Prediction of Breast Cancer Survival Rate

PROBLEM

Breast cancer is a common problem among women worldwide, and early detection and treatment are essential for improving survival rates. However, there are limited resources that can accurately predict a patient's survivability rate.

SOLUTION

We propose to develop a bioinformatics tool that can predict breast cancer survivability rate by integrating omics data. We will use exploratory data analysis to identify the most important features for predicting survivability rate, and we will develop machine learning models to train the tool.

DATASET

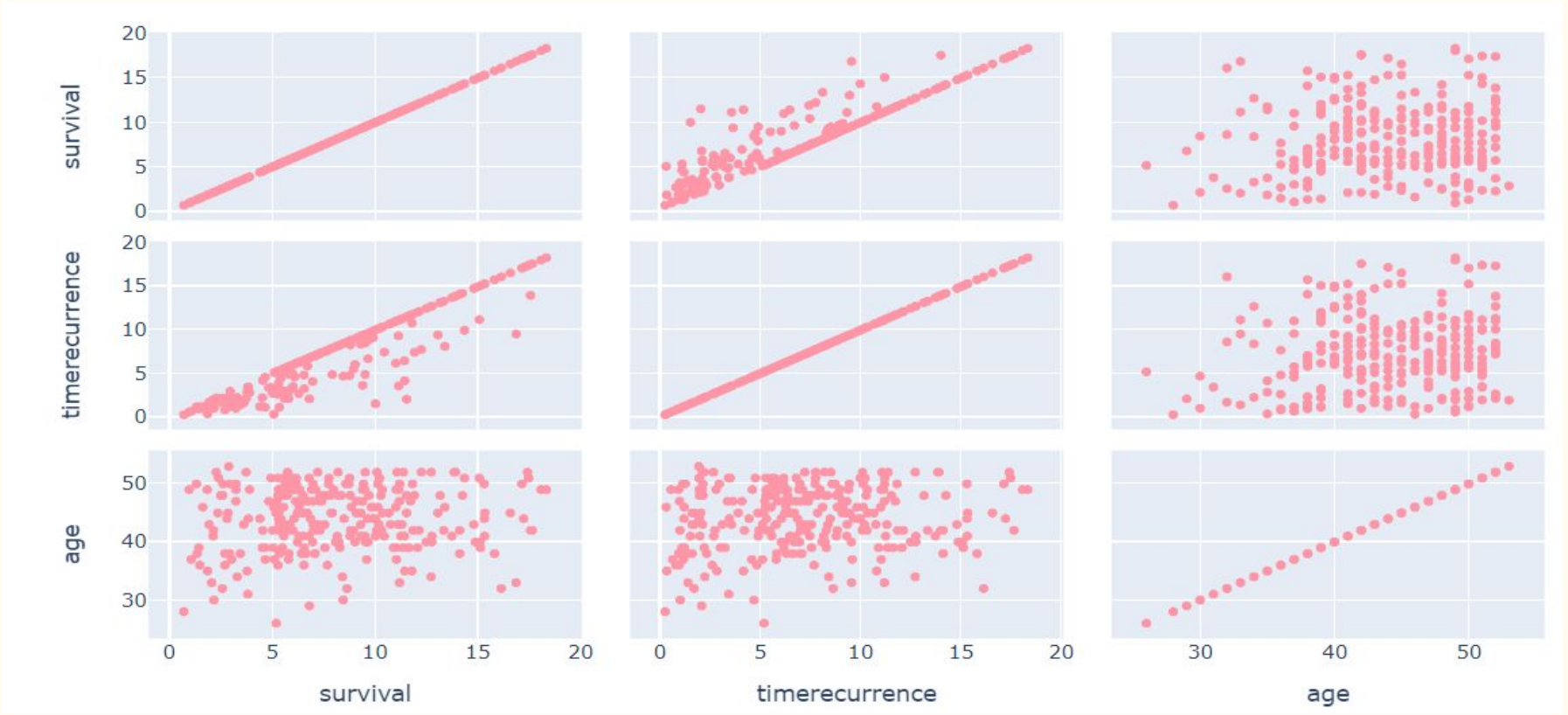
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Patient	ID	age	eventdeath	survival	timerecurrence	chemo	hormonal	amputation	histtype	diam	posnodes	grade	angioinv	lymphinfil	barcode	esr1	G3PDH_570	Contig45645_RC	Contig44916_RC	D25272
2	s122	18	43	0	14.817248	14.817248	0	0	1	1	25	0	2	3	1	6274	-0.413955	-0.954246	0.051024	-0.111203	-0.050(
3	s123	19	48	0	14.261465	14.261465	0	0	0	1	20	0	3	3	1	6275	0.195251	0.244626	-0.199602	-0.111397	-0.135:
4	s124	20	38	0	6.644764	6.644764	0	0	0	1	15	0	2	1	1	6276	0.596177	0.082434	-0.156199	-0.08498	-0.179(
5	s125	21	50	0	7.748118	7.748118	0	1	0	1	15	1	2	3	1	6277	0.501286	-1.071614	-0.206041	-0.051775	-0.049(
6	s126	22	38	0	6.436687	6.31896	0	0	1	1	15	0	2	2	1	6278	-0.066771	-0.982276	-0.514666	-0.118483	-0.086:
7	s127	23	42	0	5.037645	2.743326	1	0	1	1	10	1	1	1	1	6279	0.121859	-0.432543	-0.436983	-0.087399	-0.166(
8	s128	24	50	0	8.73922	8.73922	1	1	0	1	25	1	1	1	1	6280	0.27341	-1.117449	-0.191362	0.009781	-0.082(
