Identification of Patient Profiles Associated with Improved Neuroblastoma Survival: An Unsupervised Clustering Analysis of Clinical Data

Introduction

Neuroblastoma is a rare but aggressive pediatric cancer that originates from immature nerve cells, primarily affecting children under the age of five. Despite advances in treatment, the prognosis varies significantly depending on multiple clinical and biological factors, including patient age, tumor stage, histology, and response to therapy. Understanding which combinations of these factors are most strongly associated with positive outcomes remains a crucial challenge in pediatric oncology.

This study, titled "Identification of Patient Profiles Associated with Improved Neuroblastoma Survival: A Clustering Analysis of Clinical Data", employs unsupervised clustering techniques to identify groups of patients with similar characteristics, without relying on predefined labels such as clinical outcomes. The primary objective is to explore potential associations between the resulting clusters and relevant clinical variables, such as tumor differentiation grade, tumor location, and patient survival. This analysis aims to:

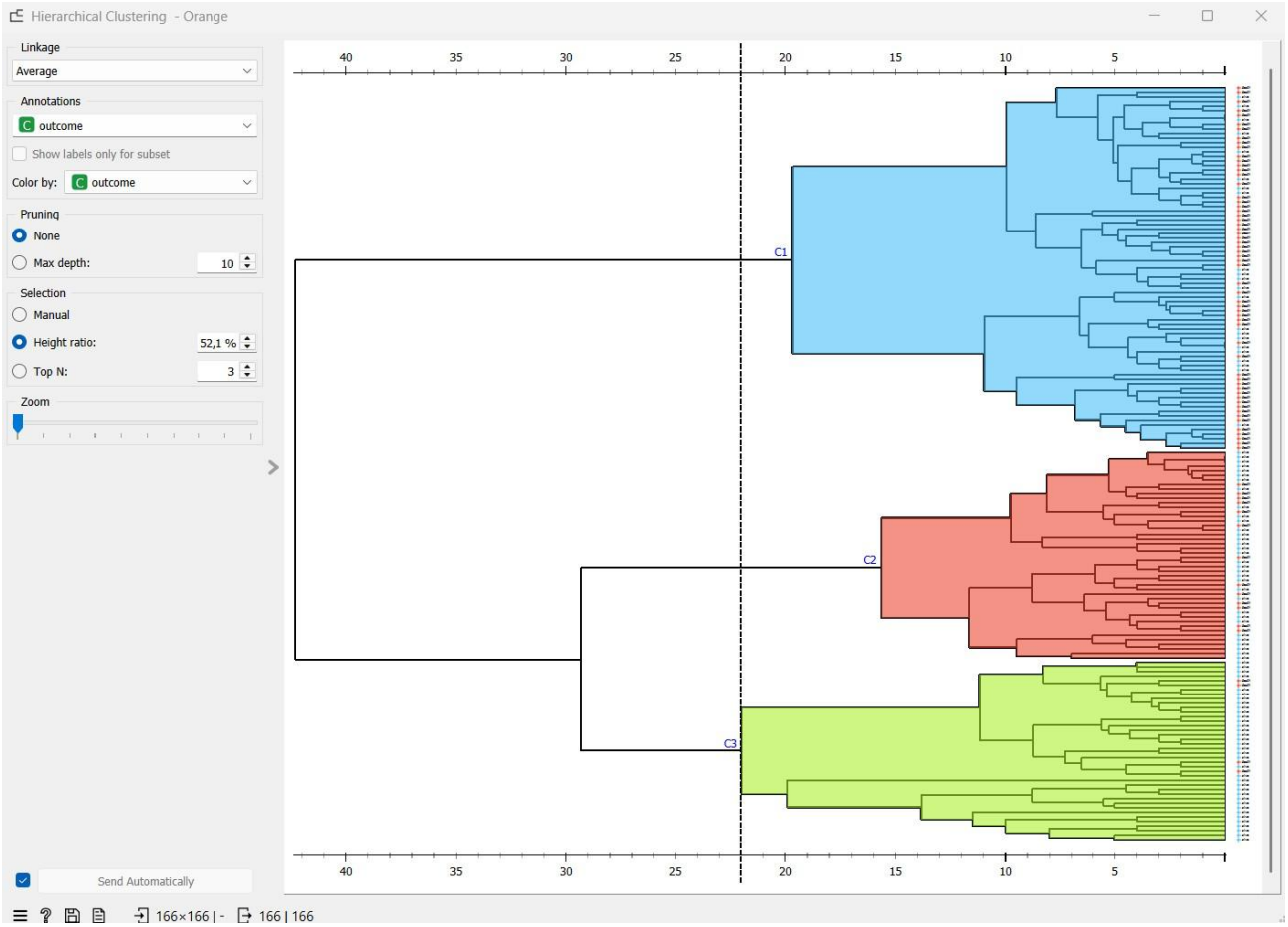
1. Identify subgroups of patients based on clinical and biological characteristics.
2. Assess whether significant correlations exist between the clusters and prognostic factors such as survival and differentiation grade.
3. Explore the potential use of unsupervised clustering to support clinical research and the personalization of therapies.

It is important to note, however, that the dataset used in this analysis is relatively small, consisting of only 169 records (or tuples) and 12 features. While the insights gained are valuable, the limited sample size imposes certain constraints on the generalizability and statistical robustness of the findings. The patterns identified should therefore be interpreted as exploratory rather than definitive, serving as a foundation for future studies involving larger cohorts and more comprehensive datasets.

