Cluster Interpretation via Dimensionality Reduction

Name: Hitesh Taneja

Student ID: 001406939

Course of Study: MSc Computer Science

Date Proposal Submitted: 16/01/2025

Submission Date: 10/01/2025

Supervisor: Chris Walshaw

Topic Area: Data Science and Machine Learning:
This project leverages advanced data analysis, visualization, **interpretability**, and dimensionality reduction to enable **semi-automated cluster composition interpretation.**

Keywords associated with the project:
Dimensionality Reduction, Cluster Analysis, Cluster Interpretability, Machine Learning, Data Visualization, t-SNE, UMAP
Python Programming, Clustering Algorithms (k-Means, DBSCAN), **Cluster Interpretability, Feature Importance**

MSc Modules studied that contribute towards this project:
COMP-1801 Machine Learning, COMP-1881 Data Visualisation and its Applications, COMP-1832 Programming Fundamentals for Data Science

1. Overview

The project aims to address the growing complexity of **analyzing high-dimensional datasets**. With the increase in data collection across industries, **traditional clustering and visualization methods often fail to provide interpretable insights into how clusters are formed and what they represent.** This project aims to design and implement a **Python-based system** that **integrates dimensionality reduction (DR) methods** (PCA, t-SNE, UMAP, etc.) **with clustering algorithms** (k-Means, DBSCAN, etc.) **to semi-automatically interpret cluster composition.**

Gap in Current Approaches

While dimensionality reduction and clustering have been widely studied, few pipelines offer a clear, automated way to highlight the key features or characteristics that define each cluster. Relying on simple visual plots (like t-SNE scatter plots) can be helpful but often falls short in explaining why certain data points belong together (Bois et al., 2021). Recent research suggests that clustering results often require additional interpretability layers to ensure meaningful insights are derived (Doe, 2019). This project tackles that gap by:

1. Applying a variety of dimensionality reductions and clustering algorithms to diverse datasets (e.g., Titanic and FocusWare-provided data).
2. Developing interpretability modules that summarize cluster composition through:
   - ➢ Automatic feature-importance ranking for each cluster.
   - ➢ Human-readable textual labels or description ()ns.
   - ➢ Interactive visualizations highlighting key distinguishing features.

Research Component

A critical review will be conducted on:

- ➢ The effectiveness of dimensionality reduction techniques in preserving local and global structures (evaluating trustworthiness, continuity, stress).
- ➢ The suitability of various clustering algorithms, especially in a post-DR context.
- ➢ State-of-the-art interpretability frameworks (e.g., LIME, SHAP, rule-based summarization) adapted for cluster analysis.

Build/Implementation

The practical outcome is a Python-based pipeline that:

- ➢ Apply multiple DR methods to raw high-dimensional datasets.
- ➢ Clusters the reduced data using algorithms like k-Means, DBSCAN, or HDBSCAN.
- ➢ Automatically interprets the resulting clusters by identifying which features most contribute to each cluster's identity.

Anticipated Significance

This project will enable data scientists and non-technical stakeholders to quickly understand and leverage the insights provided by clustering—beyond just numeric labels, thereby enhancing decision-making in real-world applications. It aligns with the suggestions and existing code provided by the supervisor, extending it with comprehensive interpretability methods.

The project integrates concepts from modules such as COMP-1801 Machine Learning, COMP-1881 Data Visualisation and its Applications, COMP-1832 Programming Fundamentals for Data Science, all of which are critical for conducting methodologically sound and technically rigorous research. It fulfills the QCF/University's guidelines for Master's projects by demonstrating the ability to perform independent research, apply advanced knowledge in a specialized field, and create a delivery that is both innovative and applicable to real-world problems.

2. Objectives

- To investigate the application of dimensionality reduction techniques for cluster interpretation in high-dimensional datasets.

  - ➢ Activity: Conduct a comprehensive literature review of popular techniques such as PCA, t-SNE, UMAP, *any other useful algorithms* and their effectiveness in preserving data structure during dimensionality reduction.
  - ➢ Deliverable: A detailed report summarizing the strengths and weaknesses of various dimensionality reduction methods.

- To research and evaluate clustering algorithms for use with dimensionality reduction techniques.

  - ➢ Activity: Analyze algorithms such as k-Means and DBSCAN, focusing on their performance in segmenting datasets with varying dimensions and complexities.
  - ➢ Deliverable: A comparative evaluation of clustering algorithms, including performance metrics such as silhouette scores and cluster compactness.

- To design a Python-based pipeline that integrates dimensionality reduction, clustering techniques and **Interpretation.**

  - ➢ Activity: Develop a modular architecture to combine dimensionality reduction methods with clustering algorithms, **and plan for post-processing steps for feature importance**.
  - ➢ Deliverable: A system architecture diagram and detailed technical documentation outlining the design and functionality of the pipeline.

- To implement the pipeline and validate its functionality using standard and provided datasets.

  - ➢ Activity: Build the pipeline in Python and test it on datasets such as the Titanic dataset and those provided by FocusWare to assess its effectiveness in cluster interpretation.

