PROJECT REPORT
FOR EEE 4518

GROUP NAME: BELA BISKUT

GROUP MEMBERS

Aseer Imad Keats	ID: 190021208
Mohammad Abrar Kabir	ID: 190021207
Md. Saifur Rahman	ID: 190021133
Md. Mahfuzul Islam	ID: 190021113

Contents

Objective: **3**

Previous Literature works **3**

Dataset **4**

Proposed Approach **6**

Discussion & Result: **8**

Single Survival Function using Kaplan-Meier 8

Kaplan-Meier Estimation Considering the Attributes 8

Cox Proportional Hazard Model 14

Predicting risk scores of a few patients using the RSF model based on age and tumor size 15

Contribution 17

Objective:

1. Estimate breast cancer survival probability - We will establish some observable relationships between the dataset's attributes using various algorithms in order to anticipate the likelihood of survival following a diagnosis.
2. Analyze the time event dataset
3. Implementation of different survival models – We will implement basically Kaplan-Meier model, Cox regression and Random survival forest.
4. Guiding clinicians to make informed decisions - Clinicians can choose the appropriate level of therapy intensity by being aware of potential dangers. Finding low risk patients may help lessen patient worry and the extra expense of care.

Previous Literature works

Survival analysis is a statistical method for predicting the time to an event. The presence of censored data, indicating that the event of interest did not occur during the study period, is an important aspect of survival analysis. The presence of censored data necessitates the application of specialized techniques. Kaplan-Meier Survival Model, Cox proportional hazards model and Random Survival Forest Model are such specialized techniques.

In this section we review two previous literature works related to these survival models. The first one is “A Case Study on Risk Prediction in Heart Failure Patients using Random Survival Forest” and the second one is “Effect of Age on Breast Cancer Patient Prognoses”. The later one is based on the SEER breast cancer dataset.

Asif Newaz et al. in their research work created a risk prediction system for heart failure patients using the Random Survival Forest model. They worked with the Faisalabad Institute of Cardiology dataset. They first identified the features that are most important for heart failure patient survival prediction, and then they built the Random Survival Forest model to forecast the risk of heart failure patients and compared the results to the Cox Proportional Hazard model. Their RSF model outperforms the Cox Proportional Hazard model, which has a C-Index of 0.81.

Hai-long Chen et al. identify age as a crucial risk factor for breast cancer and presented a study that evaluated the effect of age on breast cancer prognosis. They used the Kaplan-Meier method and the Cox regression model to investigate the relationships between age and overall survival and breast cancer specific survival.

Dataset

This dataset of breast cancer patients was obtained from the 2017 November update of the SEER Program of the NCI, which provides information on population-based cancer statistics.

The dataset involved female patients with infiltrating duct and lobular carcinoma breast cancer (SEER primary cites recode NOS histology codes 8522/3) diagnosed in 2006-2010. Patients with unknown tumor size, examined regional LNs, regional positive LNs, and patients whose survival months were less than 1 month were excluded; thus, 4024 patients were ultimately included.

This dataset has total 4024 entries and 16 attributes.

There is no null value.

Event attribute - Status

Time attribute - Survival months

Age - This data item shows the patient's age at the time of diagnosis for this cancer. The code denotes the patient's true age in years.

Race – A race is a classification of people into groups that are typically considered as unique within a given civilization and is based on similar physical characteristics. There are 3 unique values for this attribute.

- White
- Black
- Others

Marital Status - This informational item reveals the patient's marital status at the time of the reportable tumor's diagnosis. There are 5 unique values for this attribute.

- Single
- Married
- Separated
- Divorced
- Widowed

T Stage - The T refers to the size and extent of the main tumor. There are 4 unique values for this attribute.

- T1 - No evidence of primary tumor
- T2 – Tumor is 2 cm or less across
- T3 - Tumor is more than 2 cm but not more than 5 cm (2 inches) in across
- T4 - Tumor is more than 5 cm in diameter

N Stage - The N stands for the number of cancerous lymph nodes nearby. There are 3 unique values for this attribute. The extent of lymph node involvement increases as the N number rises.

- N1
- N2
- N3

6th Stage - There are 5 unique values for this attribute. The extent of risk increases as the number rises.

- IIA
- IIB
- IIIA
- IIIB
- IIIC

Grade - A low grade score typically indicates that the cancer is less likely to spread and is growing more slowly. A high grade number indicates a malignancy that is spreading more quickly. There are 4 unique values for this attribute.

- Grade I
- Grade II
- Grade III
- Grade IV

Estrogen status - Early breast cancer patients who test positive for ER typically outlive those who test negative for ER.

- Positive
- Negative

Regional nodes examined - Notes the total number of regional lymph nodes that the pathologist analyzed.

Regional nodes positive - The precise number of local lymph nodes that the pathologist examined and discovered to have metastases.

