**K-Means Clustering with Process Event Logs**

**Lab Report**

**1. Introduction and Purpose**

Process mining is a technique that enables the analysis of business processes based on event logs. K-Means clustering, when applied to process event logs, allows for the identification of similar process execution patterns, helping to simplify complex process models and discover behavioral patterns.

The purpose of this lab was to apply the K-Means clustering algorithm to process event logs from DAP for job applications. By extracting numerical features from case behaviors (such as duration, number of activities, and activity transitions), we aimed to:

* Discover groups of cases that follow similar process flows
* Identify common behaviors and deviations in process execution
* Simplify the complexity of the overall process model by analyzing each cluster separately
* Gain insights that could lead to process improvements

**2. Dataset Overview**

The dataset consists of event logs from a job application process. Each event represents an activity performed within a specific case (application), including:

* **CaseID**: Unique identifier for each application process
* **Activity**: The specific task being performed
* **InitialStatus** and **FinalStatus**: States before and after the activity
* **ProcessFlow**: Description of the activity's role in the process
* **Timestamp**: When the activity occurred

In total, the dataset contains 8 cases (applications), with various activities such as portal access, document preparation, application submission, document verification, and selection processes.

**3. Methodology**

**3.1 Feature Extraction**

To apply K-Means clustering, we first needed to transform the event logs into case-level features. For each application case, we extracted the following features:

1. **Duration** (in days): Total time from the first to the last activity
2. **Number of Events**: Total count of activities in the case
3. **Number of Unique Activities**: Count of distinct activities
4. **Average Time Between Activities** (in days): Mean time interval between consecutive activities
5. **Activity Transitions**: Number of times the process changes from one activity to another

Additional features calculated but not heavily used in the final analysis included:

* Number of resources involved
* Presence of rejection activities

**3.2 K-Means Clustering Process**

The clustering process followed these steps:

1. **Data Preparation**: Extracting case-level features from event logs
2. **Feature Standardization**: Using StandardScaler to normalize the features
3. **Optimal Cluster Determination**: Applying the Elbow Method to find the optimal number of clusters (k)
4. **Cluster Assignment**: Assigning each case to its respective cluster
5. **Visualization**: Creating scatter plots and heatmaps to visualize the clusters
6. **Analysis**: Interpreting cluster characteristics and comparing process behaviors

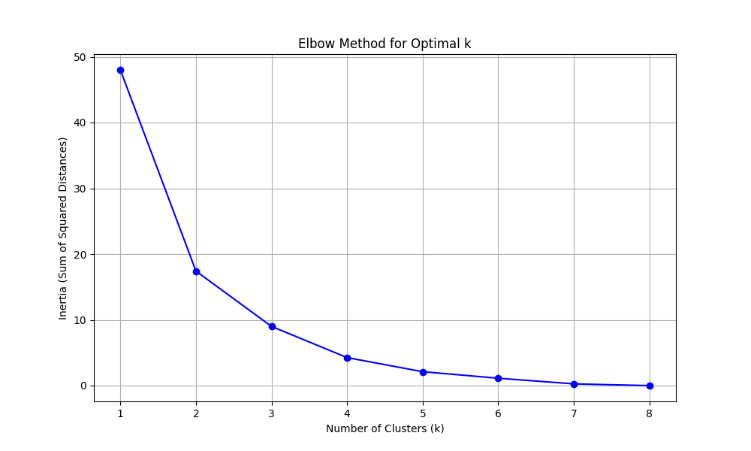
**4. Results**

**4.1 Optimal Number of Clusters**

Using the Elbow Method, we determined the optimal number of clusters to be **6**. This was achieved by plotting the sum of squared distances (inertia) against different values of k and identifying the point where adding more clusters yields diminishing returns.

**A screenshot of a chart

AI-generated content may be incorrect.**

****

**4.2 Cluster Characteristics**

| **Cluster** | **Count** | **Duration (days)** | **Num Events** | **Unique Activities** | **Avg Time Between (days)** | **Description** |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 2 | 0.11 | 7.5 | 7.5 | 0.017 | Short process with document issues |
| 1 | 2 | 3.31 | 12.0 | 12.0 | 0.301 | Full successful applications |
| 2 | 1 | 0.01 | 3.0 | 3.0 | 0.003 | Abandoned application |
| 3 | 1 | 1.13 | 11.0 | 10.0 | 0.113 | Rejected after document resubmission |
| 4 | 1 | 2.17 | 10.0 | 10.0 | 0.241 | Rejected after exam failure |
| 5 | 1 | 3.13 | 12.0 | 12.0 | 0.284 | Successful application (variant) |

**4.3 Cluster Interpretations**

**Cluster 0: Short Process with Document Issues**

* Cases 2 and 3
* Short duration (~0.11 days)
* Moderate number of events (7-8)
* Activities focused on application submission and document checking
* Process typically ends with notification about missing documents or rejection

**Cluster 1: Full Successful Applications**

* Cases 1 and 4
* Long duration (~3.31 days)
* Complete process flow (12 events)
* These applications go through the entire selection process
* End with "Selection of the winners"

