|  |  |
| --- | --- |
| **PROJECT PROPOSAL:**  **BREAST CANCER CLASSIFICATION** | Nay Zaw Lin  Meyliani Sanjaya  Javier Merino  Angeli De Los Reyes |

OBJECTIVES

The goal of this study is to:

* Examine the differences in gene expression across various stages of breast cancer.
* Predict the likelihood of survival for breast cancer patients based on gene expression.
* Identify key biomarkers for the early detection of breast cancer.
* Explore gene-gene interactions and their role in cancer progression.

DATA SOURCE

The dataset is sourced from TCGA human tumor data (transcriptomes) and is focused solely on breast cancer (BRCA) samples from cancer patients. We will be using mRNA-seq gene expression profiles of TCGA breast tumors, which are freely available in the TCGA database in TSV format. It consists of 1,231 breast tissue samples from 1,095 unique patients, with each sample containing 60,660 gene expression measurements. These measurements represent mRNA-seq counts, where higher counts indicate greater gene expression levels.

PROJECT TIMELINE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Task | Deliverable | Due | Type |
| 1 | **Data Collection:** Profile the mRNA-seq gene expression data and clinical information for breast cancer samples from TCGA, integrating them into a single data frame | Report (with visuals) | March 23, 2025 | Midterm 2 |
| 2 | **Data Preprocessing:** Explore 3 normalization techniques for gene expression data: RPKM, FPKM, and TPM. After normalization, perform EDA to better understand the data distribution, outliers, and key characteristics, focusing on the behavior of genes across different cancer stages |
| 3 | **Feature Selection & Feature Engineering:** Apply dimensionality reduction techniques (e.g., PCA, LDA, PCC) for feature selection and investigate transcription factor activities for model input | Report + Presentation (i.e., infographic or slides) | March 30, 2025 | Teamwork |
| 4 | **Model Selection & Training:** Train multiple models to predict cancer severity/stage (e.g., Ridge, RF, MLP, XGBoost, and SVM) and compare their performance using cross-validation |
| 5 | **Model Evaluation & Optimization:** Evaluate models using performance metrics (e.g., accuracy, AUC, F1-score) and optimize the selected model using techniques like hyperparameter tuning or grid search to improve performance |
| 6 | **Create a Survival Analysis Dashboard**: Develop a survival analysis dashboard leveraging **Gradio/Tableau** for the interface, **PostgreSQL** for data storage, and **Apache** **Airflow** to schedule and run the ML model for predicting survival probabilities of breast cancer patients | Report (with visuals) + Working application | April 15, 2025 | Finals |

PROJECT PIPELINE