Survival months – Was created using full dates, including days, therefore it can be different from the survival time computed using simply the year and the current month.

Status - Any patient who passes away after the follow-up cut-off date is counted as having been alive at that time and vice-versa.

- Alive
- Dead

Proposed Approach

In this work we basically introduced Kaplan Meier Estimator, Cox Proportional Model, and Random Survival Forest Model as a survivability prediction system for the Breast Cancer patients. Later we found the best model for this purpose.

First we encoded the “Status” column. This makes the column data from object type to int32 (numeric) type. Then we split the Time and Event data since these will be used for survival analysis. Now to make the data compatible with Scikit survival we convert the Time and Event data to structured array. Now, in this stage we took three different directions.

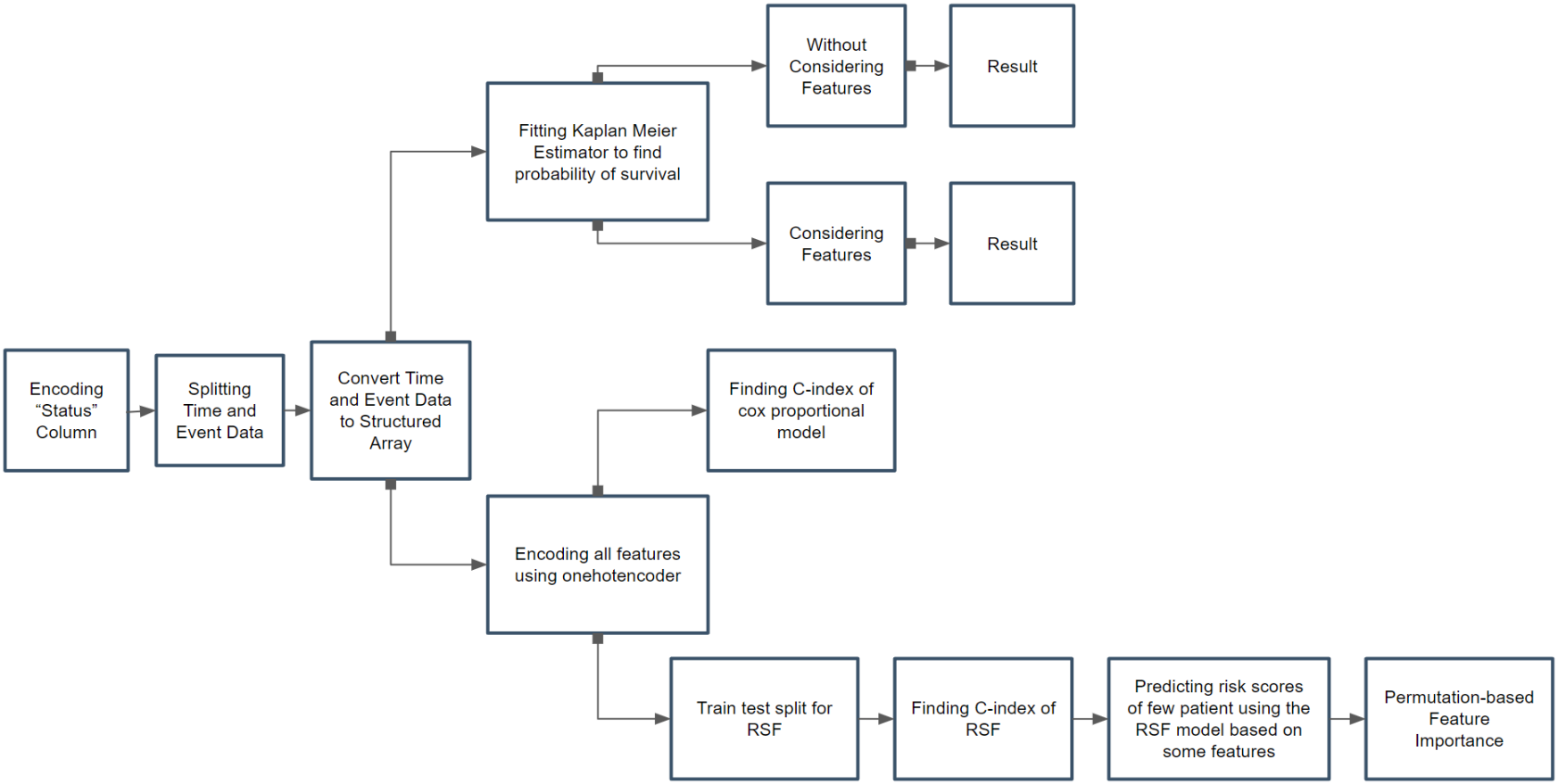
First we fitted the Kaplan-Meier estimator to find the probability of survival without considering any features. Then considering the features we found the survival probability. Kaplan-Meier estimates, however, are oversimplified because they ignore the impact of other variables on survival. Multivariate models, such as the Cox model, are therefore required for the precise calculation of risk.

