Issues in Information Systems
Volume 22, Issue 2, pp. 296-305, 2021

DOI: https://doi.org/10.48009/2_iis_2021_312-321

Breast cancer survival analysis using ML models

Parthasarathi Tumu, *University of St Thomas, part6550@stthomas.edu*
Vaghindra Manchenasetty, *University of St Thomas, vaghi.manchenasetty@stthomas.edu*
Manjeet Rege, *University of St Thomas, rege@stthomas.edu*

Abstract

The main purpose of this study is to use machine learning techniques that have potential to predict breast cancer survival more accurately and can prevent unnecessary surgical treatment procedures. Using these machine learning models on genetic data has the potential to improve our understanding of cancers and survival prediction. The dataset is taken from 'The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)' database which contains the targeted sequencing of genomic data of primary breast cancer patients. The dataset is run through various machine learning algorithms such as Logistic Regression, KNN, Decision Tree, Random Forest, Extra Trees, Adaboost, Support Vector Machines (SVM) to predict the survival status. The results are compared to the traditional survival analysis models. This study aims to showcase that machine learning models would help accurately predict the outcomes of breast cancer prognosis, which will help survival chances by providing treatment procedures in early stages of the cancer.

Keywords Survival Analysis, Survival status, Breast Cancer, Cancer Type, KNN, Logistic Regression, Decision Tree, Random Forest, AdaBoost, SVC

Introduction

All of us at some point in our lives, have met someone who fought Breast Cancer, or at least heard about the struggles that are faced by many patients who are fighting against breast cancer. Breast cancer is one of the leading causes for deaths in women historically. In 2018, the United States had more than 40,000 women that died from Breast Cancer and it impacts around 2.1 million women worldwide each year.

Traditional survival prognostication of breast cancer patients is a complex task clinically due to the similarities of breast cancer with other breast tumors and they may exhibit different clinical outcomes. The traditional methods of survival analysis, frequently utilized to study survival time, they are the distribution-free estimator of Kaplan-Meier and the proportional hazards models such as Cox's. The first gives us the estimate of survival operations, and the second allows the evaluation of covariates on hazards ratio. Though traditional mathematical survival analysis models are considered reliable, they failed to be established as the major tools for reliability analysis. In this paper we aim to use machine learning techniques that has potential to predict breast cancer survival status more accurately and can avert unnecessary complications. Using these machine learning models on genetic data has the potential to improve our understanding of cancers and survival prediction.

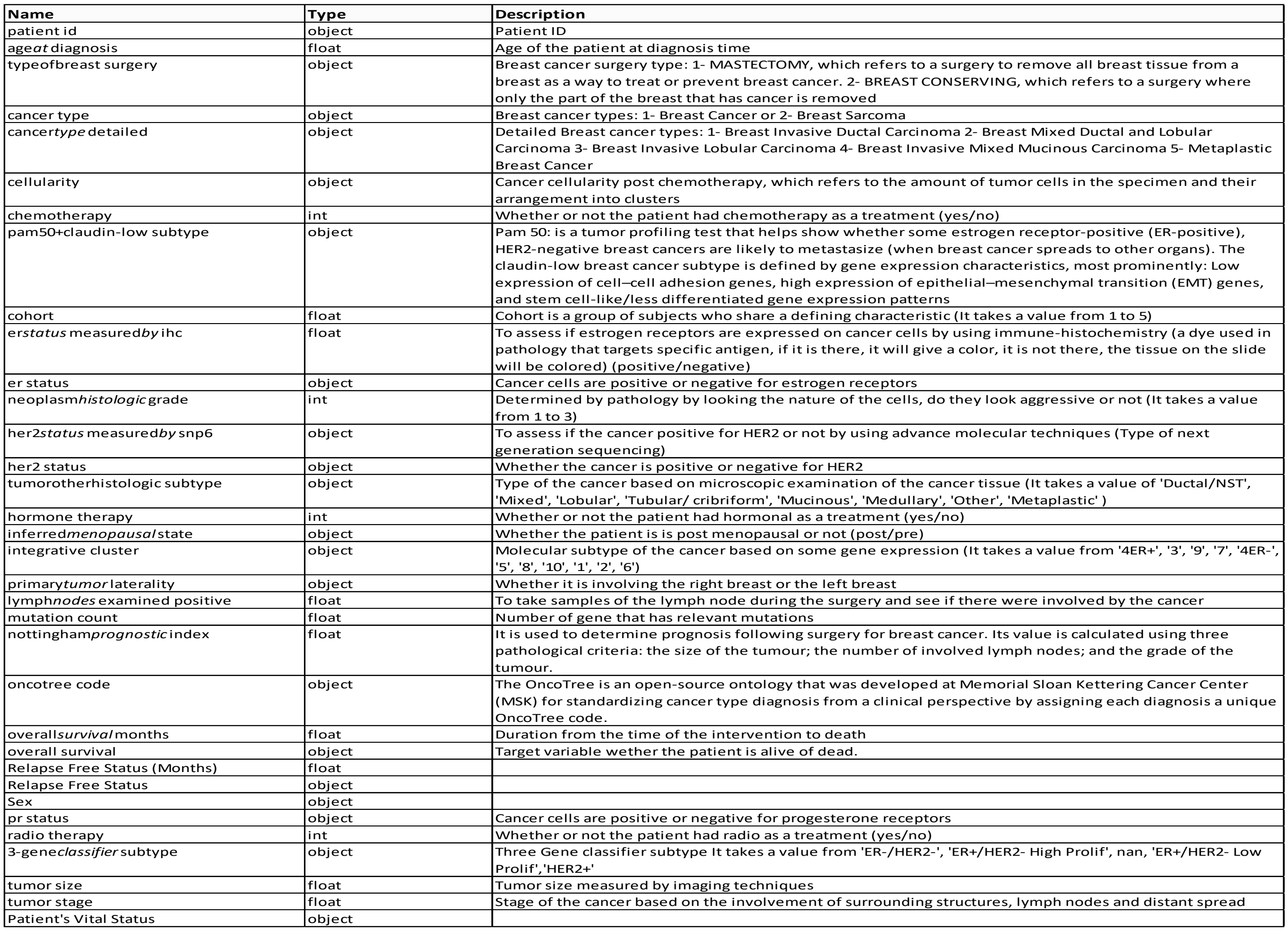
This paper explores machine learning algorithms to predict the outcome of an event (Survival Status). The generated models have the goal of predicting the event information better than traditional survival analysis models.

