Table of Contents

I. Problem 2

II. Data 2

Clinical Attributes: 3

Genetic Attributes: 4

III. Workflows 4

1. EDA & engineer features: 4

a) Preliminary cleaning of the dataset: 5

b) Relationship between clinical attributes and outcomes 6

c) Relationship between genetic attributes and outcomes 13

d) Relationship between genetic mutation attributes and outcomes 17

2. Data preprocessing: 18

3. Classification Models and Evaluation: 18

a) Classification with only clinical attributes: 18

b) XGBoost Classifier for clinical attributes only 25

4. Web Deploy: 26

IV. Web Interface 26

V. Conclusion 29

VI. Presentation 29

Breast Cancer Surgery Survival Prediction

I. Problem

In today's medical landscape, patients undergoing cancer surgeries, such as breast cancer surgery, are often required to sign a document stating that the doctors will not be held responsible for any complications that may arise during the operation. This practice can instill fear and anxiety in patients, creating an unstable mindset as they face surgery. Scientific research has demonstrated that patients with an unstable mindset going into surgery have a lower survival rate compared to those with a stable mindset. This highlights the need for a more empathetic and supportive approach to patient care, ensuring patients feel informed and reassured as they embark on their surgical journey.

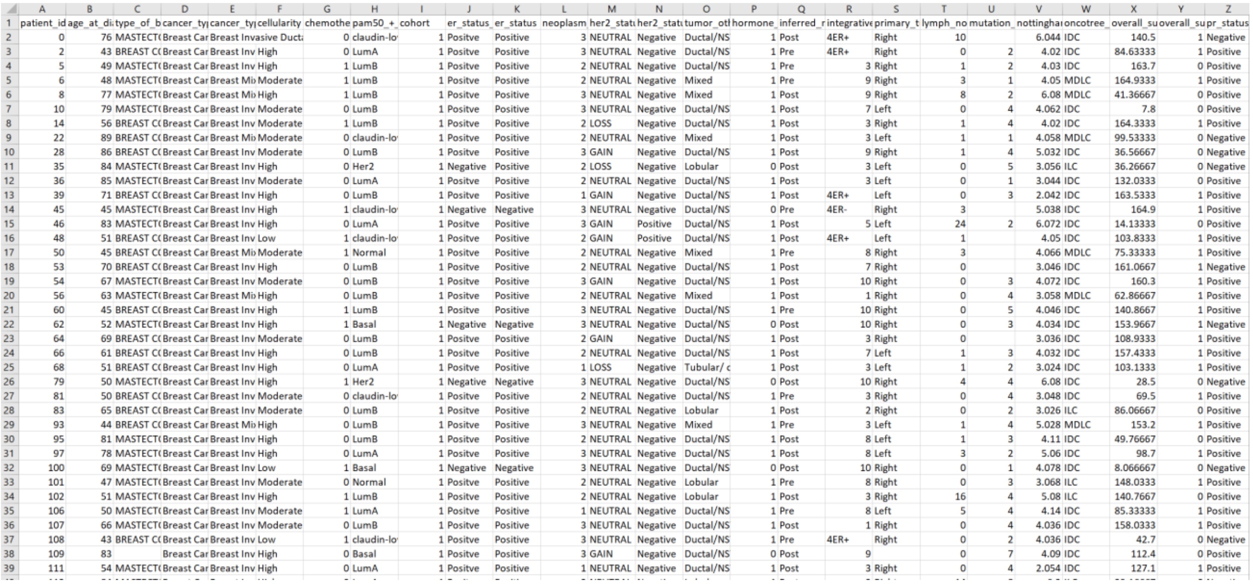
II. Data

Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Dataset

The dataset used is the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database, which is a Canada-UK Project which contains targeted sequencing data of 1,980 primary breast cancer samples.

This is a huge dataset with 1904 rows and 693 columns which among it, 31 columns are the clinical attributes and the rest are the genetic attributes.