So, now we head toward the Cox model and RSF model. For that first we change the object data to categorical data and then change the categorical data of the features to numerical type by using one hot encoding. Now we apply the Cox proportional hazard model and find the related C-index of the model.

For applying RSF model we first use the train test split technique then we apply the RSF model. After that we find the related C-index.

Along with the above works we took additional approach to find the risk scores of few random patient using RSF model base on some features. Then we applied the permutation-based feature importance to identify which feature play significant role in the Random Survival Forest Model.

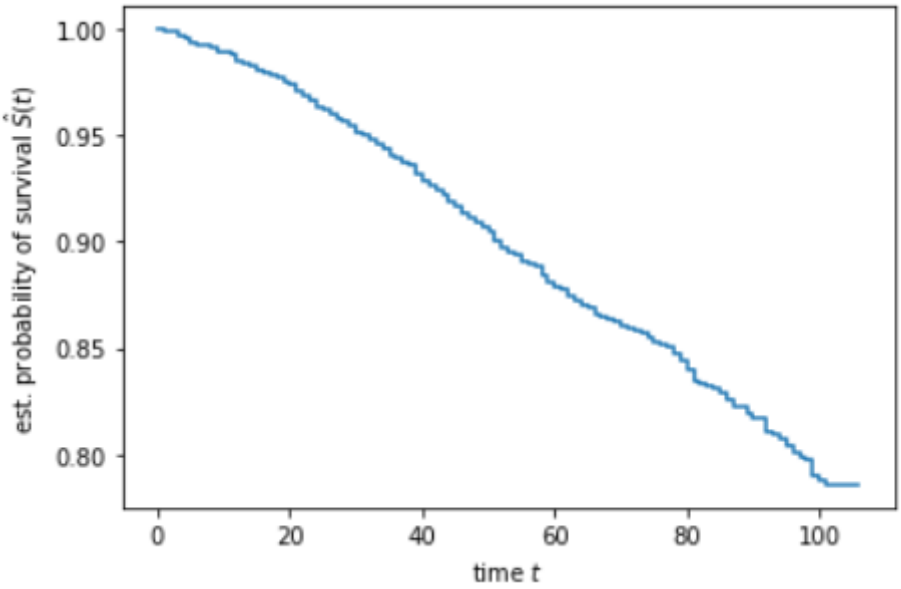
Implementation strategy



Discussion & Result:

We have received numerous graphs and output lists of our code after running the code. We will now assess our results and make an effort to align them with the project's objective.

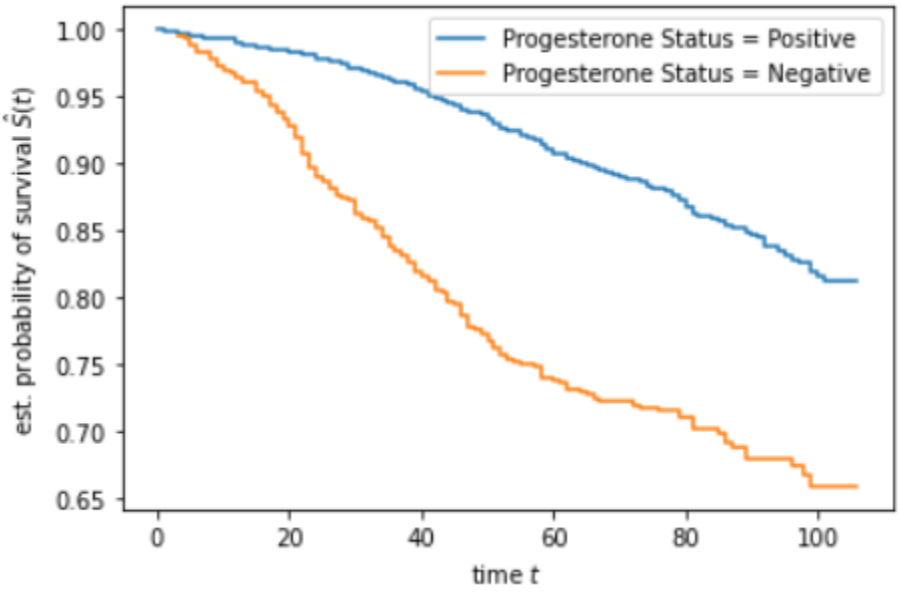
Single Survival Function using Kaplan-Meier



This graph has been plotted based on two features which are Status and Survival Months. The graph signifies the survival probability of all the patients to time in which we are not conside ring the effect of any other parameters or features like age, stage, etc. So, it is not a good choi ce for survival analysis as it is not taking account the other effects.