**Cluster 2: Abandoned Application**

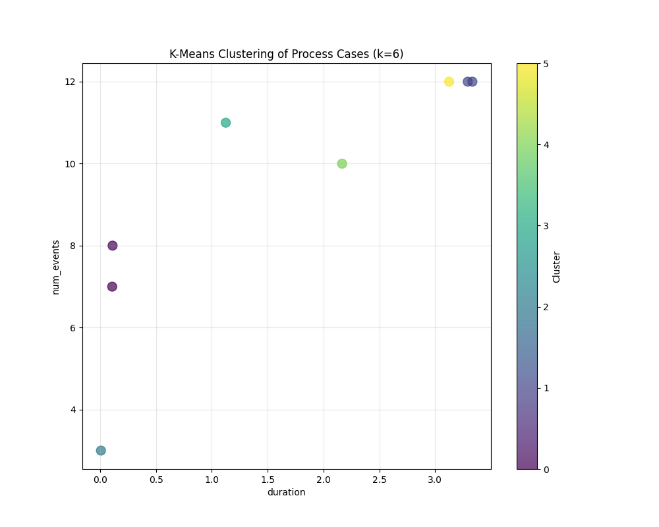
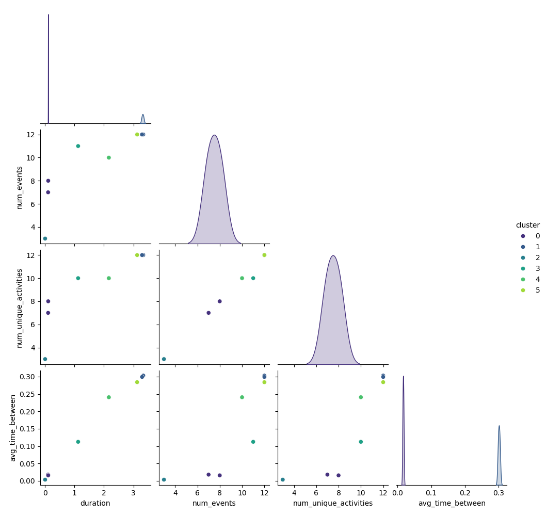
* Case 8
* Very short duration (~0.01 days)
* Only 3 events
* Process abandoned early by the applicant
* Represents incomplete applications

**Cluster 3: Rejected After Document Resubmission**

* Case 5
* Medium duration (~1.13 days)
* Has document issues but receives updated documents
* Ultimately rejected due to not meeting criteria

**Cluster 4: Rejected After Exam Failure**

* Case 6
* Medium-long duration (~2.17 days)
* Passes document verification and criteria checking
* Fails at the written exam stage

****

**Cluster 5: Successful Application (Variant)**

* Case 7
* Long duration (~3.13 days)
* Complete process flow (12 events)
* Similar to Cluster 1 but with slightly different timing patterns

**5. Analysis and Insights**

**5.1 Process Patterns**

The clustering reveals distinct patterns in the application process:

1. **Successful Applications** (Clusters 1 and 5): These follow the complete process flow with all stages and have the longest duration (~3.1-3.3 days).
2. **Early Rejections** (Cluster 0): Short-duration applications that are rejected early due to document or criteria issues, typically resolved within hours.
3. **Mid-Process Rejections** (Clusters 3 and 4): Applications that pass initial stages but are rejected after resubmission or exam stages, taking 1-2 days.
4. **Incomplete Processes** (Cluster 2): Applications abandoned very early by the applicant, representing process drop-offs.

**5.2 Process Performance Insights**

1. **Completion Time**: Successful applications (Clusters 1 and 5) take approximately 3 days to complete, while early rejections take only hours.
2. **Document Verification**: A critical phase where many applications face issues (Clusters 0 and 3).
3. **Exam Phase**: Another significant filtering stage (Cluster 4).
4. **Process Complexity**: Successful applications have the highest number of activity transitions (11), indicating complex process flows.

**5.3 Potential Process Improvements**

Based on the clustering analysis, several potential improvements can be suggested:

1. **Document Submission Guidelines**: Improve clarity of document requirements to reduce early rejections in Cluster 0.
2. **Abandoned Applications**: Investigate why applications are abandoned (Cluster 2) and implement measures to reduce drop-offs.
3. **Exam Preparation**: Provide better resources for exam preparation to reduce failures at this stage (Cluster 4).
4. **Process Timeline**: Consider optimizing the time between activities for successful applications to improve overall efficiency.

****

****

**6. Conclusion**

The K-Means clustering approach successfully identified distinct behavioral patterns in the job application process. The six clusters revealed meaningful differences in how applications progress through the system, from successful completions to early rejections and abandonments.

These insights can help process owners:

* Target specific improvement areas in the application process
* Set realistic expectations for process duration
* Identify bottlenecks and failure points
* Develop focused interventions for different types of applications

A close-up of a graph

AI-generated content may be incorrect.In summary, applying K-Means clustering to process event logs provides valuable insights that traditional process analysis might miss, enabling data-driven decision-making for process optimization.

A graph on a white background

AI-generated content may be incorrect.

A screenshot of a cluster

AI-generated content may be incorrect.