Conrad Niewienda
Marcos Sanson

November 19, 2024

Lab 2: Survival Analysis

Instructions: In this lab, you aim to run a simple survival analysis using the *Bone marrow transplant: children* dataset, which describes pediatric patients with several hematologic diseases. The dataset is publicly available in the UCI repository.

Steps:

- Do some exploratory analysis to design the patient's features to be used. What features are missing, which do not seem to correlate with the survival time, and what groups of features are jointly correlated?
- Given the set of features, run some clustering algorithms to separate the patients into groups. You can use Kmeans, DBScan, or any other. Clustering can be done over the input feature or a reduced space using PCA or a VAE.
- Compute the survival function for each group using Kaplan-Meier and Cox Regression. Analyze the differences between both methods and the survival differences between groups. In terms of interpretability, recall that Cox provides a feature importance.
- Summarize everything in a report with a link to your code (use Google Drive or Github).
- You can work individually or in groups of up to three people.

Conrad Niewienda
Marcos Sanson

November 19, 2024

Objective

The goal of this lab was to analyze the survival outcomes of pediatric patients with hematologic diseases using the *Bone Marrow Transplant: Children* dataset. The lab included feature selection, clustering patients into groups, and applying survival analysis methods such as Kaplan-Meier and Cox Regression to compare survival curves and interpret key factors affecting survival.

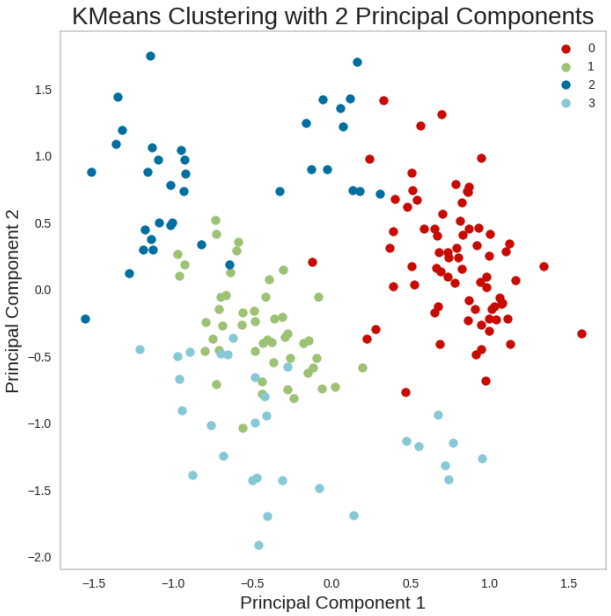
Methodology

Exploratory Data Analysis (EDA):

- A correlation matrix and feature importance analysis was used to examine relationships between features. Features such as *Disease_chronic* and *aGvHD_III_IV* showed significant correlations with survival time. Features like *RecipientGender* and *RecipientCMV* demonstrated weak correlations with survival time.

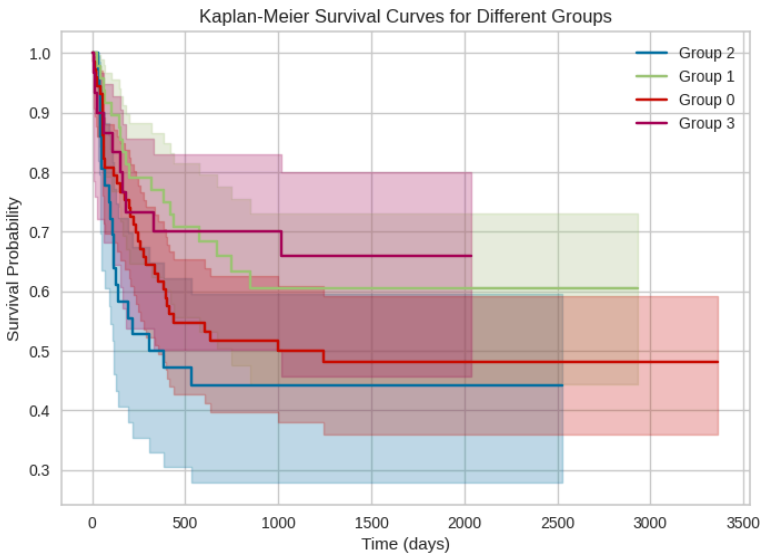
Clustering:

- PCA (Principal Component Analysis) was applied to reduce the feature space dimensionality while retaining significant variance.
- KMeans clustering grouped patients based on their feature profiles. The elbow method was used to determine the optimal number of clusters.
    - The elbow method determined the optimal number of clusters as 4. The final cluster sizes were: Cluster 0 (70), Cluster 1 (49), Cluster 2 (37), and Cluster 3 (31).
- Clusters were visualized using PCA-transformed data. The clusters showed clear separation and meaningful group distinctions.



Survival Analysis:

- Kaplan-Meier survival curves were generated for the clusters. Each cluster displayed distinct survival probabilities over time.
    - Additional Kaplan-Meier analysis was performed for recipient age groups (0.0, ~0.47, and 1.0).
    - The results showed that older recipients had significantly lower survival probabilities compared to younger patients (see code).
- Cox Regression analysis quantified the impact of individual covariates on survival. It highlighted both high-risk features, such as *Recipientage* and *Relapse*, and protective features like *PLT recovery* and *ANC recovery* (see below).



Conrad Niewienda
Marcos Sanson

November 19, 2024

Key Insights

1. **Feature Importance:**
    - Chronic diseases (*Disease_chronic*) and severe graft-versus-host disease (*aGvHD_III_IV*) significantly reduced survival probabilities.
    - *Recipient age* and *relapse* emerged as key risk factors, while *PLT recovery* and *ANC recovery* improved survival outcomes by reducing hazard.
2. **Survival Curves by Recipient Age:**
    - Kaplan-Meier survival curves showed significant differences in survival probabilities across recipient age groups (0.0, ~0.47, and 1.0).
    - Younger recipients (age 0.0) consistently demonstrated higher survival probabilities over time, while older recipients (age 1.0) exhibited lower survival rates. The baseline survival curve, representing the overall cohort, fell between these groups.
3. **Hazard Ratios for Covariates:**
    - Cox Regression quantified the survival risks associated with recipient age. Older age was strongly linked to increased hazard and reduced survival probabilities.
    - Features like *PLT recovery* and *ANC recovery* were associated with a lower hazard.
4. **Comparison of Kaplan-Meier and Cox Regression:**
    - Kaplan-Meier curves provided an effective visualization of survival differences between patient groups and recipient age categories. However, they did not offer insights into the specific contribution of individual features.
    - Cox Regression addressed this limitation by quantifying the impact of covariates such as *recipient age*, *relapse*, and *Disease_chronic*.
5. **Clustering and Patient Heterogeneity:**
    - Clustering analysis using KMeans revealed four distinct patient groups, with significant differences in survival outcomes.
    - Kaplan-Meier survival curves confirmed that these clusters represent heterogeneous subpopulations with varying survival probabilities.

Conclusion

The analysis demonstrated the utility of clustering and survival models in understanding pediatric bone marrow transplant outcomes. Critical factors such as *Disease_chronic*, *aGvHD_III_IV*, and *recipient age* were found to significantly impact survival probabilities. Kaplan-Meier curves highlighted group-level survival differences, while Cox Regression quantified the influence of individual features, identifying both risk and protective factors. Future work could enhance these findings by incorporating additional clinical features and exploring alternative clustering techniques for improved patient stratification.

Link to code:

Google Colab:
https://colab.research.google.com/drive/1ZJ66sNz2wj7xf9S61LncpyixrBPlaO2d?usp=sharing
GitHub:
https://github.com/Marcos-Sanson/UC3M-ML/blob/main/LAB02.ipynb