9	s129	25	43	0	7.56742	7.56742	1	0	0	1	15	3	2	2	1	6281	0.088803	0.226987	-0.090313	0.010277	0.051:
10	s130	26	47	0	7.296372	7.296372	1	0	0	1	18	1	3	1	2	6282	-1.006341	0.354804	-0.097944	0.023238	0.090:
11	s131	27	39	1	4.66256	1.114305	0	0	0	1	17	0	3	1	1	6283	-1.324176	-0.104224	-0.4278	-0.003295	-0.060(
12	s132	28	47	0	6.718686	5.867214	1	1	0	1	15	1	2	3	1	6284	-0.202664	0.165731	-0.058181	0.041886	0.128-
13	s133	29	32	0	8.648871	8.648871	0	0	0	1	18	0	1	2	1	6285	-0.252099	0.125784	-0.002987	-0.061983	-0.03(
14	s134	30	38	1	7.093771	6.995209	1	0	0	1	12	1	3	1	1	6286	-0.162349	0.190224	0.157729	-0.110239	-0.022:
15	s135	31	45	0	9.330595	9.330595	0	1	1	1	20	2	3	1	3	6287	-1.160662	-0.409659	-0.179898	-0.03023	-0.089:
16	s136	32	31	1	3.82204	3.438741	0	0	1	1	40	2	3	1	1	6288	-0.661729	0.09664	-0.258875	-0.105883	-0.178:
17	s137	33	41	0	15.329227	15.329227	1	0	0	1	18	1	1	1	1	6289	0.098739	-0.960866	-0.018987	-0.015122	0.030(
18	s138	34	44	1	3.849418	3.474333	0	0	1	1	15	0	3	1	1	6290	-0.435775	-0.799154	0.081538	-0.013132	0.034:
19	s139	35	41	0	12.766598	12.766598	0	0	1	1	45	0	3	3	1	6291	-0.048179	-0.342376	-0.071084	0.000781	0.017:
20	s140	36	46	0	5.555099	5.555099	0	0	0	1	18	0	1	1	1	6292	0.034581	-0.469436	-0.139353	-0.080099	0.008
21	s141	37	33	1	2.064339	1.40178	0	0	1	1	30	0	3	1	3	6293	-1.255379	-0.006519	-0.201504	-0.1204	-0.077:
22	s142	38	39	0	15.134839	15.134839	0	0	1	1	20	0	3	1	2	6294	-0.3873	-0.042295	0.233913	0.151385	0.201(
23	s144	39	38	0	14.12731	14.12731	0	0	1	1	30	0	3	1	3	6296	-1.088931	0.081496	-0.546575	-0.033396	-0.172:
24	s145	40	48	0	5.486653	5.486653	1	1	0	1	19	2	2	3	1	6297	-0.164632	0.271391	-0.119456	-0.073048	0.072(

