Machine Learning In Healthcare

Lab 1:
Survival Analysis

Jorge Chamorro Pedrosa – 100496527
Juan José Jiménez De Juan – 100496468
Mario Fernández Busto – 100496459

1. Text Preprocessing

1.1 Missing Data

The variables with the highest proportion of missing values were extcGvHD (16.5%), CMVstatus (8.5%), and RecipientCMV (7.4%). Moderate missingness was also observed in CD3dkgx10d8 and CD3dCD34 (around 2–3%).

1.2 Imputation Strategy

Missingness in categorical variables such as extcGvHD and CMVstatus was handled by creating missing flags and imputing a value of −1. Continuous variables like Rbodymass were imputed with the median, while RecipientRh was filled with the mode (value 1). Records with missing information in RecipientABO, Antigen, or Allele were removed due to low frequency. For CD3dkgx10d8 and CD3dCD34, regression-based imputation was used.

1.3 Correlation with Survival Time

The strongest positive association with survival time was found in extcGvHD ($r \approx 0.47$), suggesting longer survival among patients with this condition. PLTrecovery showed a moderate negative correlation ($r \approx -0.33$), indicating earlier events or faster recovery. All other variables displayed weak or negligible linear relationships with survival time.

2. Clustering and PCA

2.1 Principal Component Analysis (PCA)

PCA was performed on 35 standardized numeric features. The first ten components explained approximately 69% of the total variance (PC1–PC10 cumulative). The first two components captured 26% of the variance, mainly reflecting HLA-related and demographic variability. Visual inspection of PC1 vs PC2 indicated moderate separation of samples by survival status, suggesting partial biological relevance of the principal components.

2.2 Feature Contribution

Loadings from PC1 and PC2 highlighted HLAmatch, HLAgrI, Allele, Recipientage, Rbodymass, Antigen, Recipientageint, and HLAmismatch as the most influential variables. These features were selected for subsequent clustering, as they contributed most to inter-patient variation captured by the PCA.

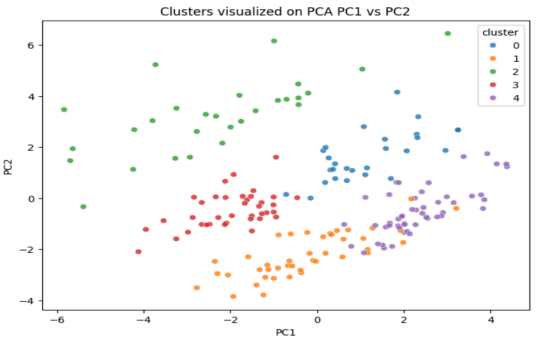
2.3 Clustering

KMeans clustering was applied on the selected features using standardized values. The optimal number of clusters was $k = 5$, determined by the highest silhouette score (0.5062). Cluster visualization in the PCA space (PC1 vs PC2) showed distinct but overlapping groups, consistent with moderate separability.

2.4 Cluster Interpretation

Clusters captured demographic and immunologic heterogeneity:

- Cluster 0 grouped older, heavier recipients (higher Recipientage and Rbodymass).
- Cluster 1 represented younger patients with low HLA matching and low antigen levels.
- Cluster 2 was characterized by strong HLA mismatch and higher HLAgrI values.
- Cluster 3 contained low-age, low-mass recipients with intermediate antigen values.
- Cluster 4 included older recipients with contrasting antigen and HLA profiles.



3. Kaplan-Meier and Cox Regression

3.1 Kaplan–Meier Curves

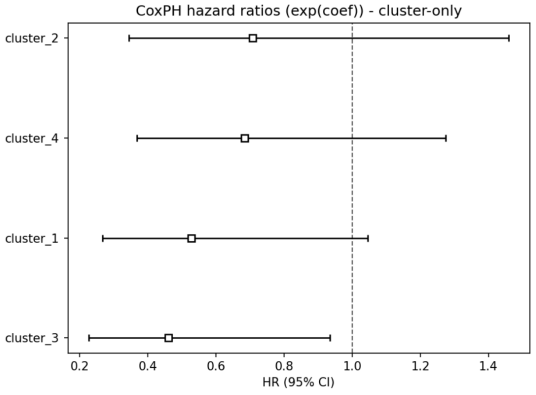
Patients were divided into five clusters (sizes: 27, 42, 28, 38, 50). The Kaplan–Meier plot ("km_clusters.png") showed distinct survival patterns among clusters, with visible separation suggesting heterogeneous survival behavior. Pairwise log-rank testing revealed one statistically significant difference (Cluster 0 vs Cluster 3, p = 0.043), while other comparisons were not significant after unadjusted testing.

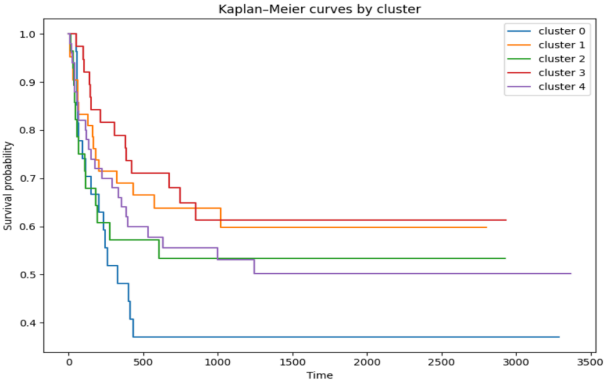
3.2 Cox Proportional Hazards Model

A Cox model including only cluster indicators (reference = Cluster 0) estimated relative hazard ratios. Cluster 3 displayed a lower hazard (HR ≈ 0.50, coef = −0.699, p ≈ 0.056), indicating roughly half the risk compared to Cluster 0, though the result was marginally non-significant. The concordance index (≈0.57) reflected moderate discriminative ability.

3.3 Interpretation

Both Kaplan–Meier and Cox analyses suggest that the unsupervised clusters capture differences in survival outcomes. In particular, Cluster 3 appears associated with improved survival relative to others. However, as clustering and survival assessment were performed on the same dataset, these findings are descriptive and require external validation.





Link GitHub → [joorgee14/Machine-Learning-In-Healthcare]