The clinical data utilized in this study are derived from the research conducted by Ma et al., titled "Neuroblastomas in Eastern China: a retrospective series study of 275 cases in a regional center" . This retrospective analysis provides valuable insights into the clinical and pathological features of neuroblastoma patients in Eastern China, including MYCN status, surgical methods, and prognosis. The study highlights the significance of MYCN amplification as an adverse prognostic factor and underscores the importance of surgical approaches in patient outcomes. By leveraging this dataset, our analysis aims to build upon these findings through the application of clustering techniques to identify patient profiles associated with improved survival.

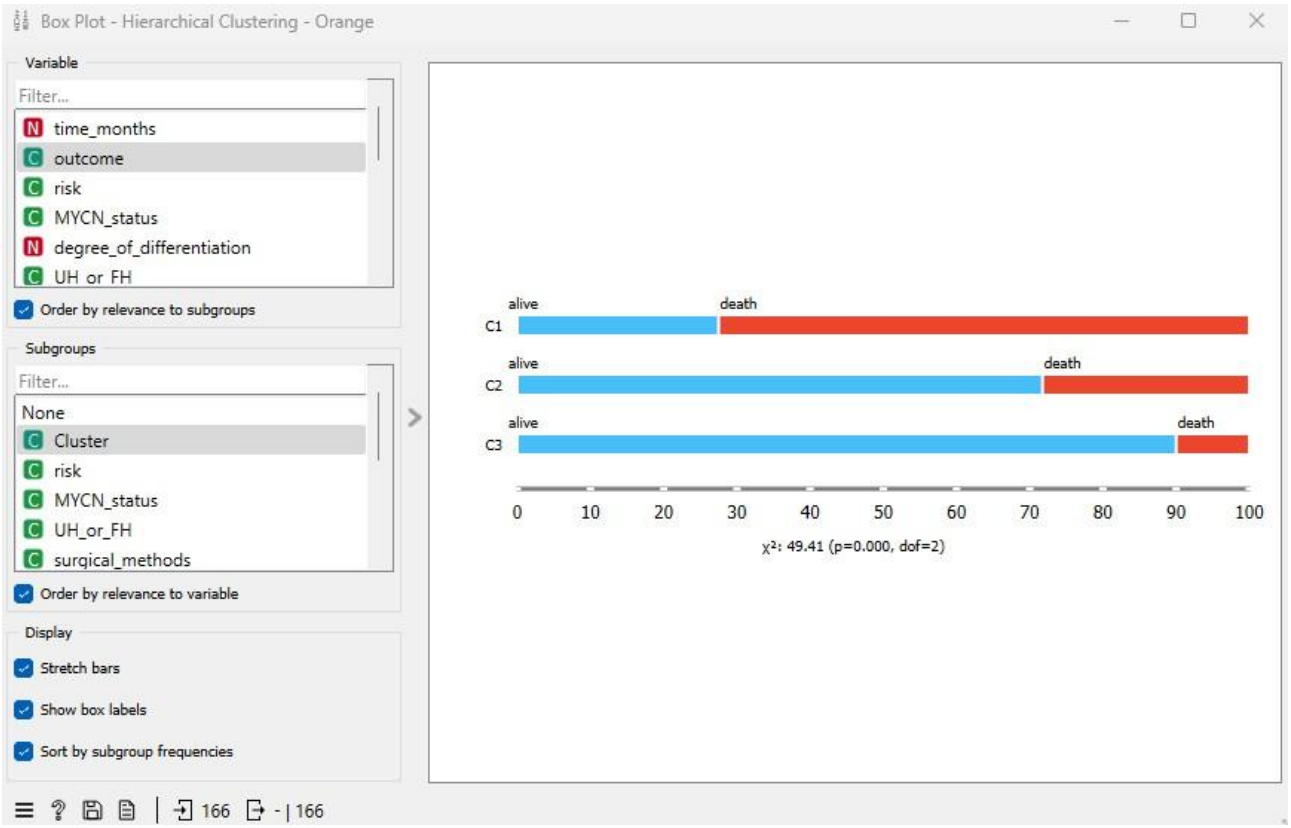
Process

I standardized the variables by centering the mean at 0. This approach yielded the best **visual** division of homogeneous clusters in **Hierarchical Clustering** and resulted in the highest **Silhouette Score** for **K-Means clustering**. This effect can be attributed to the diversity of variable types: many **boolean values** combined with **continuous variables** such as **age and time in months**. With this type of standardization, the different metrics had the least possible influence on the clustering outcome.

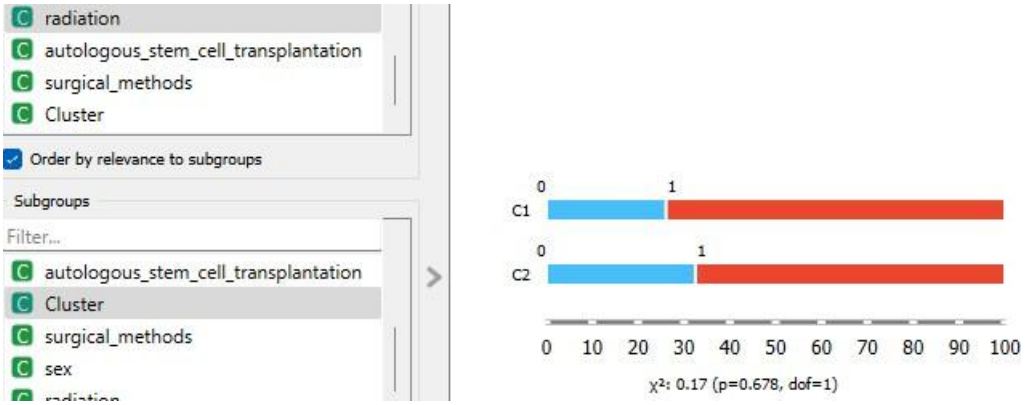


The **box plot** connected to the **hierarchical tree**, due to the significant differences in **patient outcomes** across the three identified clusters, enabled me to derive some valuable insights:

- **Cluster 1 (C1):** 80 patients, **27.5% survival rate**
- **Cluster 2 (C2):** 46 patients, **71.7% survival rate**
- **Cluster 3 (C3):** 40 patients, **90% survival rate**

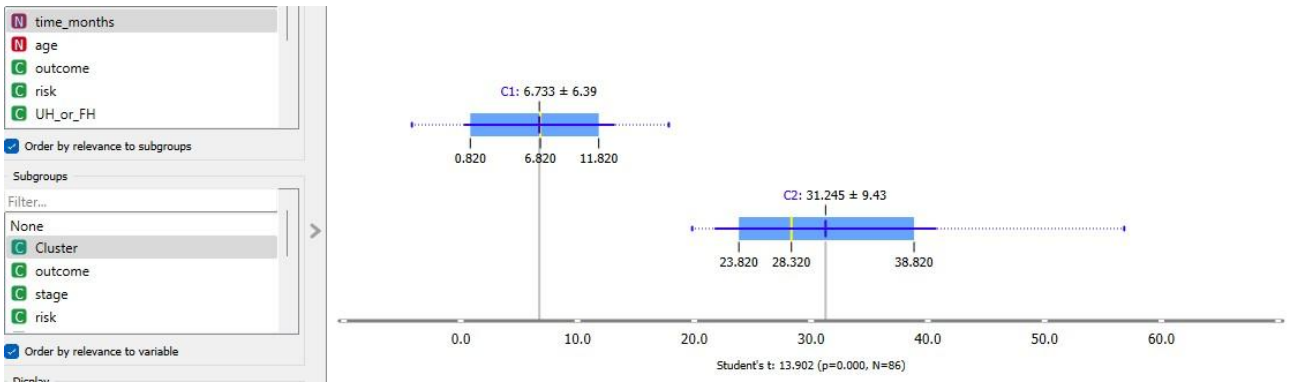


Clusters **C2 and C3** exhibit **striking similarities** regarding the **treatments administered** to patients. In both clusters, depending on the **risk level, tumor stage, and differentiation grade**, patients underwent **surgical interventions, radiotherapy, or stem cell transplantation** in almost identical proportions.