(272×1570)

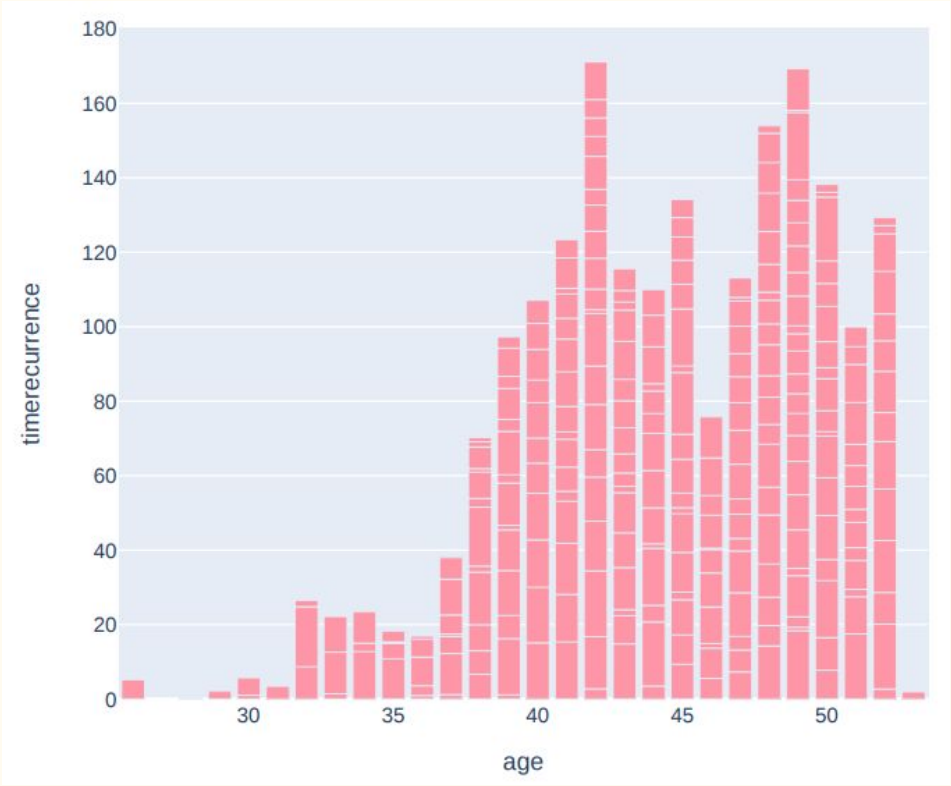
https://www.kaggle.com/datasets/nancyalaswad90/cancer-statistics-in-us-states?select=NKI_cleaned.csv