Research Methodology

Data

There are 2,509 (Table-1) unique breast cancer patients in METABRIC dataset as mentioned above. Those patients are diagnosed at ages between 21.9-96.3 and their mean diagnosis age is 60.4. Patients had undergone to two different surgeries; Mastectomy (removal of all breast tissue from a breast) or Breast-conserving surgery (removal of a part of the breast that has cancer). There are 2,506 breast cancer and 3 breast sarcoma patients in the dataset since breast sarcomas are a very rare form of breast cancer, that cover fewer than 1% of all breast cancers. The most common histological subtype of the breast cancer is invasive ductal carcinoma (IDC) with 1865 occurrences. IDC is the most common form of breast cancer, representing 80% of all breast cancer diagnoses. Those indicators show that this dataset reflects real world scenarios very accurately. Refer Table-1

Table1: Dataset attributes and Type

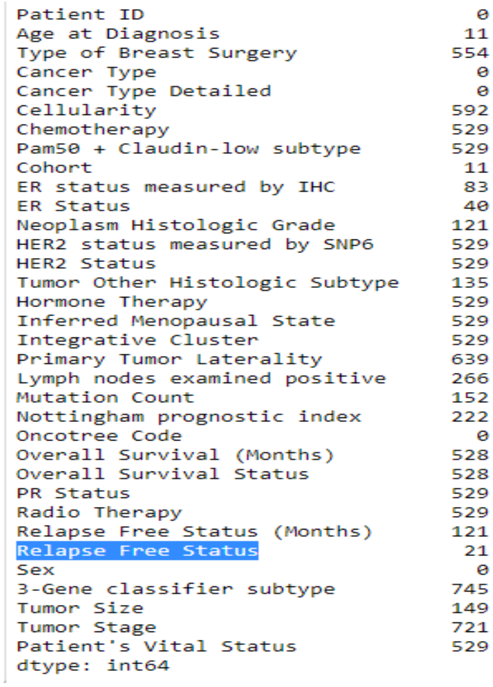


Name	Type	Description
patient id	object	Patient ID
age*at* diagnosis	float	Age of the patient at diagnosis time
typeofbreast surgery	object	Breast cancer surgery type: 1- MASTECTOMY, which refers to a surgery to remove all breast tissue from a breast as a way to treat or prevent breast cancer. 2- BREAST CONSERVING, which refers to a surgery where only the part of the breast that has cancer is removed
cancer type	object	Breast cancer types: 1- Breast Cancer or 2- Breast Sarcoma
cancer*type* detailed	object	Detailed Breast cancer types: 1- Breast Invasive Ductal Carcinoma 2- Breast Mixed Ductal and Lobular Carcinoma 3- Breast Invasive Lobular Carcinoma 4- Breast Invasive Mixed Mucinous Carcinoma 5- Metaplastic Breast Cancer
cellularity	object	Cancer cellularity post chemotherapy, which refers to the amount of tumor cells in the specimen and their arrangement into clusters
chemotherapy	int	Whether or not the patient had chemotherapy as a treatment (yes/no)
pam50+claudin-low subtype	object	Pam 50: is a tumor profiling test that helps show whether some estrogen receptor-positive (ER-positive), HER2-negative breast cancers are likely to metastasize (when breast cancer spreads to other organs). The claudin-low breast cancer subtype is defined by gene expression characteristics, most prominently: Low expression of cell–cell adhesion genes, high expression of epithelial–mesenchymal transition (EMT) genes, and stem cell-like/less differentiated gene expression patterns
cohort	float	Cohort is a group of subjects who share a defining characteristic (It takes a value from 1 to 5)
er*status* measured*by* ihc	float	To assess if estrogen receptors are expressed on cancer cells by using immune-histochemistry (a dye used in pathology that targets specific antigen, if it is there, it will give a color, it is not there, the tissue on the slide will be colored) (positive/negative)
er status	object	Cancer cells are positive or negative for estrogen receptors
neoplasm*histologic* grade	int	Determined by pathology by looking the nature of the cells, do they look aggressive or not (It takes a value from 1 to 3)
her2*status* measured*by* snp6	object	To assess if the cancer positive for HER2 or not by using advance molecular techniques (Type of next generation sequencing)
her2 status	object	Whether the cancer is positive or negative for HER2
tumorotherhistologic subtype	object	Type of the cancer based on microscopic examination of the cancer tissue (It takes a value of 'Ductal/NST', 'Mixed', 'Lobular', 'Tubular/ cribriform', 'Mucinous', 'Medullary', 'Other', 'Metaplastic')
hormone therapy	int	Whether or not the patient had hormonal as a treatment (yes/no)
inferred*menopausal* state	object	Whether the patient is is post menopausal or not (post/pre)
integrative cluster	object	Molecular subtype of the cancer based on some gene expression (It takes a value from '4ER+', '3', '9', '7', '4ER-', '5', '8', '10', '1', '2', '6')
primary*tumor* laterality	object	Whether it is involving the right breast or the left breast
lymph*nodes* examined positive	float	To take samples of the lymph node during the surgery and see if there were involved by the cancer
mutation count	float	Number of gene that has relevant mutations