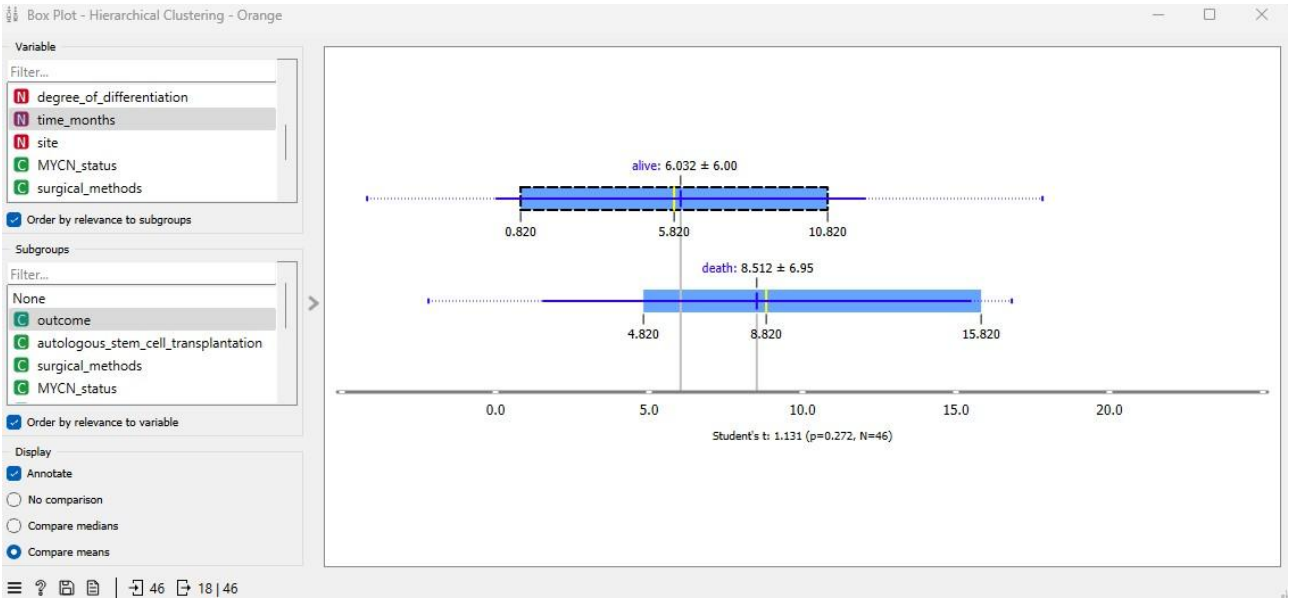


The Key Difference: Time_Months

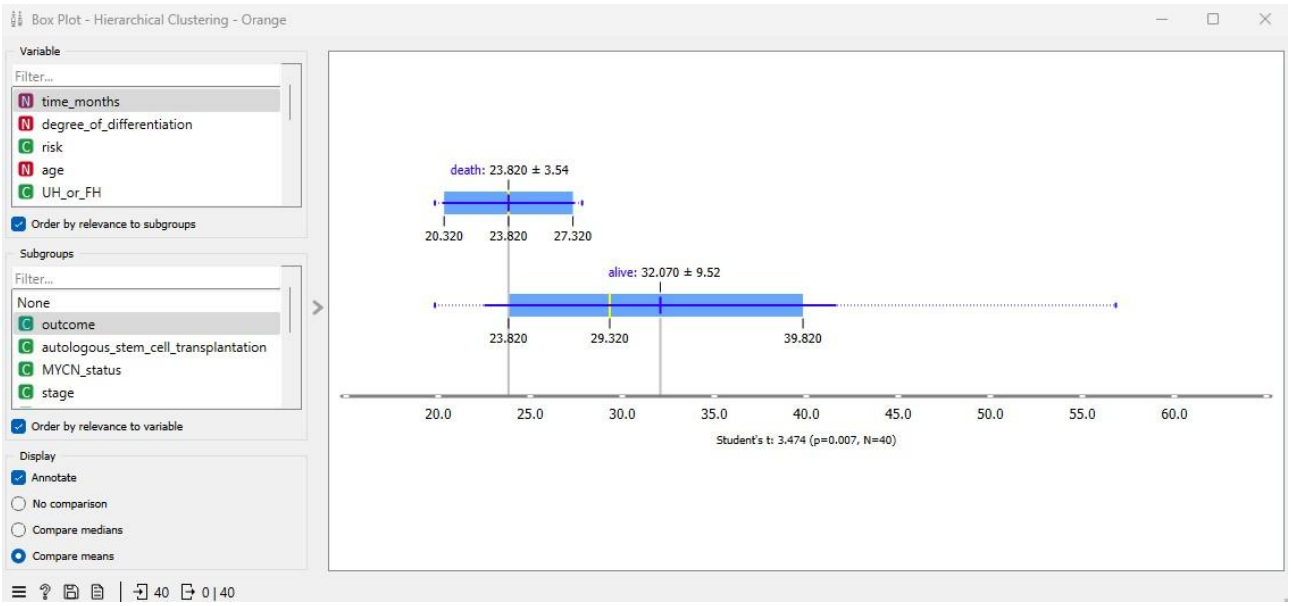
The fundamental difference between **C2 and C3** lies in the variable **"time_months."** Assuming that this variable represents the **follow-up period**, patients in **C3** were monitored for significantly longer periods than those in **C2**. The **maximum follow-up in C2 was around 18 months**, whereas in **C3, it extended up to 55 months**.



The **box plots** for **C2 and C3**, when broken down by the **"outcome"** subgroup, reveal a **considerable disparity in their medians**. This suggests that the **"time_months"** variable might directly influence the **predicted outcome**, despite similar characteristics in other clinical parameters.