Data Exploration

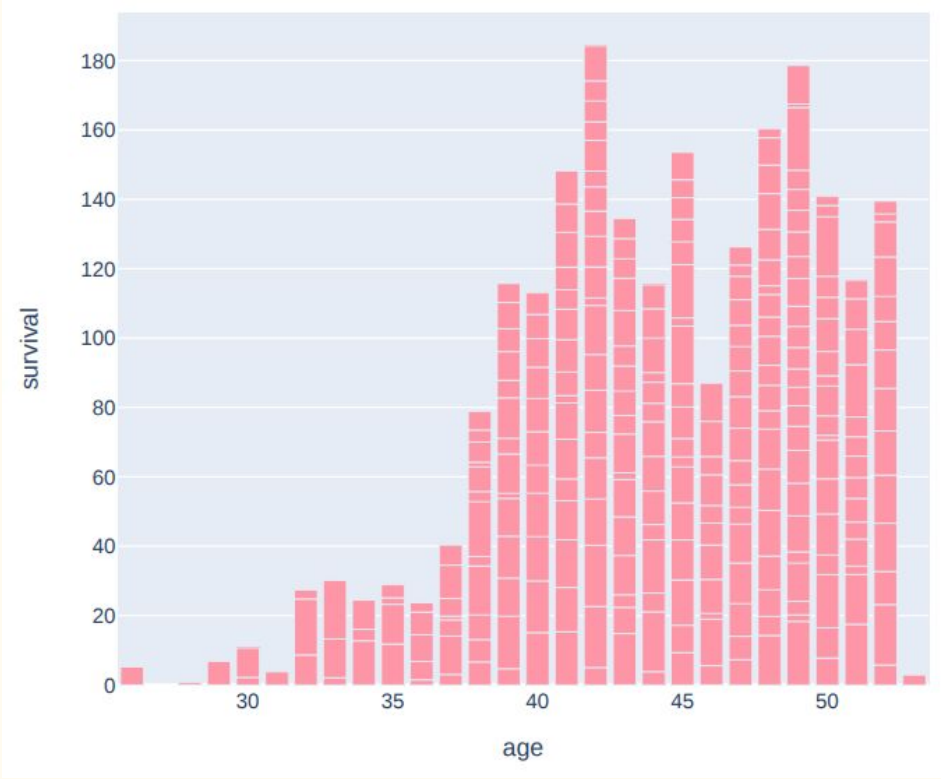
Scatter



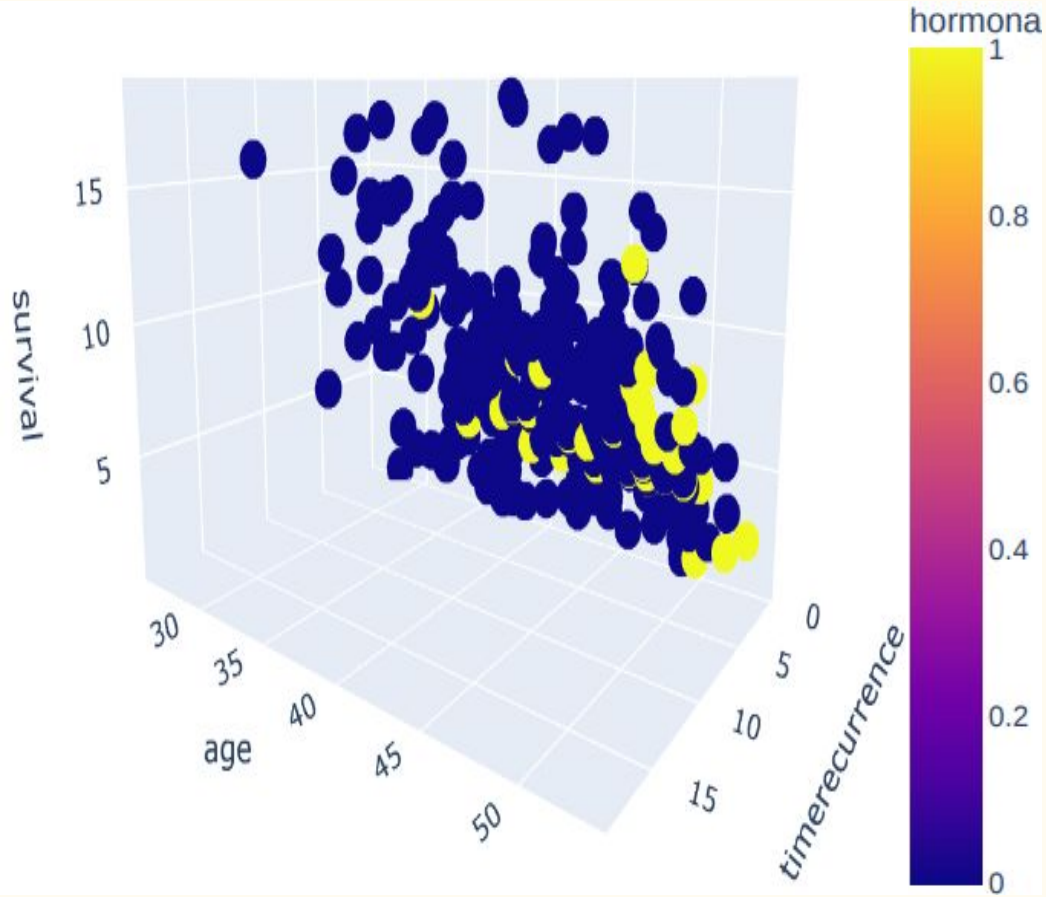
Time Recurrence vs Age



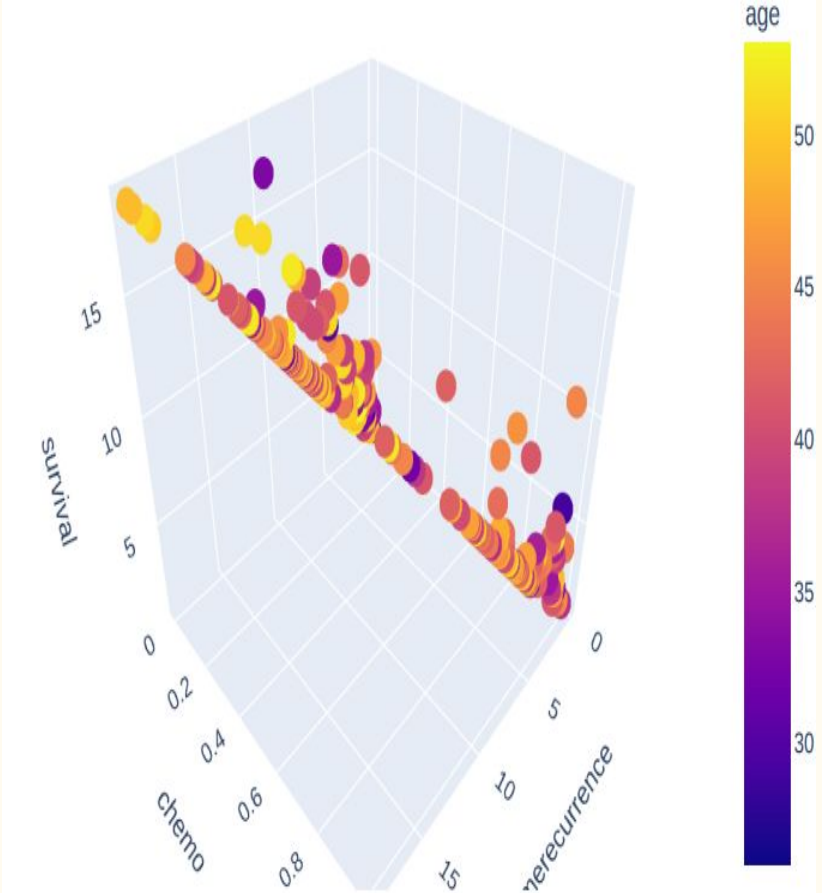
Survival vs Age



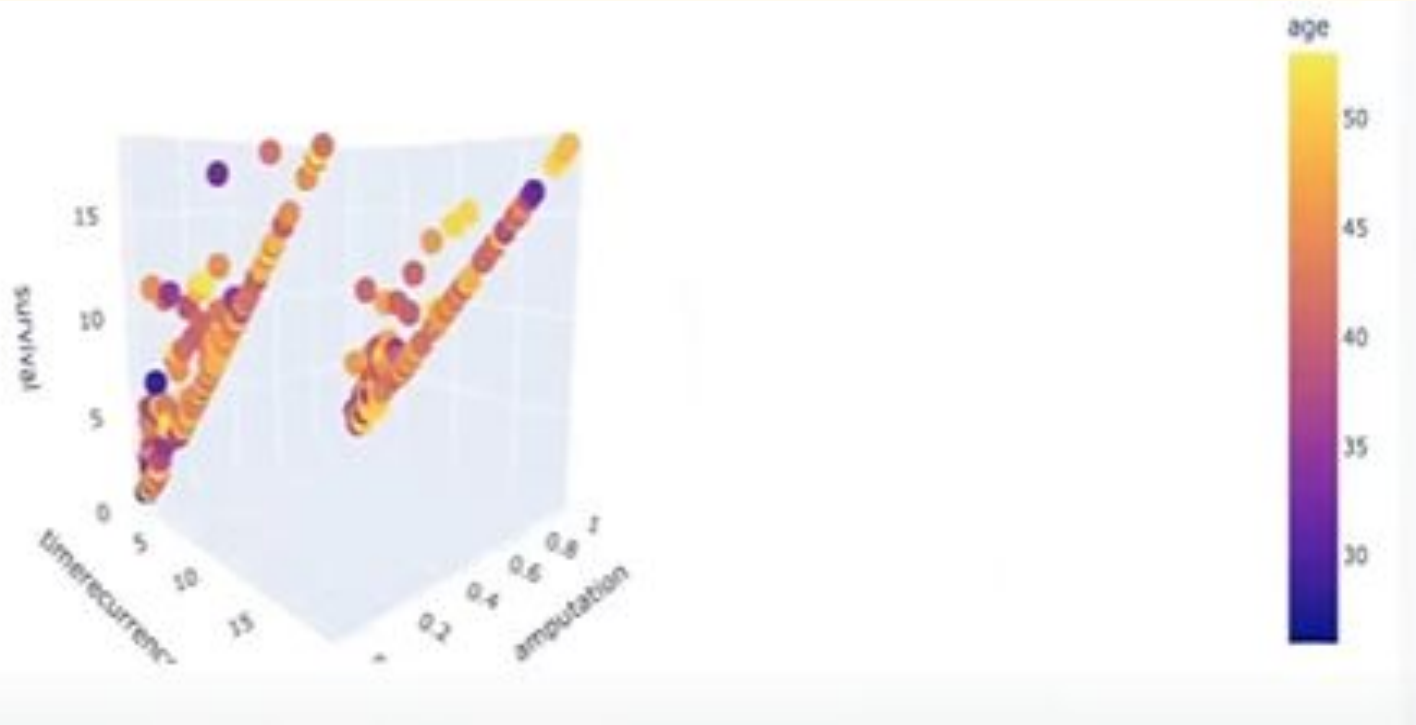