nottingham*prognostic* index	float	It is used to determine prognosis following surgery for breast cancer. Its value is calculated using three pathological criteria: the size of the tumour; the number of involved lymph nodes; and the grade of the tumour.
oncotree code	object	The OncoTree is an open-source ontology that was developed at Memorial Sloan Kettering Cancer Center (MSK) for standardizing cancer type diagnosis from a clinical perspective by assigning each diagnosis a unique OncoTree code.
overall*survival* months	float	Duration from the time of the intervention to death
overall survival	object	Target variable wether the patient is alive of dead.
Relapse Free Status (Months)	float	
Relapse Free Status	object	
Sex	object	
pr status	object	Cancer cells are positive or negative for progesterone receptors
radio therapy	int	Whether or not the patient had radio as a treatment (yes/no)
3-gene*classifier* subtype	object	Three Gene classifier subtype It takes a value from 'ER-/HER2-', 'ER+/HER2- High Prolif', nan, 'ER+/HER2- Low Prolif','HER2+'
tumor size	float	Tumor size measured by imaging techniques
tumor stage	float	Stage of the cancer based on the involvement of surrounding structures, lymph nodes and distant spread
Patient's Vital Status	object	

Data Cleaning and Validation

There are 29 columns (Fig1) with missing values and only 5 columns don't have missing values in them. Those columns are Sex, Cancer Type Detailed, Cancer Type, Oncotree Code, and Patient ID. Cancer Type and Cancer Type Detailed can be very useful for imputation, but others don't yield any information.

Dependencies between different columns are explored during the imputation stage. Missing values in event columns are filled with their most common value and missing values in duration columns are filled with most common values of Cancer Type Detailed, Event groups. Missing values in ER, PR and HER2 Status columns are filled with the most common values of their measurement technique columns (ER status measured by IHC and HER2 status measured by SNP6). Missing values in Chemotherapy, Hormone therapy, and Radio therapy are filled with the most common values in Cancer Type Detailed groups. Missing values in other columns are filled with modes or medians of different groups based on their dependencies. Some of the columns couldn't be filled with one operation so they are filled iteratively. Finally, Patient's Vital Status column is dropped because it doesn't yield any extra information for the analysis.



```
Patient ID                           0
Age at Diagnosis                    11
Type of Breast Surgery             554
Cancer Type                          0
Cancer Type Detailed                 0
Cellularity                        592
Chemotherapy                       529
Pam50 + Claudin-low subtype        529
Cohort                              11
ER status measured by IHC           83
ER Status                           40
Neoplasm Histologic Grade          121
HER2 status measured by SNP6       529
HER2 Status                        529
Tumor Other Histologic Subtype     135
Hormone Therapy                    529
Inferred Menopausal State          529
Integrative Cluster                529
Primary Tumor Laterality           639
Lymph nodes examined positive      266
Mutation Count                     152
Nottingham prognostic index        222
Oncotree Code                        0
Overall Survival (Months)          528
Overall Survival Status            528
PR Status                          529
Radio Therapy                      529
Relapse Free Status (Months)       121
Relapse Free Status                 21
Sex                                  0
3-Gene classifier subtype          745
Tumor Size                         149
Tumor Stage                        721
Patient's Vital Status             529
dtype: int64
```

Figure1: Input features with number of missing values

Data Analysis and Visualization

Data exploration was performed with the help of various pre-processing plots as part of the experimental analysis and would be helpful for the analysis and draw conclusions with these visual readings. The plots that were used in this experiment were correlation matrix to show influence of each feature with other. Other plots included box plot for outlier detection and heatmap to get correlation between input features.

