Machine Learning Project Report

Section 'F'	
Nathan Matthew Paul	PES2UG23CS368
Navneet Nayak	PES2UG23CS371

Pancreatic Adenocarcinoma Prognosis

Pancreatic cancer like Pancreatic Adenocarcinoma is often diagnosed at an advanced stage, when surgical resection is no longer possible. Accurate prognosis prediction at borderline resectable or locally advanced stages can significantly aid oncologists in risk stratification and treatment planning.

Problem Statement
This project aims to develop and evaluate a survival analysis model to predict high and low risk groups among pancreatic cancer patients by training **Cox Proportional Hazard** model on clinical features, and evaluation of performance using **Harrell's Concordance Index** and **Kaplan-Meier Survival Curves**

High Level Architecture

The project follows a modular, end-to-end **machine learning pipeline**, designed for reproducibility, interpretability, and robustness. The main stages are:

a. Data Loading and Preprocessing

- The dataset was loaded from a `.tsv` file (`paad_tcga_gdc_clinical_data.tsv`).
- Survival-related columns were standardized:
  - `Overall Survival (Months)` → `duration`
  - `Overall Survival Status` → `event`, encoded as binary (1 = deceased, 0 = censored).
- Rows with missing survival time or event data were removed.
- Columns that could cause **data leakage** (e.g., post-diagnosis outcomes or survival-related timestamps) and **identifier columns** (e.g., patient or sample IDs) were dropped.

b. Feature Selection and Cleaning

- Columns with more than **40% missing values** were removed.
- Features with **constant values** (no variability) were filtered out.
- Remaining features were divided into **categorical** and **numerical** groups.

c. Categorical Encoding

- All categorical variables were **one-hot encoded** using `pandas.get_dummies()`

d. Missing Value Imputation

- Numerical features were imputed using a **Multivariate Iterative Imputer** (`IterativeImputer`), which models each feature as a function of others, iteratively refining estimates.

e. Feature Scaling

- Features were **standardized** using `StandardScaler` to ensure numerical stability for the Cox model and improve convergence.

f. Data Splitting

- Data was divided into training and testing sets (80/20 split) to evaluate out-of-sample performance.

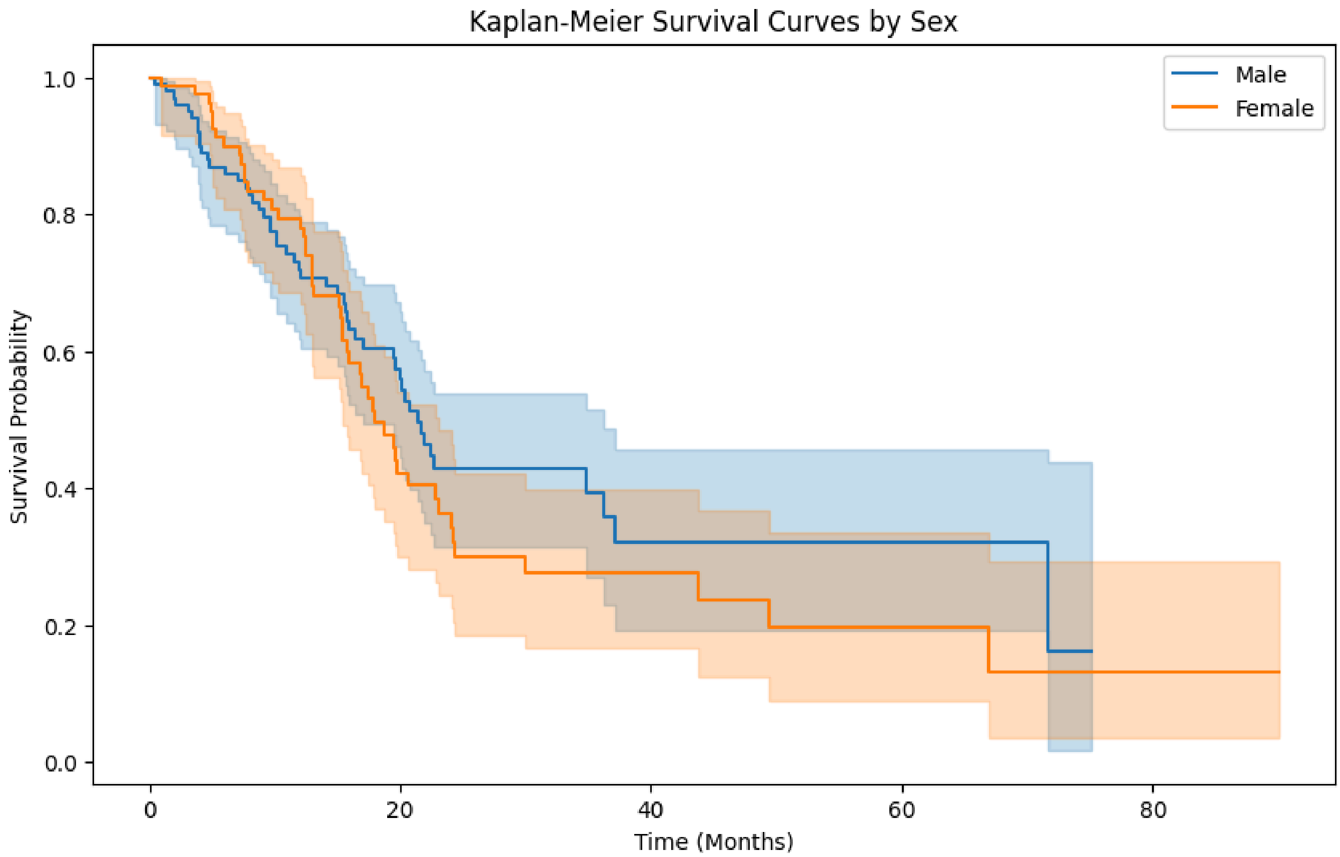
Cox Proportional Hazards Model

- Implemented using `lifelines.CoxPHFitter`.
- Regularized with an **elastic-net style penalty** (`penalizer=0.1`, `l1_ratio=0.5`) to handle multicollinearity and prevent overfitting.
- The model estimates the **hazard ratio** for each covariate, representing its influence on patient survival.

Evaluation

Performance was measured using the **Concordance Index (C-index)**, which essentially checks how often the model correctly ranks pairs of patients by their survival times. The model achieves a C-index of `0.67` on training data and `0.62` on test data.

We also visualize survival curves for each gender using **Kaplan-Meier Survival Curves**


Kaplan-Meier Survival Curves by Sex
Male
Female
Survival Probability
Time (Months)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80