THE ERDŐS INSTITUTE
Revolutionary Collaborations in
Academia and Industry

Predicting Cancer Survivability through Patient Features

Ray Lee, Samuel Ogunfuye, Karla Paulette Flores Silva,
Dilruba Sofia, Funmilola Mary Taiwo, Enayon Sunday Taiwo

Introduction

Problem

- **Breast Cancer**:
  - Most frequently diagnosed and leading cause of cancer death among females.
  - About 2.3 million new cases and 670,000 death globally in 2022 (WHO).

Goal

- Develop a data-driven robust model for classifying breast cancer patient outcomes as "survive" or "deceased."
- Possible adoption and integration of the developed model to benefit stakeholders.

Stakeholders

- Healthcare providers
- Patients
- Insurance companies
- Government institutions

Data Gathering and Cleaning

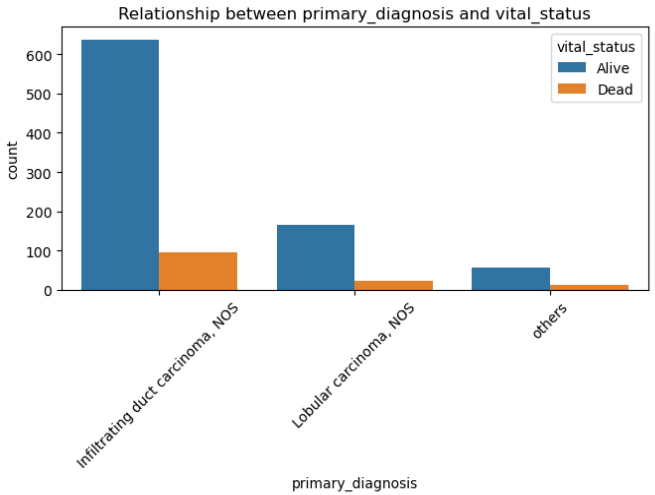
Data

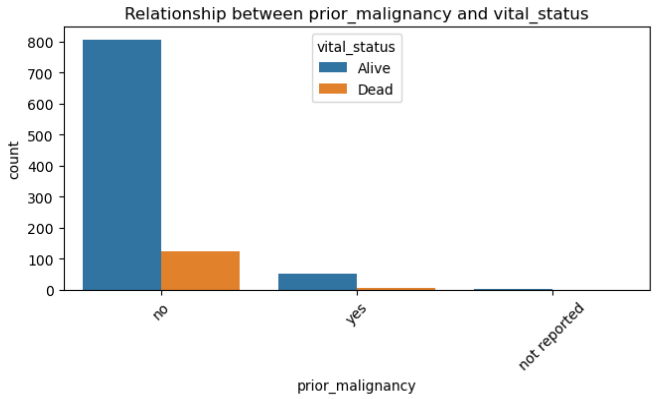
- Clinical and biospecimen datasets on breast cancer from the GDC data portal
- Size: 1006 observations X 42 variables
- Patient demographics, diagnoses, and treatment types from diagnoses
- Drop redundant and noninformative columns, and missing values

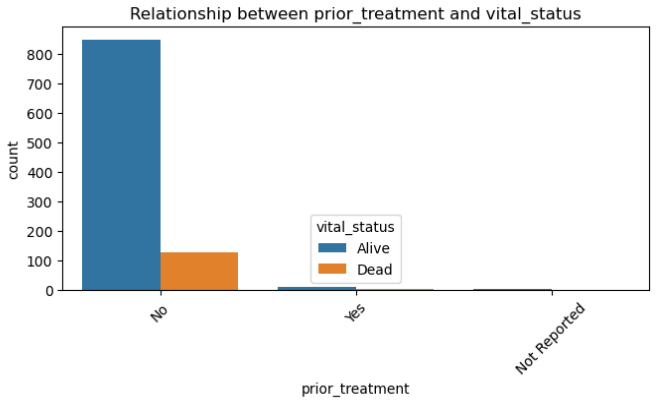
Final dataset

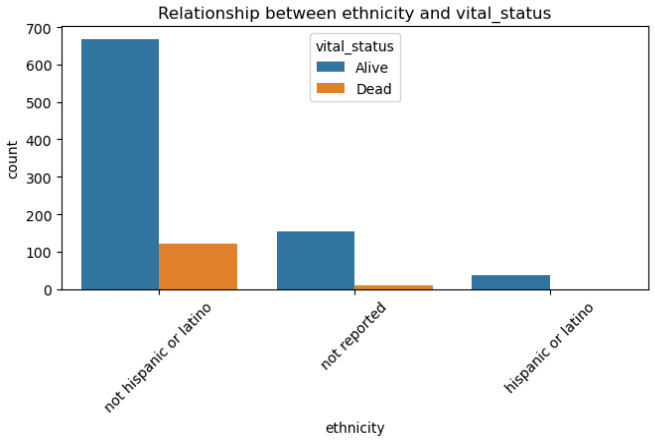
- ***Response variable***: vital status - "Alive" or "Dead"
- ***Independent variables***: Pathologic stages, prior treatment, treatment type - radiation or pharmaceutical treatment, ethnicity, race
- Size: 991 observations X 12 variables