For the distribution of all numerical data (fig2), some of them are normally distributed (like tumor_stage, and age_at_diagnosis), but most of the features are right skewed with a lot of outliers (lymph_nodes_examined_positive, mutation_count, and tumor_size). We decided to keep the outliers, as they are very important in healthcare data.

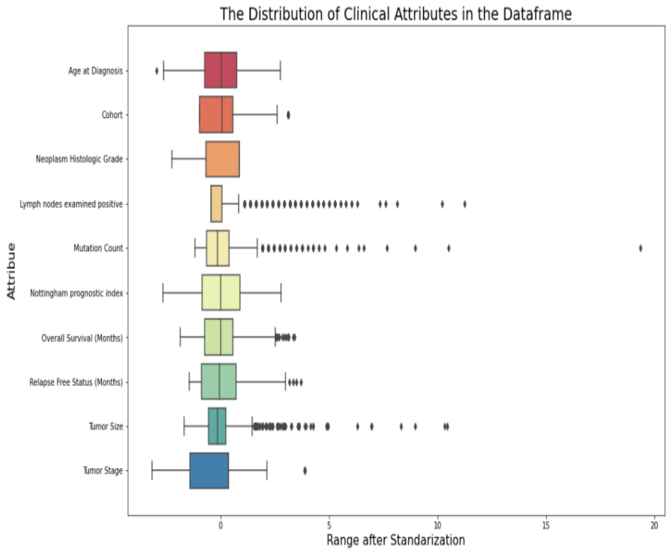


Figure2: Box plot

To compare between the two classes of patients (fig3) between who survived and patients who did not, we can see the difference between the two distributions in "age at diagnosis" column, as patients who were younger when diagnosed with breast cancer were more likely to survive. Also, the period from the time of the intervention to death or to current time is longer in the patients who survive. That means that patients are either dying early from breast cancer or surviving.

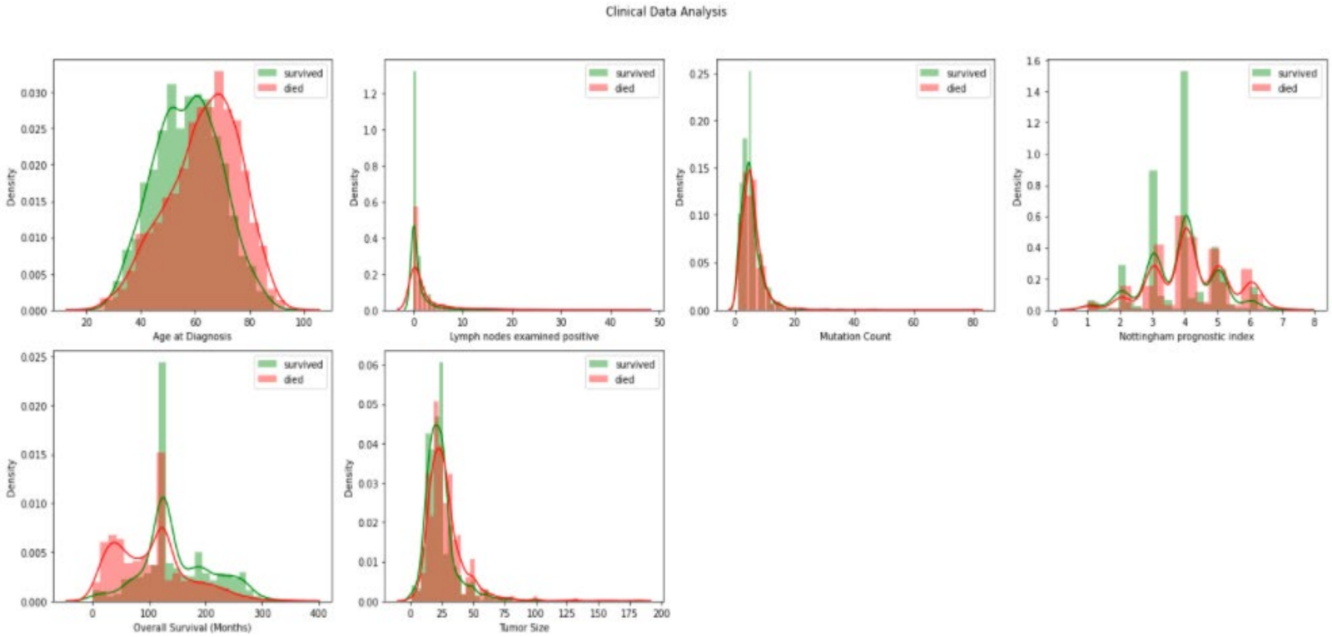


Figure 3: Clinical Data Analysis

In fig4 data shows that, patients with any type of treatment (Chemo, Hormonal, Radio) options has better survival rate than without a treatment.