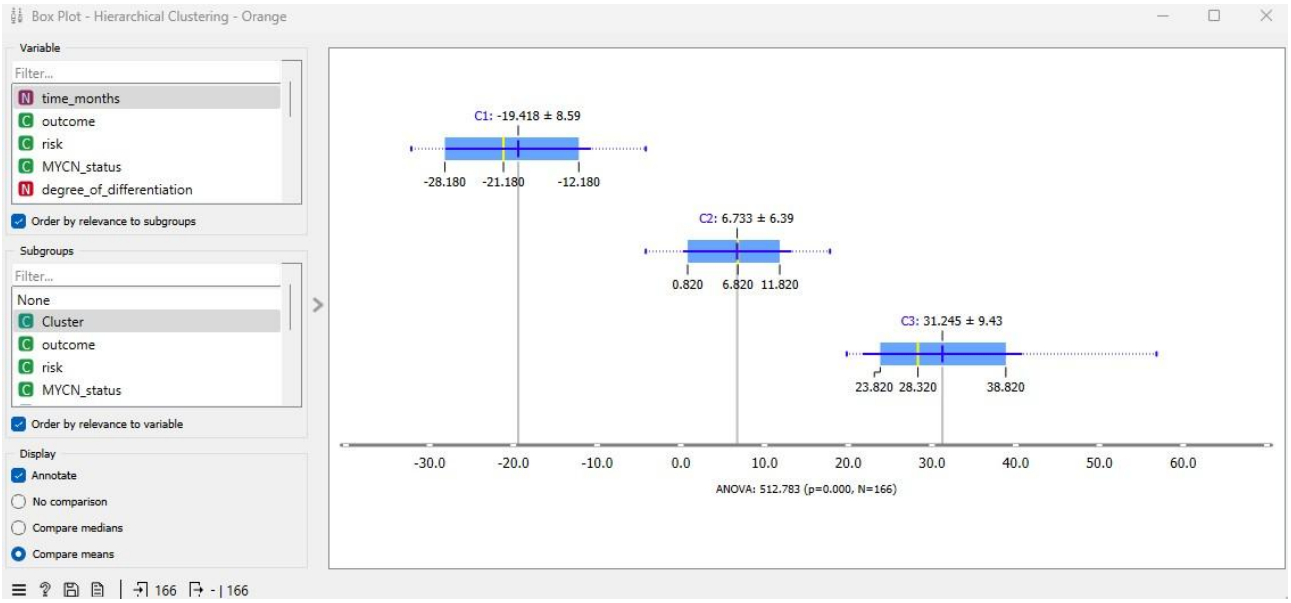
Box-plot about C2

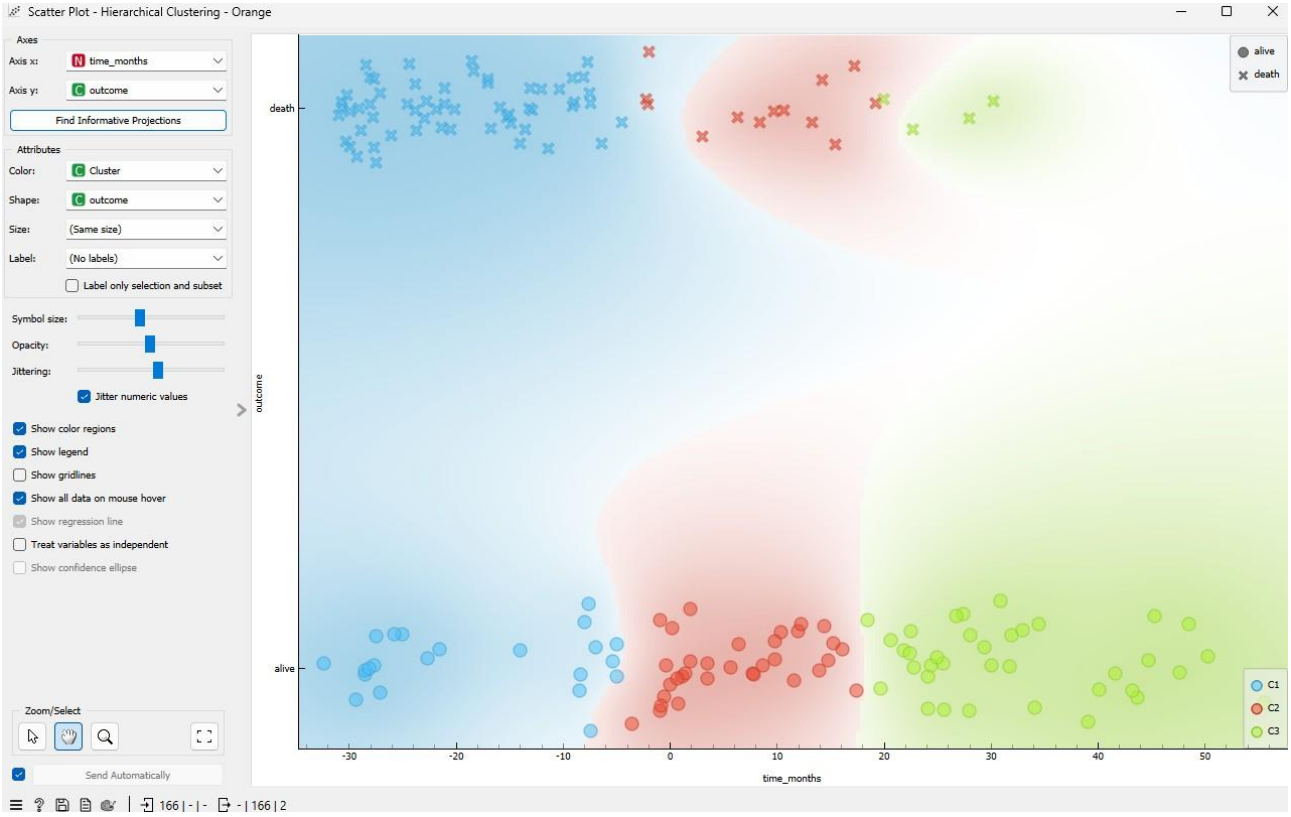


Box-plot about C3

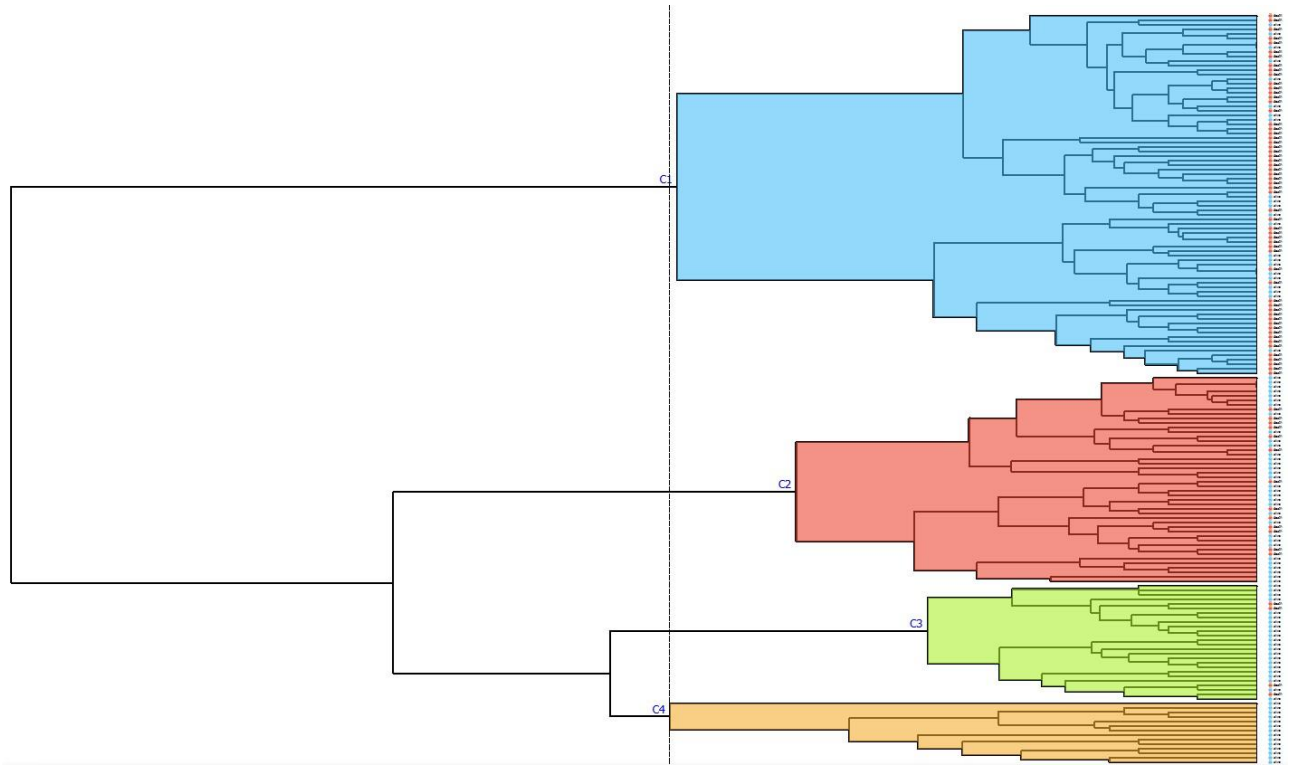
Correlation Between Outcome and Time_Months