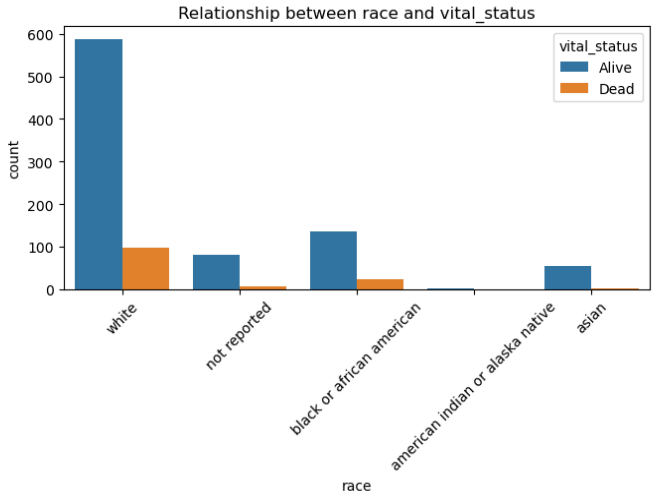
Categorical Features and Vital Status



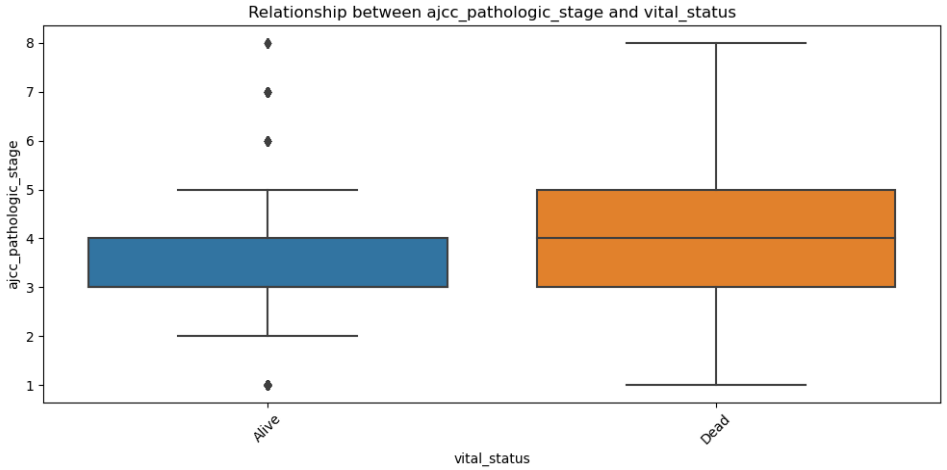


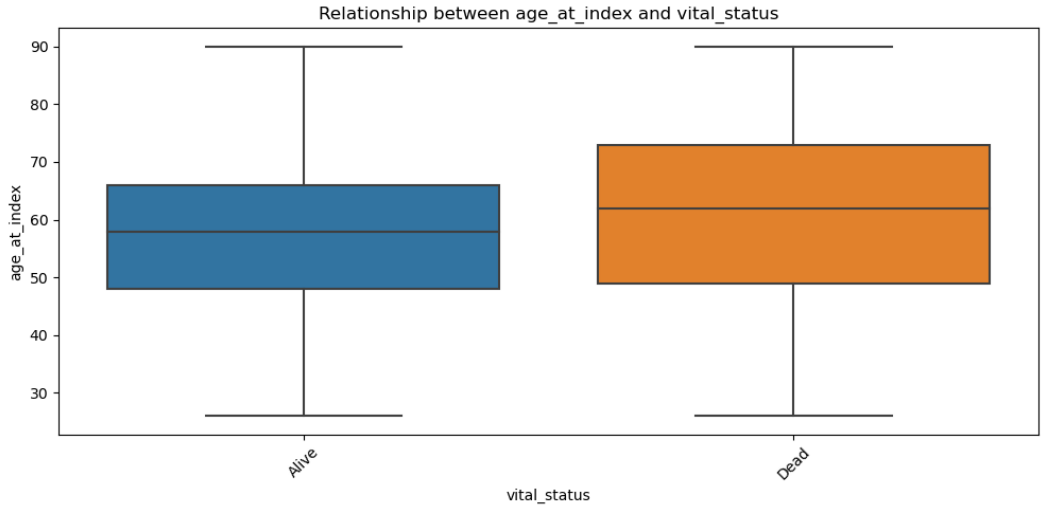






Continuous Features and Vital Status





There exist some relationships between vital status and pathologic stages and age.

Data Split and Models

Data Splitting: 80% training set and 20% test set (while stratifying the outcome). We further split the training data into a training set and validation set using StratifiedKFold (K=3).

Base Model: Decision Tree Classifier with an average validation accuracy of 84.85%.