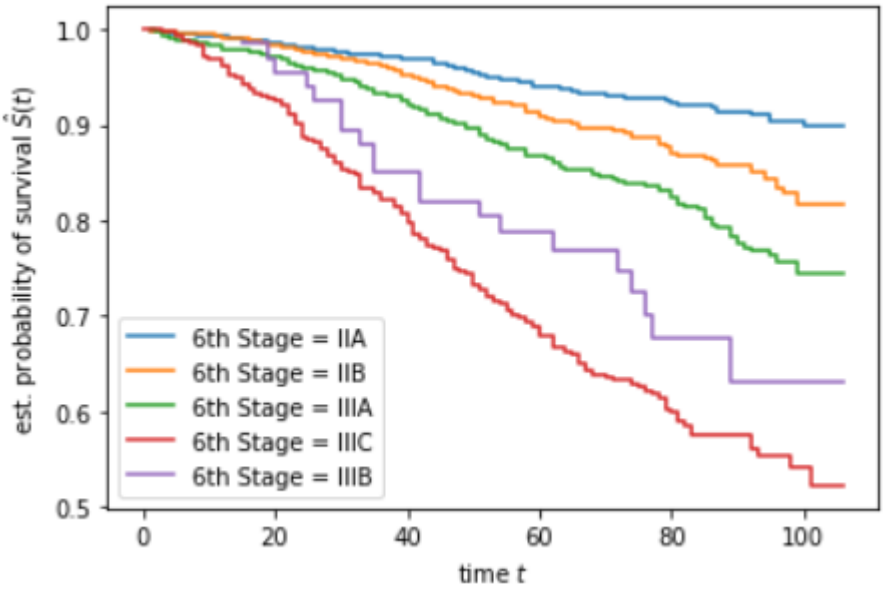
Kaplan-Meier Estimation Considering the Attributes

In this case, we can only consider the attribute which has the categorical data.

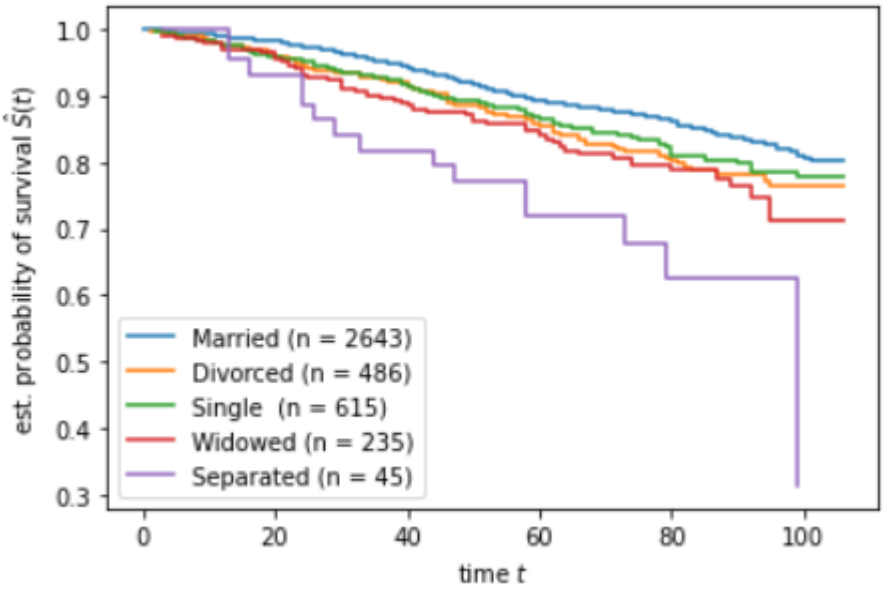


The survival function has been displayed based on progesterone status concerning time. The graph's results show two trends. The Negative status of progesterone has a higher risk than the

positive status of progesterone because the slope of the orange trend(negative status of progesterone) is steeper. At any point in time t, the survival rate is high when a patient has a positive status of progesterone.



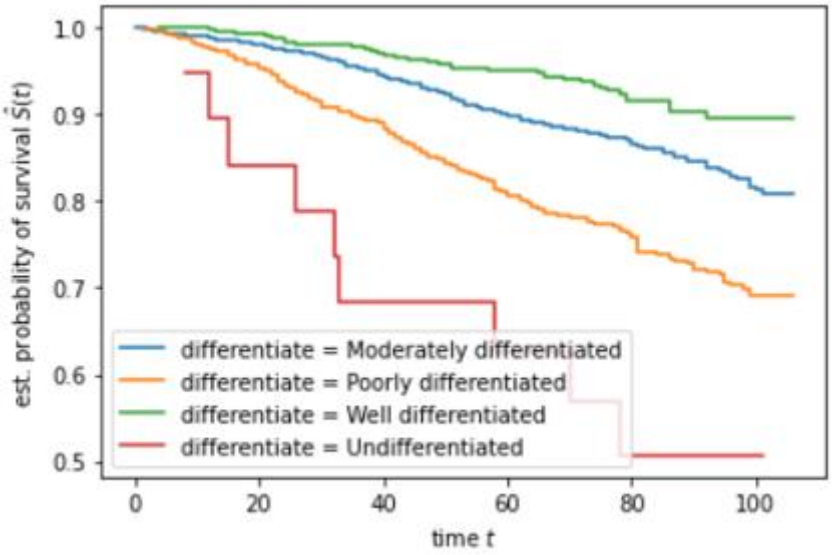
The following graph is portraying the effect of the 6th stage on the estimated probability of survival. 5 different lines of different stages of patients "IIA", "IIB", "IIIA", "IIIC", and "IIIB" have different rate of survival rates. The “IIIC” has a less probability of survival than other stages patients. On the other hand, “IIA” has a lower risk of death compared to other stages.

