://www.stratascratch.com/blog/unsupervised-clustering