Other Models:

- Random Forest Classifier (RF)
- AdaBoost Classifier
- Support Vector Classifier (SVC)
- Logistic Regression
- K-Nearest Neighbor
- Multi-Layer Perceptron (MLP)

Results: Classifying patient outcome

We evaluated our models based on their **accuracy**, **precision**, **sensitivity**, **selectivity**, **false positive rate**, and **false negative rate**, with respect to the validation sets.

We were particularly interested in:

- **Accuracy:** The model correctly classifies patient outcome.
  - Models were similarly accurate (~89-90%), and more accurate than the base model
- **Sensitivity (true positive rate):** the model correctly identifies 'deceased' outcomes as 'deceased'
  - Best model: Decision tree (53.33%), Multilayer perceptron (45.72%), Random Forest (41.9%)

Low sensitivity and high false negative rates in general imply limitations in our approach.

Random Forest: Hyperparameter Tuning

Fine-tune with grid search for max depth and number of estimators.

Choice of Hyperparameters

- Maximum depth: 10
- Number of estimators: 100

Model performance: Applying the model to the test set, we obtain:

- **Accuracy:** 92.46%
- **Precision:** 82.35%
- **Sensitivity:** 53.85%
- **False Negativity:** 46.15%

Results: Classifying patient outcome over a time interval

Goal: Consider a time frame for survival prediction.

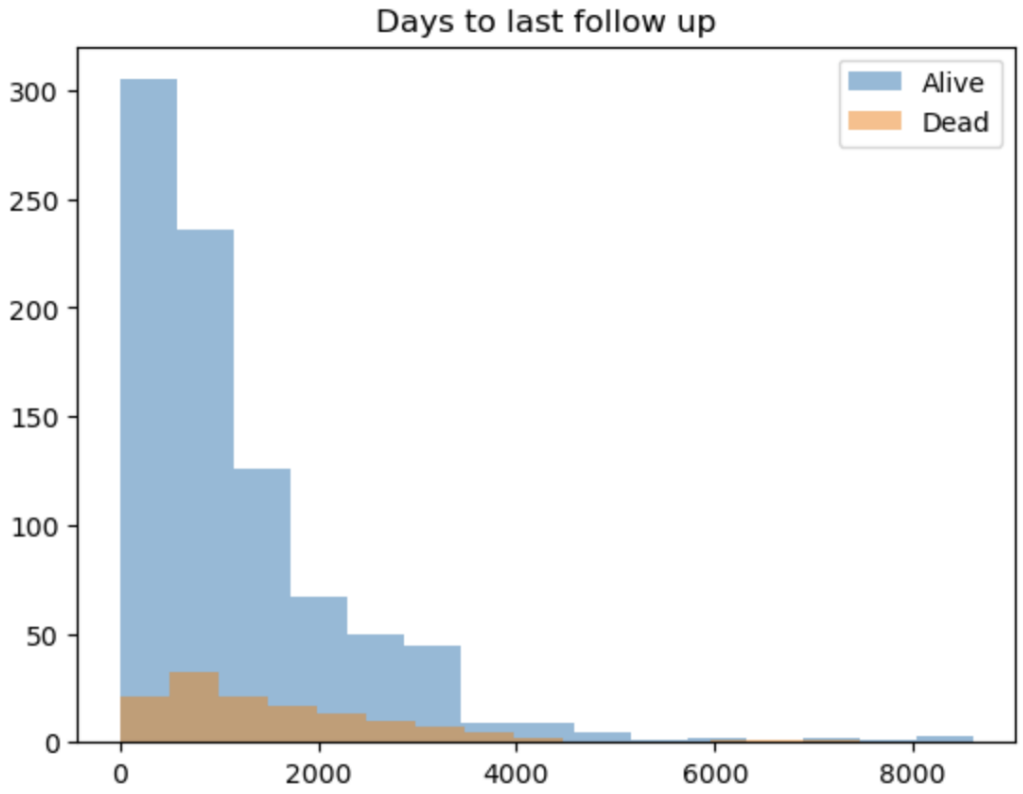
- Models: Random Forest, SVM.

- Mean of days of follow up:
  - Alive: ~ 1,200
  - Dead: ~ 1,600

- Number of days: 365, 720, ..
- Results:
  - Best model SVM

Prediction of survival within:

- 1 year - accuracy: 0.97, precision = 0.00
- 2 years - accuracy: 0.94, precision = 0.00
- 3 years - accuracy: 0.81, precision = 0.00
- 4 years - accuracy: 0.57, precision = 0.60
- 5 years - accuracy: 0.86, precision = 0.86



Discussion

Significance of results

- **Healthcare providers & patients**: Helps support physician decision-making and effective care delivery
- **Insurance companies**: Support pricing and reimbursement decisions
- **Government institutions**: Helps support resource allocation planning

Future directions

- The features in our model are limited. Other data that may be useful are gene expression data, information about immune cell abundance, and biopsy slide images.
- Explore more advanced algorithms
- Develop a web-based interface model to predict cancer survivability

Thank you!

To all the mentors, and instructors at The Erdos Institute!