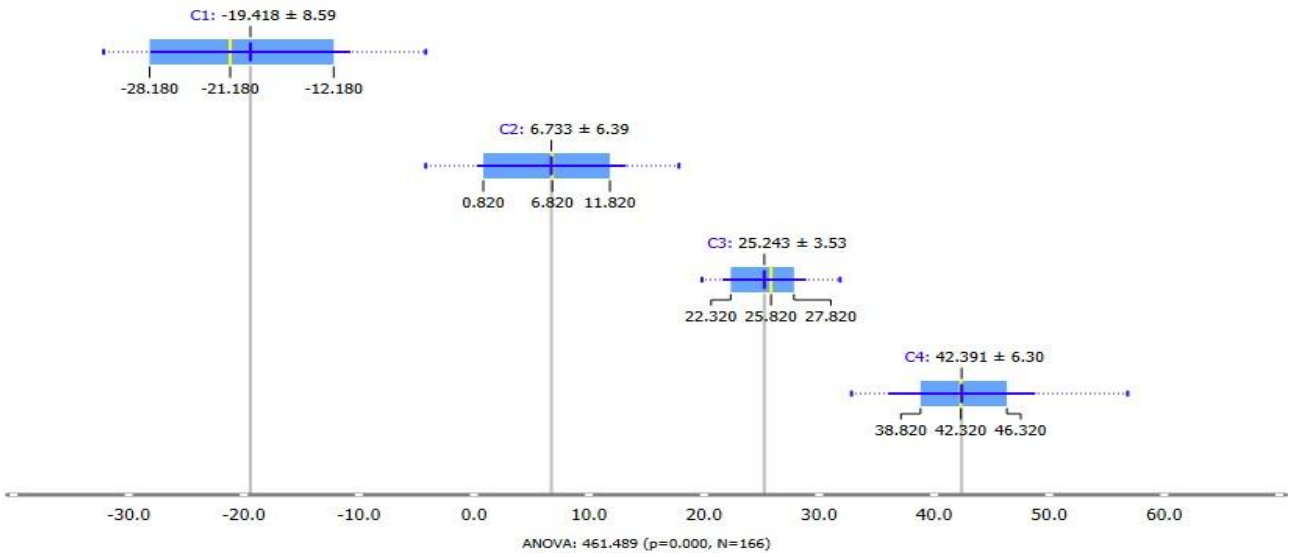
When comparing all three clusters together and considering their respective **survival rates** (**C1: 27.5%, C2: 71.7%, C3: 90%**), the correlation between **"outcome"** and **"time_months"** becomes even more apparent.

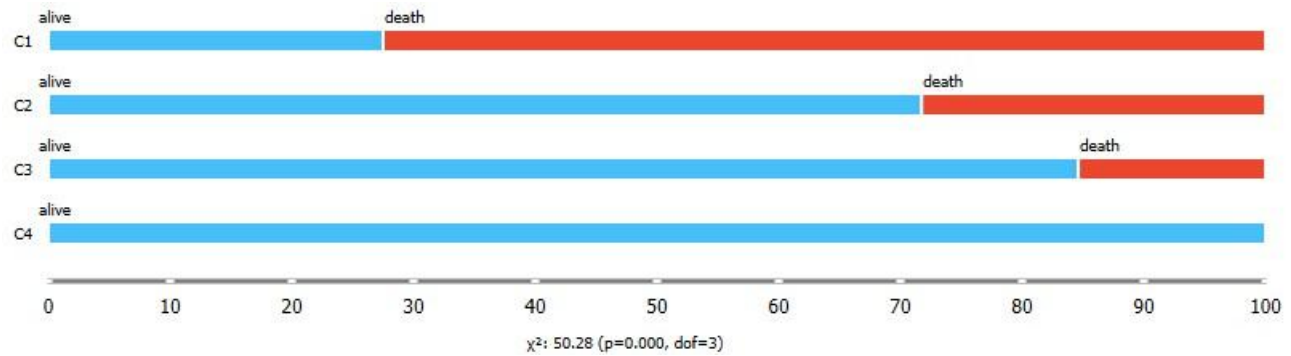




This comparison becomes even more evident if I choose to lower the evaluation height of the hierarchical clustering to 47% and analyze 4 clusters instead of 3.