Time Recurrence vs Age vs Survival



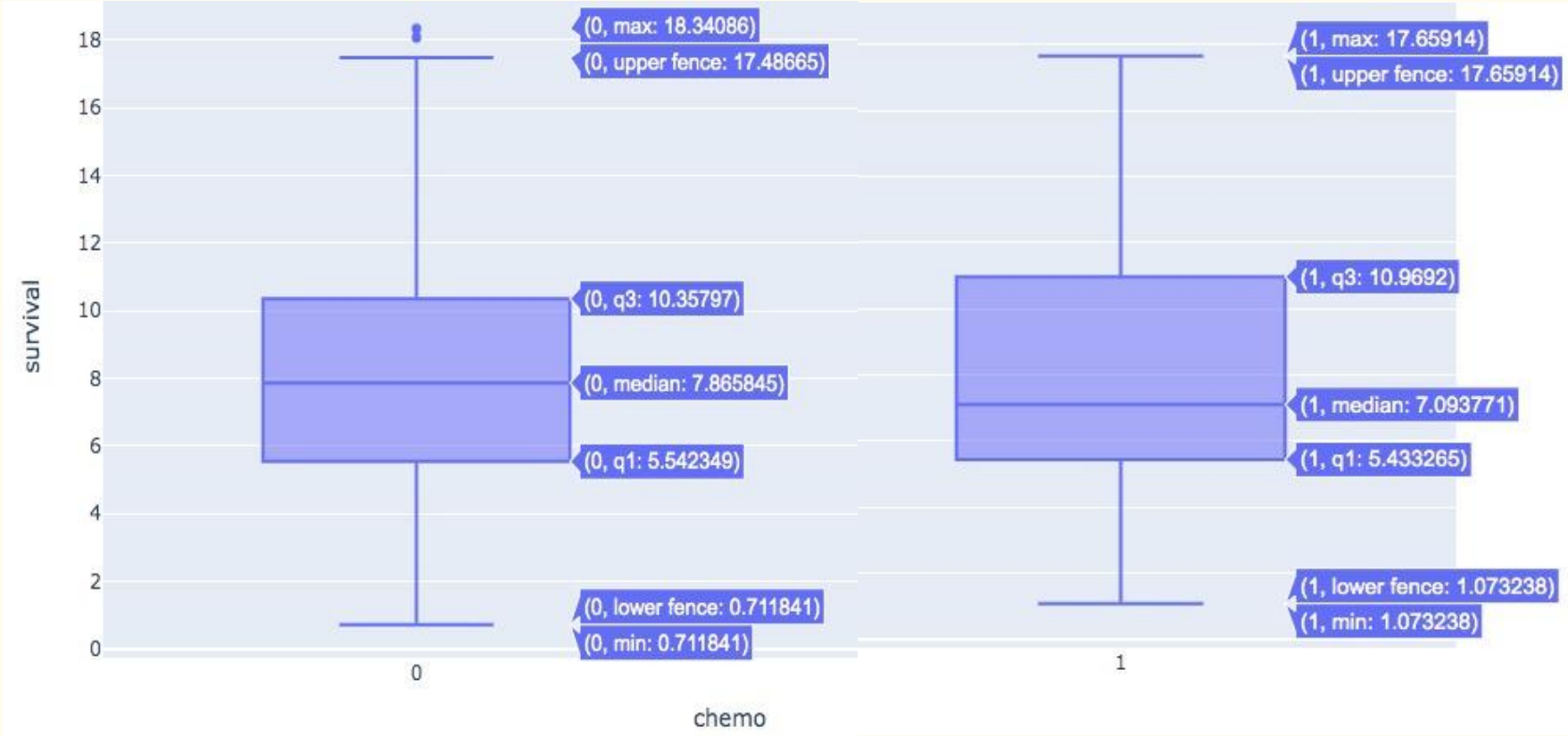
Time Recurrence vs Chemo vs Survival



Time Recurrence vs Survival Rate vs Amputation



BOX PLOT


(0, max: 18.34086)
(0, upper fence: 17.48665)
(0, q3: 10.35797)
(0, median: 7.865845)
(0, q1: 5.542349)
(0, lower fence: 0.711841)
(0, min: 0.711841)
(1, max: 17.65914)
(1, upper fence: 17.65914)
(1, q3: 10.9692)
(1, median: 7.093771)
(1, q1: 5.433265)
(1, lower fence: 1.073238)
(1, min: 1.073238)
survival
18
16
14
12
10
8
6
4
2
0
0
1
chemo

Prediction Model

We have implemented Decision Tree Regression model to predict the survivability rate of patient on the basis of their age, time recurrence and the therapies they went through.