Source: https://www.cbioportal.org/



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	patient_id	age_at_di	type_of_b	cancer_ty	cancer_ty	cellularity	chemothe	pam50_+_	cohort	er_status_	er_status	neoplasm	her2_stat	her2_stat	tumor_oth	hormone_	inferred_r	integrative	primary_t	lymph_no	mutation_	nottinghar	oncotree_	overall_su	overall_su	pr_status
2	0	76	MASTECT(	Breast Car	Breast Invasive Ducta		0	claudin-lo	1	Positve	Positive	3	NEUTRAL	Negative	Ductal/NS	1	Post	4ER+	Right	10		6.044	IDC	140.5	1	Negative
3	2	43	BREAST C(	Breast Car	Breast Inv	High	0	LumA	1	Positve	Positive	3	NEUTRAL	Negative	Ductal/NS	1	Pre	4ER+	Right	0	2	4.02	IDC	84.63333	1	Positive
4	5	49	MASTECT(	Breast Car	Breast Inv	High	1	LumB	1	Positve	Positive	2	NEUTRAL	Negative	Ductal/NS	1	Pre	3	Right	1	2	4.03	IDC	163.7	0	Positive
5	6	48	MASTECT(	Breast Car	Breast Mi	Moderate	1	LumB	1	Positve	Positive	2	NEUTRAL	Negative	Mixed	1	Pre	9	Right	3	1	4.05	MDLC	164.9333	1	Positive
6	8	77	MASTECT(	Breast Car	Breast Mi	High	1	LumB	1	Positve	Positive	3	NEUTRAL	Negative	Mixed	1	Post	9	Right	8	2	6.08	MDLC	41.36667	0	Positive
7	10	79	MASTECT(	Breast Car	Breast Inv	Moderate	0	LumB	1	Positve	Positive	3	NEUTRAL	Negative	Ductal/NS	1	Post	7	Left	0	4	4.062	IDC	7.8	0	Positive
8	14	56	BREAST C(	Breast Car	Breast Inv	Moderate	1	LumB	1	Positve	Positive	2	LOSS	Negative	Ductal/NS	1	Post	3	Right	1	4	4.02	IDC	164.3333	1	Positive
9	22	89	BREAST C(	Breast Car	Breast Mi	Moderate	0	claudin-lo	1	Positve	Positive	2	NEUTRAL	Negative	Mixed	1	Post	3	Left	1	1	4.058	MDLC	99.53333	0	Negative
10	28	86	BREAST C(	Breast Car	Breast Inv	Moderate	0	LumB	1	Positve	Positive	3	GAIN	Negative	Ductal/NS	1	Post	9	Right	1	4	5.032	IDC	36.56667	0	Negative
11	35	84	MASTECT(	Breast Car	Breast Inv	High	0	Her2	1	Negative	Positive	2	LOSS	Negative	Lobular	0	Post	3	Left	0	5	3.056	ILC	36.26667	0	Negative
12	36	85	MASTECT(	Breast Car	Breast Inv	Moderate	0	LumA	1	Positve	Positive	2	NEUTRAL	Negative	Ductal/NS	1	Post	3	Left	0	1	3.044	IDC	132.0333	0	Positive
13	39	71	BREAST C(	Breast Car	Breast Inv	High	0	LumB	1	Positve	Positive	1	GAIN	Negative	Ductal/NS	1	Post	4ER+	Left	0	3	2.042	IDC	163.5333	1	Positive
14	45	45	MASTECT(	Breast Car	Breast Inv	High	1	claudin-lo	1	Negative	Negative	3	NEUTRAL	Negative	Ductal/NS	0	Pre	4ER-	Right	3		5.038	IDC	164.9	1	Positive
15	46	83	MASTECT(	Breast Car	Breast Inv	High	0	LumA	1	Positve	Positive	3	GAIN	Positive	Ductal/NS	1	Post	5	Left	24	2	6.072	IDC	14.13333	0	Positive
16	48	51	BREAST C(	Breast Car	Breast Inv	Low	1	claudin-lo	1	Positve	Positive	2	GAIN	Positive	Ductal/NS	1	Post	4ER+	Left	1		4.05	IDC	103.8333	1	Positive
17	50	45	BREAST C(	Breast Car	Breast Mi	Moderate	1	Normal	1	Positve	Positive	2	NEUTRAL	Negative	Mixed	1	Pre	8	Right	3		4.066	MDLC	75.33333	1	Positive
18	53	70	BREAST C(	Breast Car	Breast Inv	High	0	LumB	1	Positve	Positive	2	NEUTRAL	Negative	Ductal/NS	1	Post	7	Right	0		3.046	IDC	161.0667	1	Negative
19	54	67	MASTECT(	Breast Car	Breast Inv	Moderate	0	LumB	1	Positve	Positive	3	GAIN	Negative	Ductal/NS	1	Post	10	Right	0	3	4.072	IDC	160.3	1	Positive
20	56	63	MASTECT(	Breast Car	Breast Mi	High	0	LumB	1	Positve	Positive	2	NEUTRAL	Negative	Mixed	1	Post	1	Right	0	4	3.058	MDLC	62.86667	1	Positive
21	60	45	BREAST C(	Breast Car	Breast Inv	High	1	LumB	1	Positve	Positive	3	NEUTRAL	Negative	Ductal/NS	1	Pre	10	Right	0	5	4.046	IDC	140.8667	1	Positive
22	62	52	MASTECT(	Breast Car	Breast Inv	High	1	Basal	1	Negative	Negative	3	NEUTRAL	Negative	Ductal/NS	0	Post	10	Right	0	3	4.034	IDC	153.9667	1	Negative
23	64	69	BREAST C(	Breast Car	Breast Inv	Moderate	0	LumB	1	Positve	Positive	2	GAIN	Negative	Ductal/NS	1	Post	3	Right	0		3.036	IDC	108.9333	1	Positive
24	66	61	BREAST C(	Breast Car	Breast Inv	High	0	LumB	1	Positve	Positive	2	NEUTRAL	Negative	Ductal/NS	1	Post	7	Left	1	3	4.032	IDC	157.4333	1	Positive
25	68	51	BREAST C(	Breast Car	Breast Inv	High	0	LumA	1	Positve	Positive	1	LOSS	Negative	Tubular/ c	1	Post	3	Left	1	2	3.024	IDC	103.1333	1	Positive
26	79	50	MASTECT(	Breast Car	Breast Inv	High	1	Her2	1	Negative	Negative	3	NEUTRAL	Negative	Ductal/NS	0	Post	10	Right	4	4	6.08	IDC	28.5	0	Negative
27	81	50	BREAST C(	Breast Car	Breast Inv	Moderate	0	claudin-lo	1	Positve	Positive	2	NEUTRAL	Negative	Ductal/NS	1	Pre	3	Right	0	4	3.048	IDC	69.5	1	Positive
28	83	65	BREAST C(	Breast Car	Breast Inv	Moderate	0	LumB	1	Positve	Positive	2	NEUTRAL	Negative	Lobular	1	Post	2	Right	0	2	3.026	ILC	86.06667	0	Positive
29	93	44	BREAST C(	Breast Car	Breast Mi	High	0	LumB	1	Positve	Positive	3	NEUTRAL	Negative	Mixed	1	Pre	3	Left	1	4	5.028	MDLC	153.2	1	Positive
30	95	81	MASTECT(	Breast Car	Breast Inv	High	0	LumB	1	Positve	Positive	2	NEUTRAL	Negative	Ductal/NS	1	Post	8	Left	1	3	4.11	IDC	49.76667	0	Positive
31	97	78	MASTECT(	Breast Car	Breast Inv	High	0	LumA	1	Positve	Positive	3	NEUTRAL	Negative	Ductal/NS	1	Post	8	Left	3	2	5.06	IDC	98.7	1	Positive
32	100	69	MASTECT(	Breast Car	Breast Inv	Low	1	Basal	1	Negative	Negative	3	NEUTRAL	Negative	Ductal/NS	0	Post	10	Right	0	1	4.078	IDC	8.066667	0	Negative
33	101	47	MASTECT(	Breast Car	Breast Inv	Moderate	0	Normal	1	Positve	Positive	2	NEUTRAL	Negative	Lobular	1	Pre	8	Right	0	3	3.068	ILC	148.0333	1	Positive
34	102	51	MASTECT(	Breast Car	Breast Inv	High	1	LumB	1	Positve	Positive	2	NEUTRAL	Negative	Lobular	1	Post	3	Right	16	4	5.08	ILC	140.7667	0	Positive
35	106	50	MASTECT(	Breast Car	Breast Inv	Moderate	1	LumA	1	Positve	Positive	1	NEUTRAL	Negative	Ductal/NS	1	Pre	8	Left	5	4	4.14	IDC	85.33333	1	Positive
36	107	66	MASTECT(	Breast Car	Breast Inv	Moderate	0	LumB	1	Positve	Positive	3	NEUTRAL	Negative	Ductal/NS	1	Post	1	Right	0	4	4.036	IDC	158.0333	1	Positive
37	108	43	BREAST C(	Breast Car	Breast Inv	Low	1	claudin-lo	1	Positve	Positive	3	NEUTRAL	Negative	Ductal/NS	1	Pre	4ER+	Right	0	2	4.036	IDC	42.7	0	Negative
38	109	83		Breast Car	Breast Inv	High	0	Basal	1	Positve	Positive	3	GAIN	Negative	Ductal/NS	0	Post	9		0	7	4.09	IDC	112.4	0	Positive
39	111	54	MASTECT(	Breast Car	Breast Inv	High	0	LumA	1	Positve	Positive	1	NEUTRAL	Negative	Ductal/NS	1	Post	3	Right	0	4	2.054	IDC	127.1	1	Positive