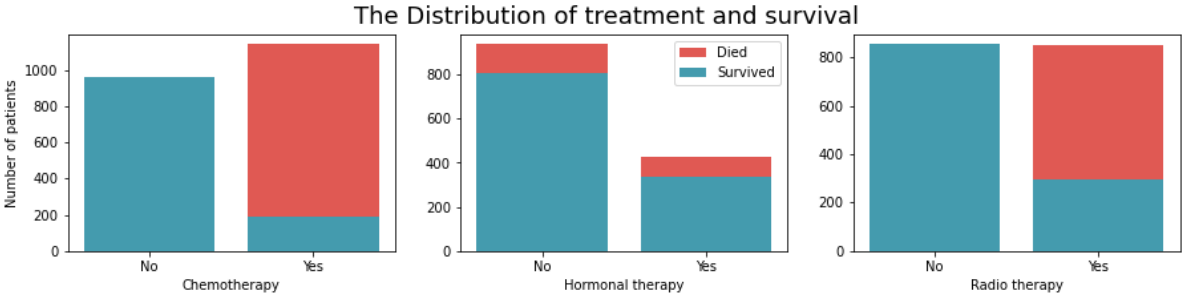


Figure 4: The Distribution of Treatment and Survival

```
Average Patient profile:
Mean age: 60.420
Most occurring tumor stage:  2
Most occurring histopathological type:  3
Mean tumor diameter: 26.220
Probability of survival: 0.002
```

In fig5 Venn diagram for the three different treatments for breast cancer and the distribution of patients amongst them. Most patients either have chemo and hormonal or chemo and radio therapy combinations. There is a group that is not shown here in the diagram, which are the patients that did not receive any of the three treatments. they were 311 patients, and their survival rate was slightly lower than the rest of patients.

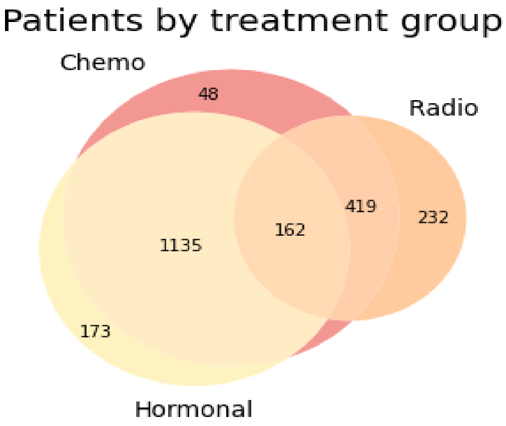
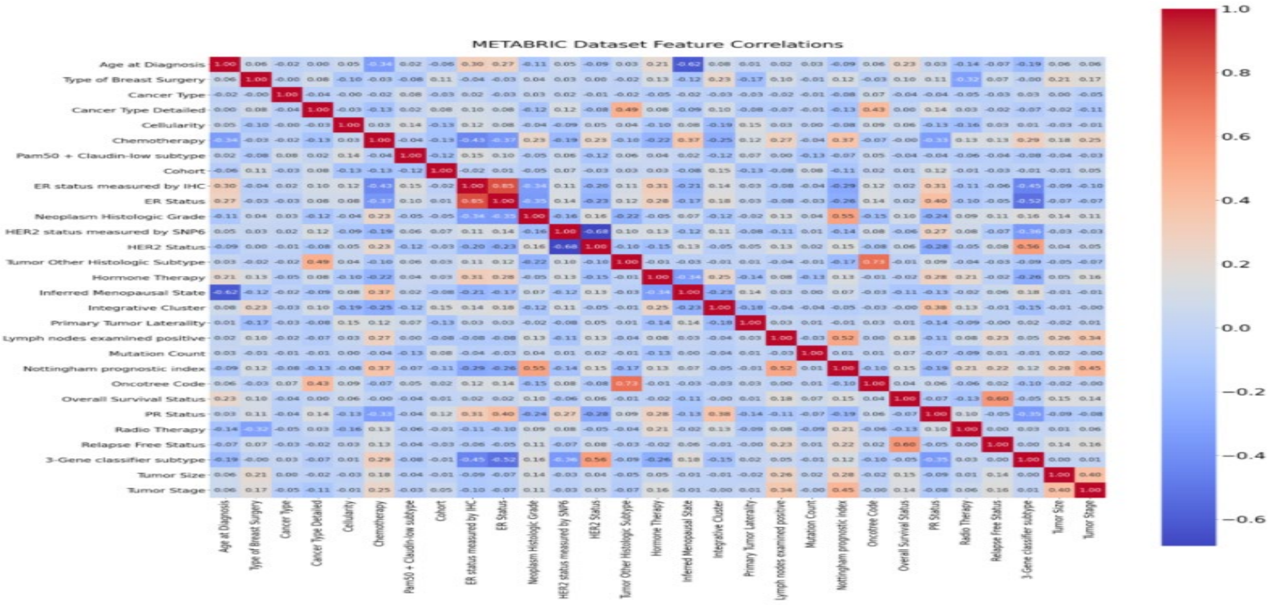