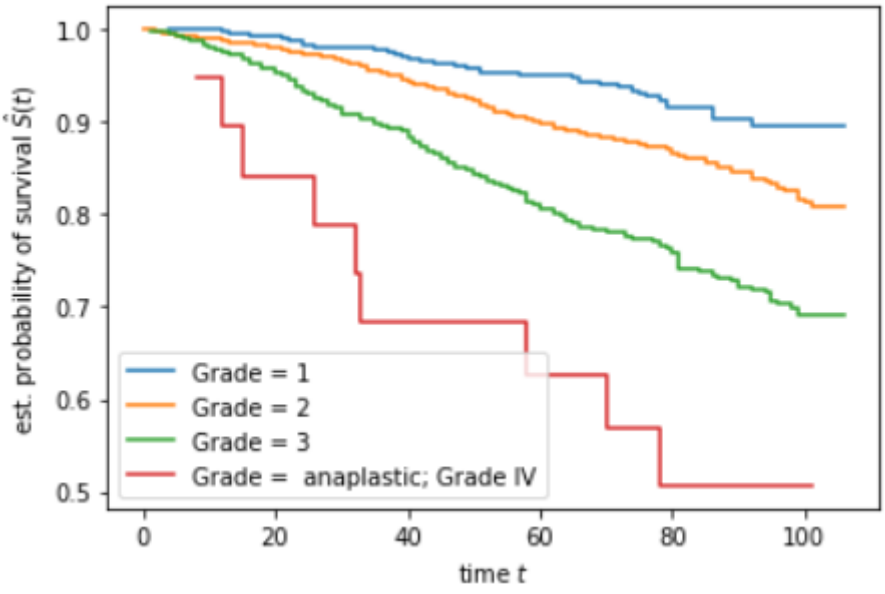


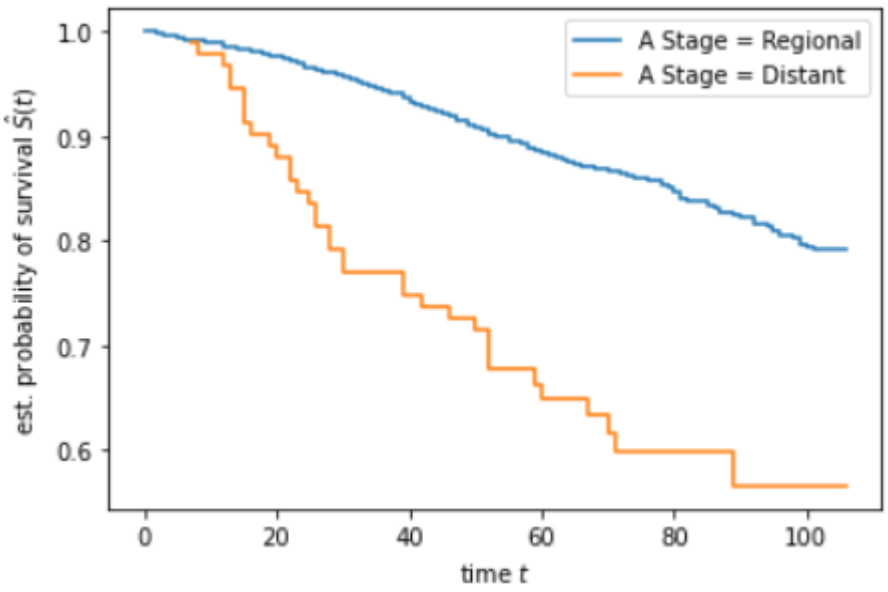
Considering the marital status feature, the succeeding graph of survival function to time has been plotted. The above graph may not be the ideal result of survival function as the number of separated women’s records is less compared to other categories like married, divorced, etc.

Except for the separated women, the survival rate of other categories are almost similar and all the trends are close to each other. So, we can neglect the effect of this column.