Introduction the Decision Tree Regression model:

A decision tree regression model is a type of machine learning model that can be used to predict a continuous value, such as the survivability rate of a patient. The model works by creating a tree-like structure of decisions, where each decision is based on a feature of the data. The model then uses the values of the features to make predictions for the target value.

The features used in the model are:

* Age: The age of the patient at the time of diagnosis.

* TimeRecurrence: The time in months from diagnosis to recurrence.

* Chemo(0/1): Whether the patient received chemotherapy.

* Hormonal(0/1): Whether the patient received hormonal therapy.

* Amputation(0/1): Whether the patient had an amputation.

* HistType(1,2,4,5,7): The histological type of the tumor.

92.6204%

Accuracy of the model after a few runs...

```
lr1.fit(X_train, y_train)
pred1 = lr1.predict(X_test)
print("Validation Accuracy: ", lr1.score(X_test, y_test))
```

```
Validation Accuracy:  0.9262042520337905
```

```
age = input("Enter age ")
timerecurrence = input("Enter timerecurrence ")
chemo = input("Enter chemo ")
hormonal = input("Enter hormonal ")
amputation = input("Enter amputation ")
histtype = input("Enter histtype ")
```

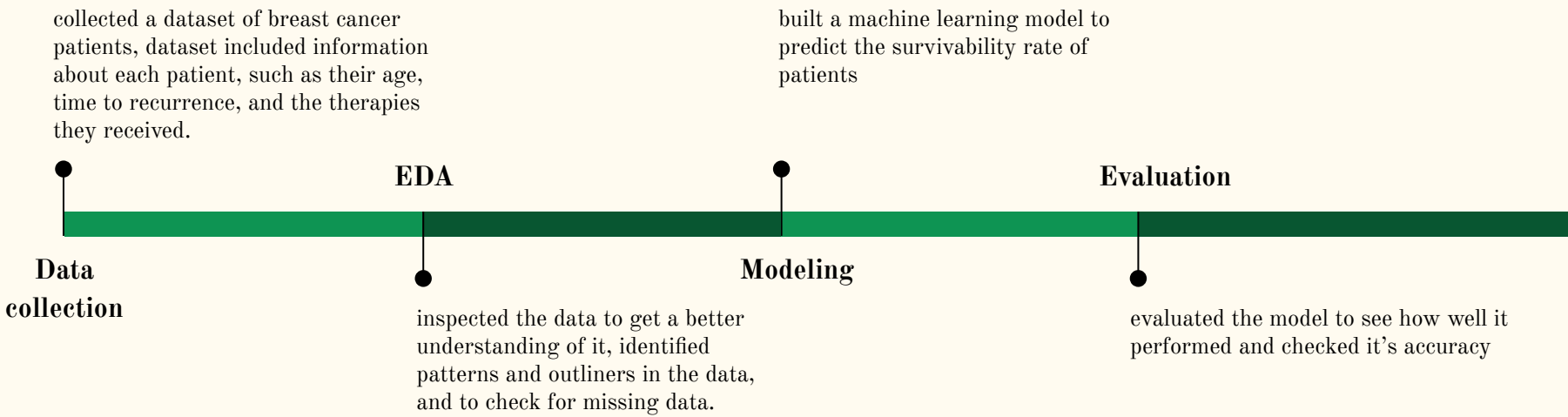
```
Enter age 24
Enter timerecurrence 2
Enter chemo 1
Enter hormonal 0
Enter amputation 0
Enter histtype 3
```

```
# Make a prediction using the custom input values
prediction = regressor.predict([[age, timerecurrence, chemo, hormonal, amputation, histtype]])

# Print the prediction
print('The predicted survivability rate is {}%'.format(prediction[0]))
```

```
The predicted survivability rate is 2.132786%
```

Project Flow



THANK YOU