Figure 5: Patients by Treatment Group

Data Correlation

There are some strong relationships between attributes with above 0.6 and below -0.6 (refer Fig6) indicating lots of features are dependent to each other. A positive correlation exists between Cellularity and Integrative Cluster, so one of them is dropped. Though there is a strong relationship exist between two attributes, Overall Survival Status is dependent Relapse Free Status. Those features shouldn't be used as covariates because of this reason.



Results

Our goal is to predict the Survival status of a given patient with clinical data. From the data Analysis, Input features ‘Sex’, ‘Patient ID’, ‘Patient's Vital Status’ hold no significance on patient survival status. So, these features are dropped from input dataset.

Input data set is split in to two datasets training (67%) and test (33%). Target Variable ‘Overall Survival status’ (Living/Deceased) is binary encoded. A Label encoding process is applied on all other categorical features.

A set of 7 classification models used to predict the Variable ‘Overall Survival status’.Fig7 shows the summary of 7 models and accuracy score . From fig7 and fig8 it is clear that Random forest and SVC models performed best.

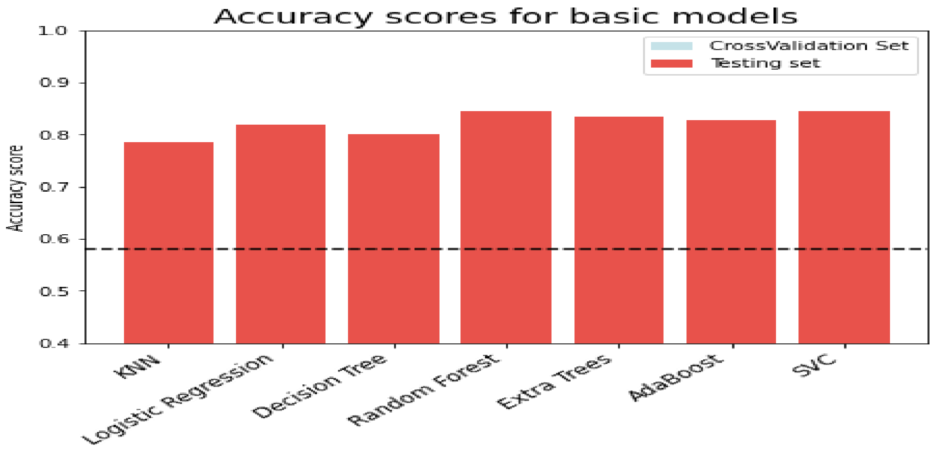


Figure: 7

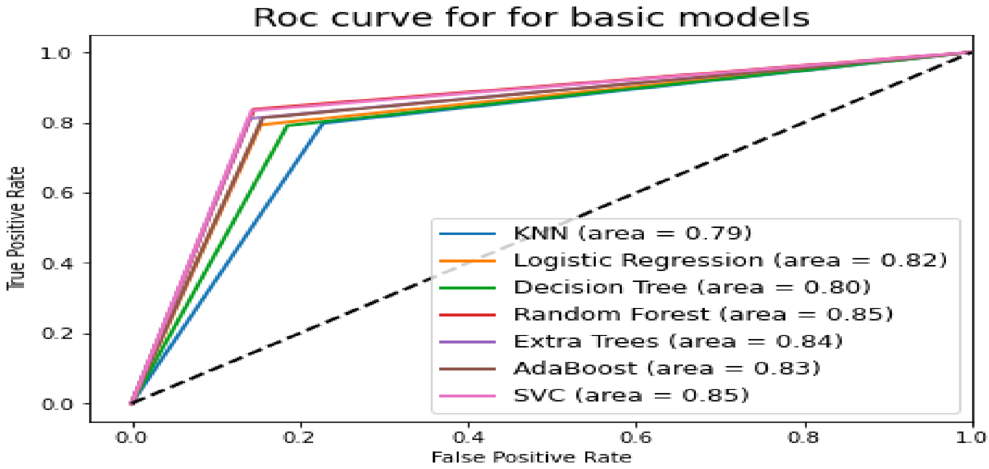


Figure: 8

Traditional Survival models

Traditional survival estimation processes are classified into non-parametric, semi-parametric, and parametric methods. In general, the popular non-parametric method is used to depict survival probabilities as function of time and to give a middling view of individual's population. Kaplan-Meier estimator is a univariate method.