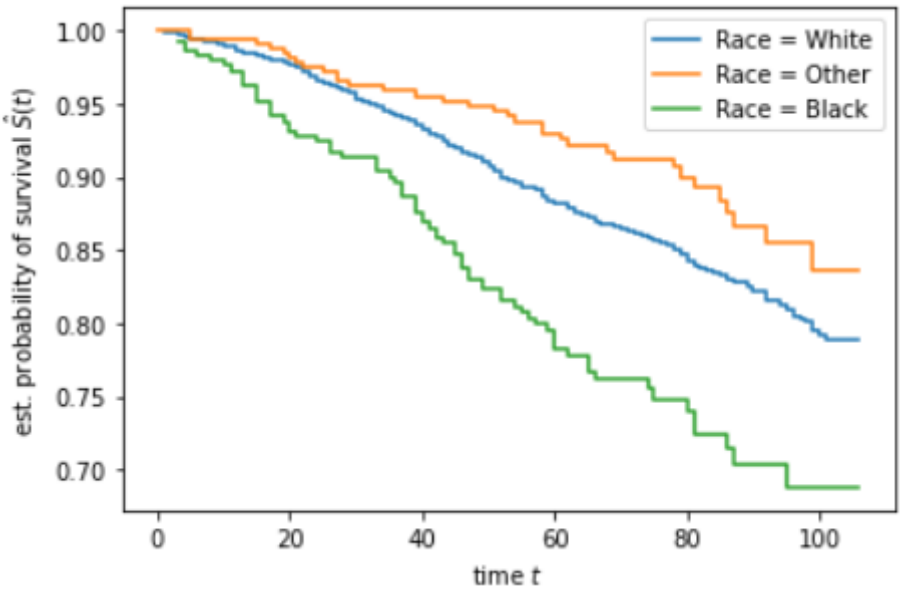
Similar kinds of trends are also gotten for some attributes like ‘differentiate’ and ‘grade’ which graphs are given below. The effects of these attributes can be neglected as they are given biased results.



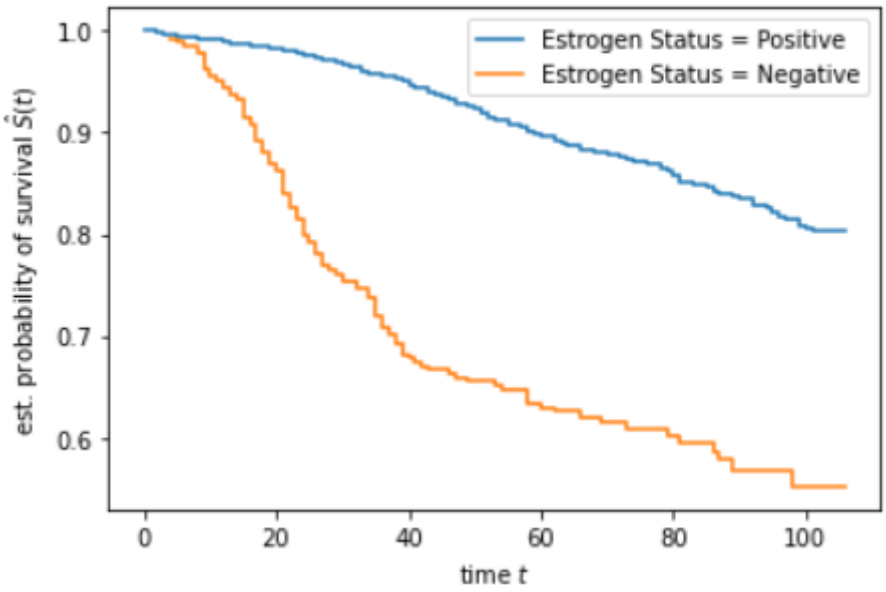




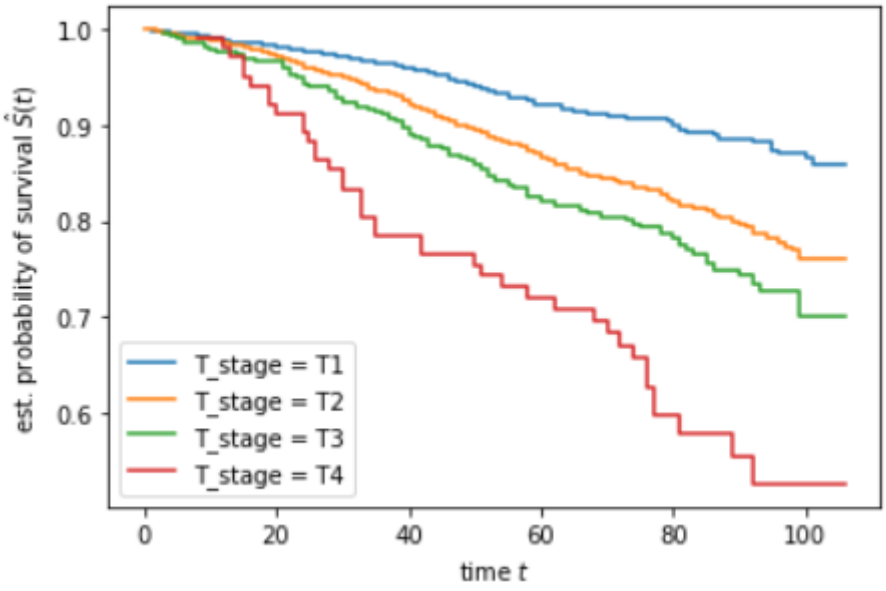
The survival probability of Reginal is high for Distant. The Kaplan-Meier estimation is being said that the women who are Distant from A Stage have a higher death risk than the Regional. From the graph, it is represented that the Distant one is very steeper than the other one.