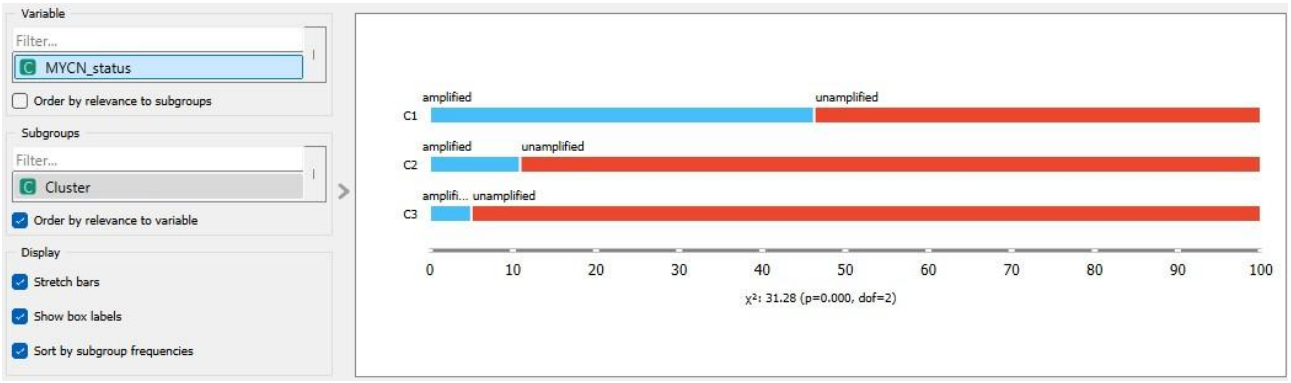


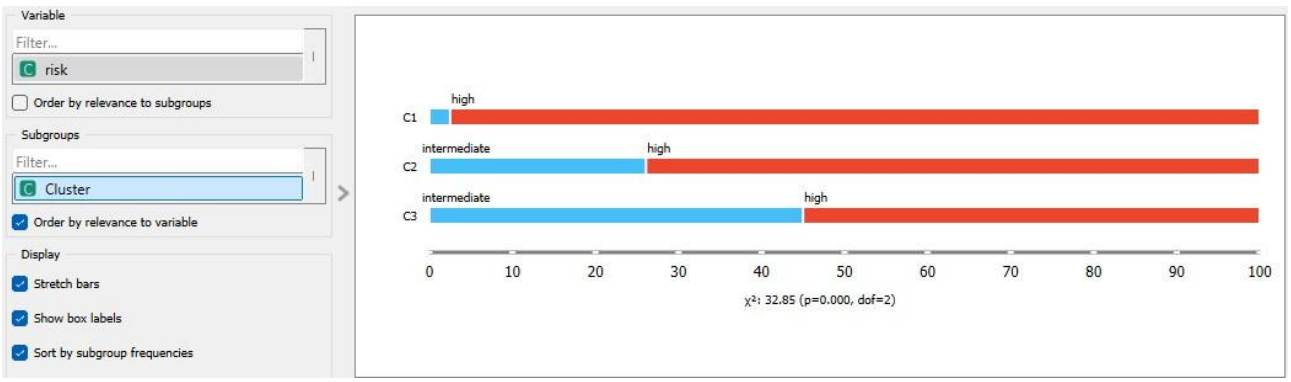
Thus, this analysis raises the question: **Does a longer follow-up inherently contribute to better survival outcomes, or is it simply an artifact of extended monitoring?** Further statistical validation is required to determine whether "time_months" plays a causative role in survival prediction or merely reflects variations in patient observation time.

Association Between Clustering, MYCN Status, Risk Levels in Neuroblastoma Prognosis

The bar plots display the distribution of MYCN amplification status and clinical risk levels across the three identified clusters (C1, C2, and C3). Cluster C1, which shows the poorest prognosis, includes a relatively high proportion of MYCN-amplified cases (blue) and is almost entirely composed of high-risk patients. This aligns with existing literature, where MYCN amplification and high-risk classification are strong indicators of aggressive neuroblastoma and poor outcomes. On the other hand, Cluster C3, associated with higher survival rates, consists mostly of unamplified MYCN cases and includes a significant share of intermediate-risk patients. Cluster C2 presents a mixed profile, with fewer MYCN-amplified cases and a majority of high-risk patients, placing it between C1 and C3 in terms of prognosis.

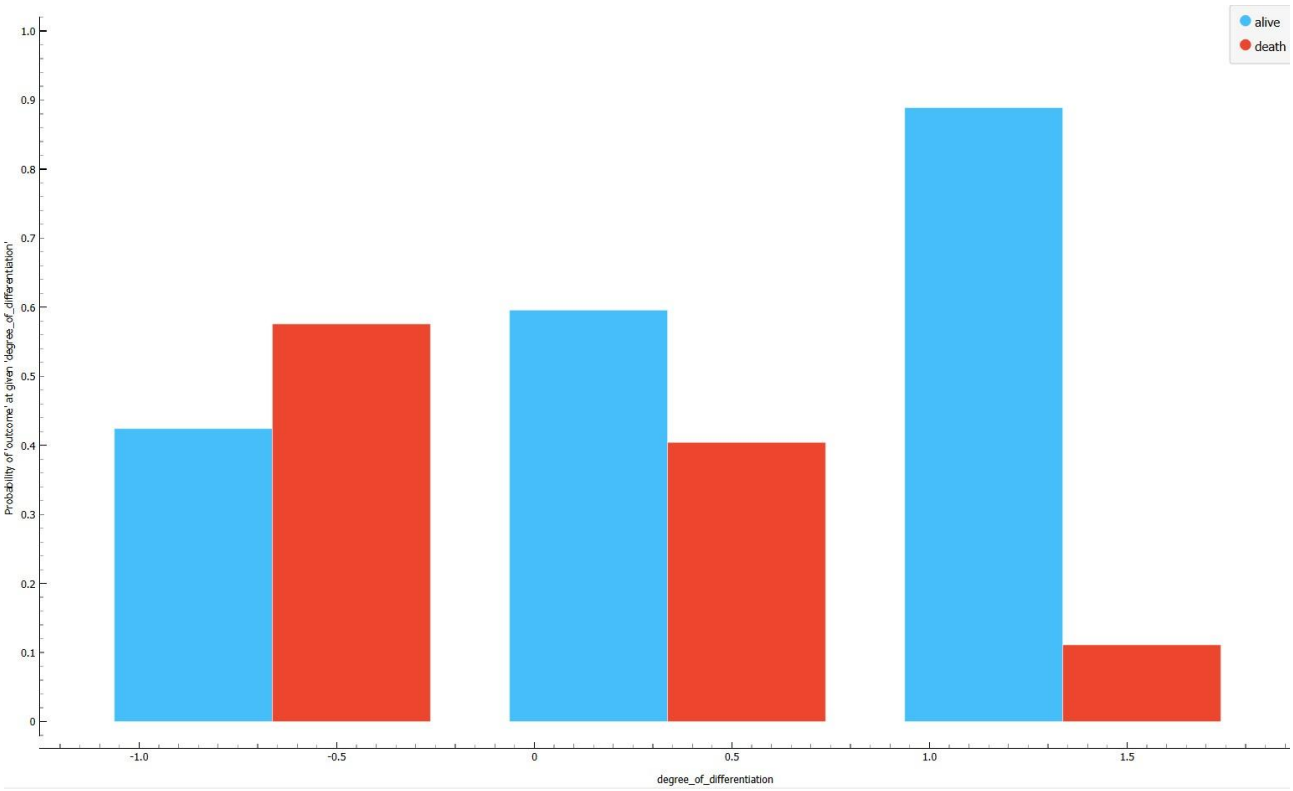
The chi-square test results (χ^2 = 31.28 and 32.85, p = 0.000 for both variables) confirm a statistically significant association between cluster assignment and both MYCN status and clinical risk. These findings support the relevance of the clustering process in effectively differentiating patient profiles based on survival-related clinical features.