The main notion of the Kaplan-Meier estimator is to split the estimation of the survival function S(t) into smaller steps depending on the observed time events. For each interval, the probability of surviving until the end of that interval is calculated by using formula:

$$S\hat{(t)} = \prod_{i:t_i<=t} \frac{n_i - d_i}{n_i}$$

Kaplan-Meier estimate can also be used for predicting survival probabilities of unseen data at given times. Kaplan-Meier Fitter Predict function takes times parameter which is a list of timesteps to predict probabilities. For example, the model below tries to predict both events happening (separately) at 12, 24, and 36 months. Model is evaluated on 5 split cross-validation and unseen test set.

From Fig9 It can be seen that probability of event not happening is close to 1 at the start of the study and decreases to 0 over time. Estimates closer to start of the study are more confident while estimates closer to end of the study are less confident since it is harder to forecast future.

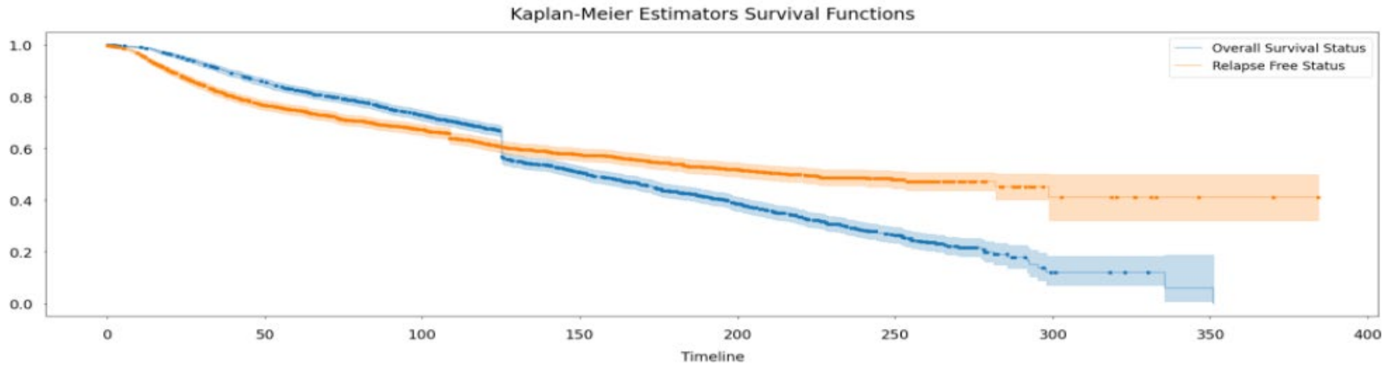
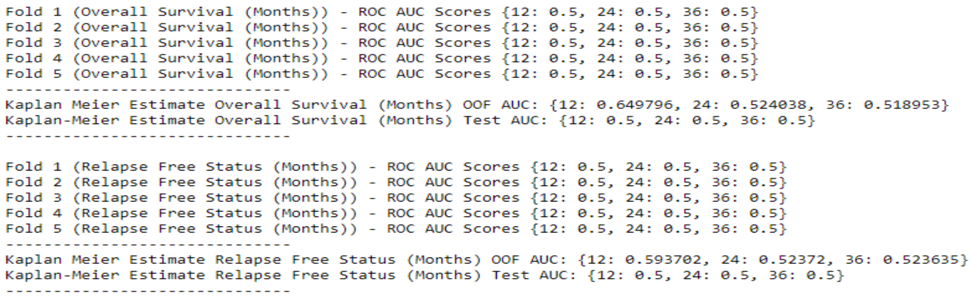