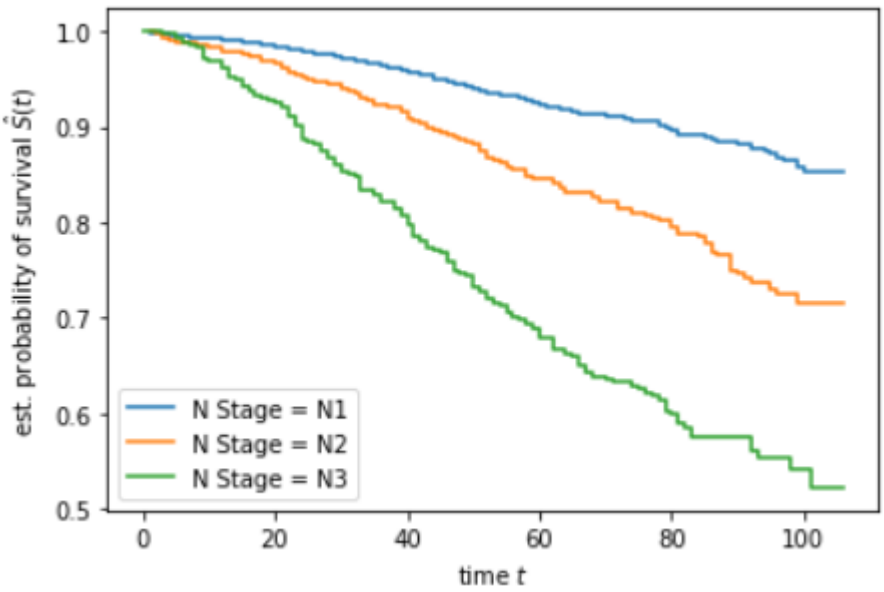
Based on race which is White, Other, and Black, we have estimated the survival probability which signifies that white women have a higher survivability of breast cancer compared to other women and black women have a higher risk of death. But this analysis on the basis of race may seem unnatural.



Based on Estrogen status across time, the survival function has been apparent. Results on the graph indicate two tendencies. Because the orange trend's slope (indicating the negative status of Estrogen) is steeper than that of the positive status of Estrogen, it carries a higher risk than the positive state. A patient who has a positive Estrogen status at any given time t has a high survival rate.



In accordance with the T_stage of four kinds which are T1, T2, T3, and T4, we have estimated the survival function. It is shown that the probability at any point of time t from where the study started is lower for T4 than other T_stage.



We have approximated the survival function in line with the N stage of three types, N1, N2, and N3. It is demonstrated that N3 has a lower probability than other N stages at any time t after the study's beginning.

A doctor can determine the survival rate for various cases based on the results above. The out come of the analysis will determine how the doctor will treat the patient. L like N-stage patien ts, N3-stage patients have a lower chance of survival. Consequently, she needs to receive mor e medication or treatment than other patients. Additionally, a doctor can reassure other stage women that they have a good chance of surviving.

The Kaplan-Meier Estimation is also not a good choice because it is univariate. For this, we h ave applied other models like Cox Proportional Hazard Model and Random Survival Forest M odel.

Cox Proportional Hazard Model

In the Cox model, the important parameter is the hazard ratio which has the following formul a, which represents the relative risk of an event occurring at time t: (t) / 0. A hazard ratio of 2 indicates that at any given period, there are twice as many incidents in the treatment group. A hazard ratio positive means, there is a negative impact on the patient which reduces survivabil ity. On the other hand, a negative hazard ratio means, it has a positive impact on the patient.

The following output represents the hazard ratio of each attribute of our dataset.

