Projet Final — Analyse de Survie

Solo ou binôme

Master 2 MIASHS

Année universitaire 2025–2026

Data limite : Dim. 9 novembre 2025 à 23h59.

Objectif du projet

Ce projet vise à mettre en pratique les méthodes vues en cours à travers la réalisation d'une analyse de survie complète sur un jeu de données : partir d'une cohorte brute, comprendre ses données, construire un modèle robuste, l'interpréter et le valider. Votre Notebook doit raconter une histoire statistique claire, du contexte à la conclusion.

Vous devez donc choisir le **jeu de données de votre choix** pour y appliquer une analyse de survie. L'objectif est de démontrer votre capacité à :

- décrire la cohorte associée à votre dataset ;
- explorer et décrire ses variables pertinentes ;
- appliquer les méthodes d'estimation, de modélisation et d'évaluation des modèles de survie ;
- interpréter correctement les résultats obtenus dans un cadre scientifique cohérent.

Consignes générales

- Le livrable final doit être un **Notebook Python** (`.ipynb`) clair, documenté et exécutable.
- Toutes les packages Python sont authorisés, je vous conseille en particulier `lifelines` et `sksurv` qui ont été vus en cours mais d'autres existente (`skglm`,`pycox`, et j'en passe).
- Les cellules Markdown doivent contenir vos explications, interprétations, définitions mathématiques et conclusions.
- La qualité de présentation, la rigueur mathématique et la clarté du raisonnement sont essentielles.

Structure attendue du Notebook

REMARQUE IMPORTANTE. Ce que je vous suggère ci-dessous n'est pas obligatoire, le sujet étant assez libre. Néanmoins, je pense avoir vous suggérer les points primordiaux à suivre lorsque l'on fait de l'inférence et en particulier de l'analyse de survie.

1. Choix et présentation du jeu de données

- Choisissez un dataset **différent de celui du TP** (autonome, open source, ou issu d'un domaine d'intérêt).
- Vérifiez qu'il contient au moins :

— une variable de durée (`time`, `duration`, etc.) ;
— une variable d'évènement (`event`, `status`, etc.).

- Présentez le contexte (source, domaine, objectif).
- Décrivez les covariables : type, distribution, pertinence biologique/clinique/statistique.
- Réalisez des statistiques descriptives (taille de l'échantillon, tableaux de grandeurs statistiques usuelles, histogrammes, boxplots, corrélations pertinentes, nombre de censurés VS non censurés, répartion de la variable T, ...).

2. Analyse de survie descriptive : Kaplan–Meier

- Estimez la fonction de survie $\widehat{\mathcal{S}}(t)$ par la méthode de **Kaplan–Meier**.
- Calculez et affichez :
  — le **temps médian de survie** et son **intervalle de confiance à 95%**,
  — la fonction de survie par sous-groupe (selon une covariable pertinente : sexe, traitement, etc.).
- Comparez les courbes entre groupes que vous aurez vous-même choisis à distinguer à l'aide du **test du log-rank**.
- Interprétez les différences observées.

3. Modélisation paramétrique et semi-paramétrique

- Implémentez un **modèle de Cox proportionnel** : $h(t|\mathbb{X}) = h_0(t)\exp(\mathbb{X}^\top \widehat{\boldsymbol{\beta}})$.
- Interprétez les coefficients $\widehat{\beta}_j$ via les **hazard ratios** : $\mathrm{HR}_j = \exp(\widehat{\beta}_j)$.
- Commentez les effets significatifs, les intervalles de confiance et le test de Wald.
- Comparez plusieurs modèles : Cox standard (lifelines) ; Cox régularisé (Elastic Net, Ridge ou LASSO) : Cox neuronal (`CoxPHSurvivalAnalysis` ou modèle NN-sksurv, si souhaité) ; Random Survival Forest : ...

4. Évaluation et validation

- Évaluez la **discrimination** :

$$\text{C-index} = \frac{\text{nombre de paires ordonnées correctement selon le risque}}{\text{nombre total de paires comparables}}$$

Autrement dit : le C-index mesure la proportion de paires comparables pour lesquelles l'ordre des scores de risque est cohérent avec l'ordre des temps observés.

- Évaluez la **calibration** :
  — courbe de calibration (Calibration plot) à un horizon choisi $t^\star$;
  — Brier Score à ce même instant d'horizon $t^\star$ (*e.g.*, temps médian de survie).

5. Discussion et conclusion

- Interprétez vos résultats en lien avec le contexte du dataset.
- Discutez des limites du modèle et des hypothèses (proportionnalité des risques, taille d'échantillon, etc.).
- Concluez sur l'intérêt et la portée de l'analyse réalisée.

Barème indicatif (Note / 20)

Critère d'évaluation	Points
Choix du dataset, originalité et pertinence	2
Présentation claire du contexte et des variables	3
Analyse descriptive et Kaplan–Meier (courbes, IC, log-rank test)	2
Construction et interprétation du modèle de Cox (ou variantes)	2
Qualité de l'analyse inférentielle (tests, IC, HR, interprétation)	3
Évaluation du modèle (C-index, calibration, Brier Score, AIC/BIC) et **discussion du modèle appliqué au contexte**	2
Clarté du code, des figures et **surtout de la rédaction** (comment vous justifiez, en format Markdown, propreté du Notebook)	6
Total	**/ 20**

Conseil : La qualité de la rédaction (figures propres, commentaires clairs, explications concises et précises) comptera autant que la qualité statistique. Un Notebook bien structuré et interprété vaut toujours mieux qu'un code long et peu commenté.

Suggestions de dataset

Un exemple de lien où piocher des datasets : `https://lifelines.readthedocs.io/en/latest/lifelines.datasets.html`. Vous pouvez choisir un dataset autres que ceux présentés ici mais je vous propose des références pour des projets académiques :

Dataset	Domaine	Lien / Source
`Rossi`	Récidive criminelle après sortie de prison	lifelines.load_rossi()
`Lymphoma`	Étude de survie sur le lymphome	lifelines.load_lymphoma()
`Canadian Senators`	Étude parlementaire de longévité	sksurv.datasets.load_canadian_senators
`Veterans Lung Cancer`	Cancer du poumon	sksurv.datasets.load_veterans_lung_cancer
`GBSG2`	Cancer du sein	sksurv.datasets.load_gbsg2
`AIDS dataset`	Étude sur le VIH/SIDA	sksurv.datasets.load_aids
`NHANES I`	Cohorte populationnelle américaine	shap.datasets.nhanesi
`SUPPORT Study`	Soins intensifs / pronostic vital	DeepSurv paper (SUPPORT study)
`METABRIC`	Génomique du cancer du sein	DeepSurv paper (METABRIC)
`WHAS500`	Post-infarctus (Worcester Heart Attack Study)	sksurv.datasets.load_whas500
`FLCHAIN`	Protéine sérique et mortalité	sksurv.datasets.load_flchain
`SEER`	Registre national américain du cancer	Kaggle SEER dataset

Bon travail !