Figure: 9

From Fig10 Kaplan-Meier model returns 0.5 Test ROC AUC scores suggest that Kaplan-Meier estimate fails to generalize on an unseen test set. This was expected because this model doesn't use patient covariates and it outputs same probabilities for every patient in the population. That's why Kaplan-Meier estimate shouldn't be used for predictions, but it is very useful for exploratory data analysis.



```
Fold 1 (Overall Survival (Months)) - ROC AUC Scores {12: 0.5, 24: 0.5, 36: 0.5}
Fold 2 (Overall Survival (Months)) - ROC AUC Scores {12: 0.5, 24: 0.5, 36: 0.5}
Fold 3 (Overall Survival (Months)) - ROC AUC Scores {12: 0.5, 24: 0.5, 36: 0.5}
Fold 4 (Overall Survival (Months)) - ROC AUC Scores {12: 0.5, 24: 0.5, 36: 0.5}
Fold 5 (Overall Survival (Months)) - ROC AUC Scores {12: 0.5, 24: 0.5, 36: 0.5}
------------------------------
Kaplan Meier Estimate Overall Survival (Months) OOF AUC: {12: 0.649796, 24: 0.524038, 36: 0.518953}
Kaplan-Meier Estimate Overall Survival (Months) Test AUC: {12: 0.5, 24: 0.5, 36: 0.5}
------------------------------

Fold 1 (Relapse Free Status (Months)) - ROC AUC Scores {12: 0.5, 24: 0.5, 36: 0.5}
Fold 2 (Relapse Free Status (Months)) - ROC AUC Scores {12: 0.5, 24: 0.5, 36: 0.5}
Fold 3 (Relapse Free Status (Months)) - ROC AUC Scores {12: 0.5, 24: 0.5, 36: 0.5}
Fold 4 (Relapse Free Status (Months)) - ROC AUC Scores {12: 0.5, 24: 0.5, 36: 0.5}
Fold 5 (Relapse Free Status (Months)) - ROC AUC Scores {12: 0.5, 24: 0.5, 36: 0.5}
------------------------------
Kaplan Meier Estimate Relapse Free Status (Months) OOF AUC: {12: 0.593702, 24: 0.52372, 36: 0.523635}
Kaplan-Meier Estimate Relapse Free Status (Months) Test AUC: {12: 0.5, 24: 0.5, 36: 0.5}
------------------------------
```

Figure 10

Summary

Using machine learning models on a clinical data has the potential to improve our understanding of cancers and survival prediction. Huge open-source datasets are available for the public to analyze and hopefully, get some insights. The model with the best performance was Random Forest and SVC. To enhance this project, increase the number of samples, include mutations and raw genetic data into the modeling part, and maybe try some deep learning models.

The advanced study of survival analysis using ML models gives more procedural freedom. With right hyperparameter tuning methods it is likely to attain more accurate predictions of the time-to-event target variable. The format of the dataset is very crucial. To apply survival analysis techniques, the data must meet the requirements of the characteristic survival analysis data points: event, duration and valuable feature

References

K. S. Bhangu, J. K. Sandhu and L. Sapra, "Improving diagnostic accuracy for breast cancer using prediction-based approaches," *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, Waknaghat, India, 2020, pp. 438-441, doi: 10.1109/PDGC50313.2020.9315815.

P. Liu, B. Fu, S. X. Yang, L. Deng, X. Zhong and H. Zheng, "Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer," in IEEE Transactions on Biomedical Engineering, vol. 68, no. 1, pp. 148-160, Jan. 2021, doi: 10.1109/TBME.2020.2993278.

D. Kaushik, B. R. Prasad, S. K. Sonbhadra and S. Agarwal, "Post-Surgical Survival Forecasting of Breast Cancer Patient: A Novel Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, 2018, pp. 37-41, doi: 10.1109/ICACCI.2018.8554745.

B. Hela, M. Hela, H. Kamel, B. Sana and M. Najla, "Breast cancer detection: A review on mammograms analysis techniques," 10th International Multi-Conferences on Systems, Signals &

ZhongXin Ding, "The application of support vector machine in survival analysis," 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011, pp. 6816-6819, doi: 10.1109/AIMSEC.2011.6011384.

T. Lim, "Applying Survival Analysis for Customer Retention: A U.S. Regional Mobile Service Operator," 2020 Zooming Innovation in Consumer Technologies Conference (ZINC), 2020, pp. 338-342, doi: 10.1109/ZINC50678.2020.9161811.

Wenbin Zhang, Jian Tang and Nuo Wang, "Using the machine learning approach to predict patient survival from high-dimensional survival data," 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016, pp. 1234-1238, doi: 10.1109/BIBM.2016.7822695.

S. M. Husband and J. Roberts, "Survival Random Forest to Predict Time to Fill," 2017 IEEE International Conference on Data Mining Workshops (ICDMW), 2017, pp. 195-198, doi: 10.1109/ICDMW.2017.32.

B. Sansaengtham, V. C. Barroso and P. Phunchongharn, "Survival Analysis For Computing Systems Using A Deep Ensemble Network," 2020 IEEE 6th International Conference on Control Science and Systems Engineering (ICCSSE), 2020, pp. 57-62, doi: 10.1109/ICCSSE50399.2020.9171987.

Z. Ma and A. W. Krings, "Survival Analysis Approach to Reliability, Survivability and Prognostics and

Health Management (PHM)," 2008 IEEE Aerospace Conference, 2008, pp. 1-20, doi: 10.1109/AERO.2008.4526634.

STATISTICS FOR BIOLOGY AND HEALTH, Series Editors: Gail, Mitchell, Samet, Jonathan M. Publisher: SpringerLink