**Executive Summary**

This project aims to perform survival analysis and identify clinical predictors of poor treatment outcomes in breast cancer patients using machine learning. The goal is to build a web-based application that integrates clinical and genomic features to stratify patients by risk level, thereby supporting precision oncology. The application will allow users to input clinical and genomic data and receive predictions on survival probabilities and risk classifications, ultimately informing treatment strategies.

**Motivation**

Breast cancer remains one of the leading causes of cancer-related mortality in women globally. Despite advancements in treatment, predicting individual outcomes continues to be a challenge due to the complex and heterogeneous nature of the disease. Personalized treatment decisions—guided by data analysis—have the potential to significantly improve prognosis and quality of life. This project is especially meaningful to me, as my best friend is currently undergoing treatment for breast cancer. Witnessing her journey has deepened my commitment to contributing to tools that can enhance clinical decision-making. Existing prognostic models often fall short by not fully integrating comprehensive genomic data. This project aims to bridge that gap by leveraging the METABRIC dataset to develop a machine learning–based risk prediction model and a user-friendly application that supports precision oncology.

**Data Question**

**Question 1:**  
How can clinical and genomic data be used to predict survival outcomes and identify risk factors for poor prognosis in breast cancer patients?

**Question 2:**  
Which clinical and genomic features are most predictive of poor survival outcomes? How can we present risk information in an accessible and interpretable format for clinical use?

**Question 3:**  
What machine learning models (e.g., Cox Proportional Hazards, Random Survival Forests, DeepSurv) are best suited for survival analysis with high-dimensional genomic data, and how do their performances compare?

**Question 4:**  
Can integrating treatment-specific variables (e.g., chemotherapy, hormonal therapy) improve the model’s ability to predict differential survival outcomes across treatment regimens?

**Relevant Literature:**

* Curtis et al., Nature (2012): Introduced METABRIC dataset and molecular subtypes of breast cancer.
* Pereira et al. , Nature Communications (2016): Improved classification of breast cancer subgroups by integrating clinical and genomic data using deep learning-based survival models to demonstrate promising performance of machine learning in oncology research.

**Minimum Viable Product**

1. A trained machine learning model. I will compare XGBoost with a survival loss function, neural networks (e.g., DeepSurv), and Random Survival Forests to predict survival probabilities and stratify patients into risk categories.
2. An interactive web application that:

* Accepts clinical and genomic inputs
* Displays survival probability curves
* Identifies top contributing risk factors using interpretability tools
* Stratifies patients into low-, medium-, and high-risk categories

**Schedule**

1. Get the Data (5/1/2025)
2. Clean & Explore the Data (5/30/2025)
3. Create Presentation (6/15/2025)
4. Internal Demos (6/28/2025)
5. Graduation (7/3/2025)
6. Demo Day (7/10/2025)

**Data Sources**

METABRIC (Molecular Taxonomy of Breast Cancer International Consortium), The data contains gene expression, clinical data, copy number gene variations and survival outcomes of breast cancer patients. This dataset was originally published in Nature journal in 2012 and Nature communication journal in 2016.

**Known Issues and Challenges**

* + - 1. How can I handle missing or incomplete clinical/genomic data without compromising model accuracy or fairness?
      2. What are the ethical implications of deploying predictive models in clinical settings, particularly when informing patients of high-risk classifications?
      3. How can dimensionality reduction or feature selection techniques (e.g., PCA, LASSO) be effectively applied to reduce noise and highlight meaningful genomic predictors?
      4. How can model interpretability be improved to ensure clinicians understand and trust the model’s predictions?