Age	0.020279
Race	-0.144992
Marital Status	0.044947
T_stage	0.305074
N Stage	0.351682
6th Stage	-0.009137
differentiate	-0.008689
Grade	0.293472
A Stage	-0.092808
Tumor Size	-0.000208
Estrogen Status	-0.702602
Progesterone Status	-0.475178
Regional Node Examined	-0.033007
Reginol Node Positive	0.056949

Fig: Log Hazard Ratio of each feature in Cox

To find the accuracy of the Cox model or the performance of the Cox model we have found t he C-Index which is `0.7358041287445589`
Then we have performed feature importance concerning C-Index:

6th Stage	0.668648
Reginol Node Positive	0.659404
N Stage	0.651261
Tumor Size	0.609325
Progesterone Status	0.602014
T_stage	0.599802
Grade	0.597362
Estrogen Status	0.571949
Age	0.544994
Regional Node Examined	0.524454
A Stage	0.521829
Race	0.516124
Marital Status	0.514455
differentiate	0.469465

dtype: float64

Fig: CoxPHSurvival C-index of each feature

From the output table, we are noticing that the 6th stage has high importance than other features. After that Regional Node Positive has higher importance. The importance of each feature is sequentially represented in the following output.

Predicting risk scores of a few patients using the RSF model based on age and tumor size

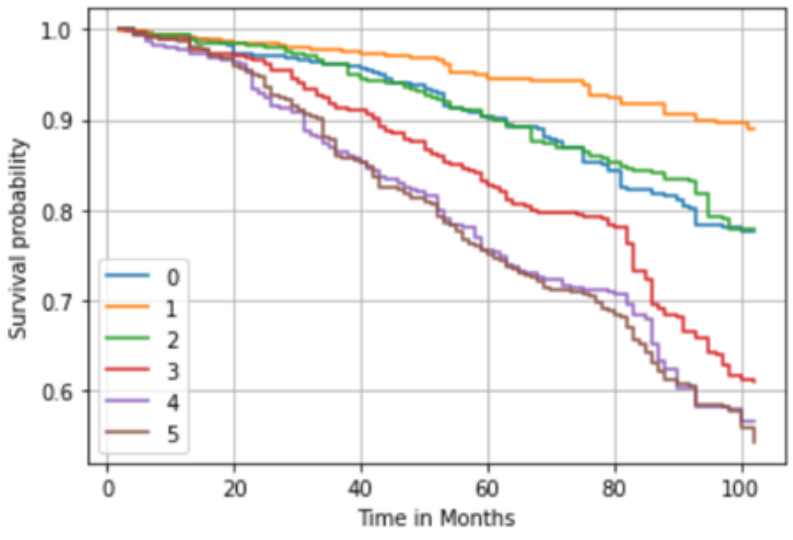
The following table given below is manifested sorted data according to age and tumor size, the samples of 6 patients have been extracted.

	Age	Race	Marital Status	T_stage	N Stage	6th Stage	differentiate	Grade	A Stage	Tumor Size	Estrogen Status	Progesterone Status	Regional Node Examined	Reginol Node Positive
702	30	2	3	1	1	2	0	2	1	28	1	1	19	7
998	31	1	3	0	0	0	0	2	1	19	1	1	20	3
1143	31	2	0	1	0	1	0	2	1	42	1	1	9	3
3503	69	2	0	2	0	2	1	3	1	90	1	1	4	2
155	69	2	4	2	1	2	0	2	1	100	1	0	16	6
3775	69	2	0	3	1	3	0	2	1	120	1	1	9	7

The Risk score of 6 patients are given below:

```
0     9.639917
1     4.662552
2     9.151795
3    17.281357
4    23.570835
5    24.019244
dtype: float64
```

Here 0, 1, 2 .. signifies the patient number. From the risk score, we find that patient 6 has a higher risk of death and a higher risk score whose age is 69 and tumor size is 100. The result is matched our intuition. If we compare the risk score of patients 1 and 2, the risk score is higher for patient 1, though she is younger than patient 2. From intuition, we can say that the reason is tumor size. Patient 1 has a bigger tumor size than patient 2. Then we plotted the survival function for 6 patients and analyze it accordingly which is mentioned above.



After this, we performed permutation-based features importance which is similar to features selection. Then we compared the result with our previous result of C-Index-based feature importance.

	importances_mean	importances_std
Reginol Node Positive	0.033329	0.008343
6th Stage	0.021256	0.008207
Progesterone Status	0.019592	0.003800
Age	0.018736	0.006898
N Stage	0.016184	0.006780
Estrogen Status	0.011395	0.004381
Regional Node Examined	0.008639	0.004098
Grade	0.007865	0.002384
Tumor Size	0.007492	0.001959
differentiate	0.004394	0.001730
Race	0.004119	0.001509
T_stage	0.003733	0.001444
A Stage	0.000691	0.000527
Marital Status	0.000667	0.001561

In this case, we got that the Regional Node Positive feature is more important than other features and 2nd one is the 6th stage feature. These feature selection techniques gave us a close result. In the case of the C-Index, the important feature was 6th stage

Contribution

Aseer Imad Keats	Mohammad Abrar Kabir	Md. Saifur Rahman	Md. Mahfuzul Islam
Team Co-ordination	Research	Research	Research
Research	Coding	Coding	Documentation
Coding	Debugging	Code Modification	Coding
Documentation	Result Analysis	Code Assembling	Data Analysis

END