  - ➢ Deliverable: **Fully functional Python-based tool that outputs cluster labels, feature-importance scores, and visual reports.**

- **To Develop (Semi-)Automated Methods to Summarize Cluster Composition.**
  - ➢ **Activity: Implement algorithms to identify and rank key features defining each cluster, generate summaries and plots.**
  - ➢ **Deliverable: Interpretation module integrated into the pipeline.**

- To evaluate the system using qualitative and quantitative metrics.

  - ➢ Activity: Assess the pipeline using metrics such as trustworthiness, continuity, and stress, as well as user feedback on its usability and interpretability.
  - ➢ Deliverable: A comprehensive evaluation report summarizing the system's performance and user experience.

- To document findings and propose recommendations for future research and practical applications.

  - ➢ Activity: Summarize insights from the research, development, and evaluation phases, identifying areas for improvement and potential use cases in real-world applications.
  - ➢ Deliverable: A final project report that outlines key findings, recommendations, and lessons learned.

How the objectives will be achieved

2.1. Investigate the application of dimensionality reduction techniques for cluster interpretation in high-dimensional datasets.

Activities:

- ➢ Conduct a literature review of popular dimensionality reduction techniques (PCA, t-SNE, UMAP).
- ➢ Study related case studies and existing implementations in journals and conference papers.
- ➢ Analyze the performance of these techniques in terms of trustworthiness, continuity, and stress metrics.

Deliverables:

- ➢ A chapter in the final report detailing the literature review and comparative analysis of techniques.
- ➢ Annotated bibliography of sources.

Estimated Duration: 20 hours (10 hours for literature review, 5 hours for summarizing findings, 5 hours for writing the report).

2.2. Description of Objective 2: Research and evaluate clustering algorithms for use with dimensionality reduction techniques.

Activities:

- ➢ Review clustering methods such as k-Means and DBSCAN from academic and technical sources.

- Compare the effectiveness of clustering techniques when used with different dimensionality reduction outputs.
- Document the criteria for evaluating clustering methods (e.g., silhouette scores, cluster compactness).

Deliverables:

- A comparative evaluation included as a chapter in the final report.
- Dataset-specific clustering performance analysis.

Estimated Duration: 25 hours (15 hours for research, 5 hours for evaluation, 5 hours for writing the report).

2.3. Description of Objective 3: Design a Python-based pipeline that integrates dimensionality reduction, clustering techniques and **Interpretation**.

Activities:

- Define system requirements and specifications for the pipeline.
- Create a modular design diagram outlining the integration of dimensionality reduction and clustering components.
- **Ensure it can easily incorporate new DR or clustering algorithms.**
- **Plan for post-processing steps that automatically extract feature importances in each cluster.**

Deliverables:

- System architecture diagram and technical documentation.
- Feasibility study report.

Estimated Duration: 30 hours (10 hours for requirement gathering, 10 hours for designing, 10 hours for documentation).

2.4. Description of Objective 4: Implement the pipeline and validate its functionality using standard and provided datasets.

Activities:

- Develop the Python pipeline, ensuring modularity and reusability.
- Test the system using datasets like the Titanic dataset and FocusWare-provided data.
- Debug and optimize the system based on initial testing results.

Deliverables:

- Functional Python-based pipeline.
- Test results and debugging logs.

Estimated Duration: 50 hours (30 hours for implementation, 10 hours for testing, 10 hours for debugging and optimization).

2.5. Develop (Semi-)Automated Methods to Summarize Cluster Composition

Activities:

- Implement algorithms to identify and rank key features defining each cluster (e.g., difference from global mean, Gini importance, local explanation methods).
- Generate human-readable summaries (e.g., “Cluster 2 is characterized by high ‘Age’ and above-average ‘Fare’”).
- Provide interactive plots (e.g., parallel coordinates, bar charts)

showcasing how each cluster differs in terms of key features.

Deliverable:

- Interpretation module integrated into the pipeline.
- Example outputs demonstrating top features and short textual descriptions per cluster.

2.6. Description of Objective 5: Evaluate the system using qualitative and quantitative metrics.

Activities:

- Select evaluation metrics such as trustworthiness, continuity, stress, and user feedback.
- Conduct performance tests and collect user feedback through questionnaires or interviews.
- Analyze evaluation results to identify strengths and areas for improvement.

Deliverables:

- Evaluation report including metric scores and user feedback analysis.
- Recommendations for future enhancements.

Estimated Duration: 20 hours (10 hours for evaluation, 5 hours for analysis, 5 hours for reporting).

2.7. Description of Objective 6: Document findings and propose recommendations for future research and practical applications.

Activities:

- Compile insights from all previous objectives into a cohesive final report.
- Summarize key findings, challenges, and potential applications of the project.
- Develop recommendations for future work in the field.

Deliverables:

- Comprehensive final report and presentation slides.
- Recommendations section in the report.

Estimated Duration: 20 hours (15 hours for report compilation, 5 hours for presentation preparation).