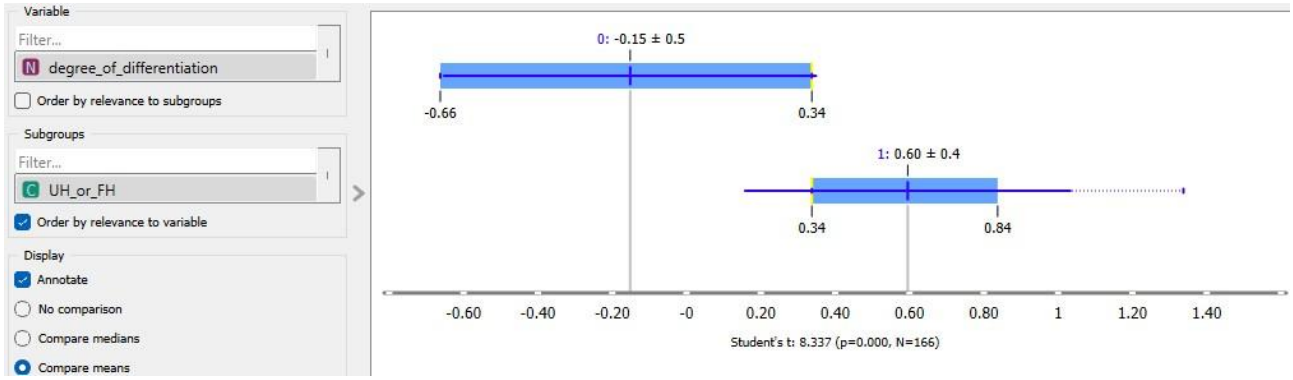




Higher Degree of Differentiation, Higher Probability of Survival

When considering the entire dataset as a whole, without dividing it into clusters, a notable pattern emerges: a higher degree of differentiation in neuroblastoma cells generally corresponds to a higher probability of survival. This correlation can be observed across various scenarios.





CONCLUSIONS

Algorithm used

In my study, I chose to primarily use *Hierarchical Clustering* and *K-Means*, two well-known algorithms for their effectiveness and interpretability, especially in contexts with a limited number of observations, as in my case (169 tuples and 12 features). Hierarchical Clustering provided a clear representation of the relationships between the data thanks to its tree structure, which was useful for naturally identifying the groups. K-Means, on the other hand, proved efficient in refining these groupings, also providing a good level of separation between the clusters, as indicated by the Silhouette coefficient values.

I also experimented with the *Mean-Shift* algorithm, but the results did not lead to conclusions substantially different from those already observed with K-Means and Hierarchical Clustering. For this reason, I decided not to explore its use further. I excluded *DBSCAN*, as it tends to be unstable on small datasets with varying densities, and *Birch*, which is designed for much larger datasets and thus not suitable for deriving practical benefits in my case. In light of these considerations, Hierarchical Clustering and K-Means proved to be the most suitable choices for achieving the objectives of my analysis.

Observations

The analysis conducted through unsupervised clustering techniques highlighted significant correlations between clinical variables and outcomes in patients with neuroblastoma. Standardizing the variables allowed for better cluster separation, confirmed by high Silhouette scores.

In cases where the tumor is highly aggressive, mortality tends to occur within the first few months following diagnosis. Conversely, if the hospitalization or follow-up period extends beyond the early critical phase, the likelihood of survival generally increases, suggesting that early endurance may be indicative of a more favorable prognosis.

The results clearly show that there are homogeneous patient groups with different survival probabilities, associated with clinical and biological features such as MYCN gene amplification, risk classification, and tumor differentiation grade. In particular, MYCN amplification, a well-known marker of tumor aggressiveness, is predominantly present in clusters associated with poorer outcomes. Similarly, patients classified as high-risk are more likely to belong to clusters with lower survival, while those with intermediate-risk profiles tend to show better prognoses. Additionally, the degree of tumor differentiation correlates with survival trends: poorly differentiated or undifferentiated tumors are more often found in clusters with higher mortality, whereas differentiated tumors are linked to more favorable outcomes. These findings reinforce the clinical relevance of clustering in supporting patient stratification and guiding treatment decisions.