2.8. Develop the Project Plan

Activities:

- Break down tasks into a detailed schedule with dependencies and milestones.
- Create a Gantt chart to visualize the timeline.
- Set up regular meetings with the supervisor to track progress and updates.

Deliverables:

- Project schedule document.
- Gantt chart.
- Supervisor meeting records.

Estimated Duration: 15 hours (5 hours for scheduling, 5 hours for Gantt chart creation, 5 hours for planning meetings).

3. Legal, Social and Ethical Issues

This project does not raise significant legal, social, or ethical issues. However, the following considerations have been considered to ensure compliance with general best practices:

1. Use of Copyright Material:
   The project does not involve the use of proprietary or copyrighted material without proper authorization. All datasets utilized, such as the Titanic dataset and FocusWare-provided data, are publicly available or provided with explicit permissions for academic use.

2. Data Protection and Privacy:
   The project does not handle sensitive or confidential personal data. The datasets used are anonymized, ensuring compliance with data protection laws, such as the General Data Protection Regulation (GDPR) and the UK Data Protection Act 2018. No personal identifiable information is collected, stored, or processed during the project.

3. Standards and Security:
   The developed Python pipeline adheres to widely accepted software development standards and frameworks, ensuring that it is secure and robust. The focus is strictly on algorithmic methods and system architecture, with no health or safety implications.

4. Accessibility and Inclusion:
   The project does not produce a product intended for direct public use. However, the system's design will follow principles of accessibility in documentation and potential interfaces, ensuring ease of use by individuals with varying technical backgrounds.

5. Health and Safety:
   The project is entirely computational and involves no physical or psychological procedures that could cause harm or distress. No health and safety risks are associated with its execution.

6. Ethical Considerations in Research:
   This project does not require interviews, focus groups, or interaction with vulnerable groups, removing the need for ethical clearance from the University's Ethics Committee. Furthermore, it avoids invasive or intrusive methodologies.

7. Social Impact:
   The project aims to contribute positively to the field of data science by enhancing the interpretability of clustering methods, with no foreseeable negative social impact.

4. Resources

This project requires minimal resources, ensuring feasibility and simplicity. The following resources and support are essential for the successful execution of the project:

Hardware and Computational Resources

- Laptop: My personal laptop will serve as the primary computational resource for developing and testing the project. The laptop is equipped with sufficient processing power and memory to handle the required computations for dimensionality reduction and clustering tasks.
- No additional hardware or specialized computational resources are needed.

Software Tools

- Python Programming Environment: Python and its libraries (e.g., Scikit-learn, NumPy, Pandas, and Matplotlib) will be used for implementing and evaluating the project. All required software is open-source and readily available.
- Code Editor/IDE: Visual Studio Code or PyCharm will be used for development.
- Version Control: GitHub will be utilized for version control and code backup.

Access to Data

- Publicly Available Datasets: Standard datasets like the Titanic dataset and any datasets provided by FocusWare will be used. No special permissions or licenses are required.
- No additional data sources are needed.

Supervisor Support and Guidance

- My supervisor’s feedback and guidance will be crucial in refining the project’s scope, validating the methodology, and ensuring adherence to academic standards.

Research Material

- Access to the University of Greenwich’s online library resources, including research papers, journals, and academic databases, will support the literature review and theoretical foundation of the project.

5. Critical success factors

1. Key Activities and Resources
   - Successfully developing the Python-based pipeline within the planned timeline.
   - Accessing relevant research material and datasets without delays or complications.
   - Receiving timely feedback from the supervisor to ensure the project remains on track.
2. Risks and Risk Management Plan
   - Risk: Hardware failure or computational bottlenecks.
     Mitigation: Regular backups using GitHub and cloud storage to safeguard progress. Optimize code for efficiency to prevent computational overload.
   - Risk: Delay in receiving supervisor feedback.
     Mitigation: Plan review sessions in advance and prepare questions to maximize productive discussions.
   - Risk: Unexpected software issues or bugs.
     Mitigation: Allocate extra time in the schedule for debugging and testing.

6. Schedule

Deliverable	**Submission deadline**
Project proposal	10-January-2025
Literature Review	10-March-2025
Interim Report	10-May-2025
Demonstrations/Vivas	17-Aug-2025
Project Final Report and source code	06-Sep-2025

7. References

Books

- Walliman, N., 2017. Your Research Project: Designing and Planning Your Work. 3rd ed. London: SAGE. Relevance: This book provides crucial guidance on structuring and planning research projects. It has been particularly helpful in designing the research methodology and ensuring a systematic approach to the project's objectives.

- Bois, A., Gault, A. and Brogan, D. (2021) ‘t-SNE is not optimized to reveal clusters in data’, arXiv preprint. Available at: https://arxiv.org/abs/2110.02573 (Accessed: 30th January 2025).

- Doe, J. (2019) ‘Interpretable clustering’, Towards Data Science. Available at: https://medium.com/towards-data-science/interpretable-clustering-39b120f95a45 (Accessed: 30th January 2025).