Clinical Attributes:

patient_id: Patient ID

age_at_diagnosis: Age of the patient at diagnosis time

type_of_breast_surgery: Breast cancer surgery type

cancer_type: Breast cancer types

cancer_type_detailed: Detailed breast cancer types

cellularity: Cancer cellularity post chemotherapy, which refers to the amount of tumor cells in the specimen and their arrangement into clusters

chemotherapy: Whether or not the patient had chemotherapy as a treatment

pam50_+_claudin-low_subtype: 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

cohort: Cohort is a group of subjects who share a defining characteristic

er_status_measured_by_ihc: 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)

er_status: Cancer cells are positive or negative for estrogen receptors

neoplasm_histologic_grade: Determined by pathology by looking the nature of the cells, do they look aggressive or not

her2_status_measured_by_snp6: To assess if the cancer positive for HER2 or not by using advance molecular techniques (Type of next generation sequencing)

her2_status: Whether the cancer is positive or negative for HER2

tumor_other_histologic_subtype: Type of the cancer based on microscopic examination of the cancer tissue

hormone_therapy: Whether or not the patient had hormonal as a treatment

inferred_menopausal_state: Whether the patient is is post menopausal or not

integrative_cluster: Molecular subtype of the cancer based on some gene expression

primary_tumor_laterality: Whether it is involving the right breast or the left breast

lymph_nodes_examined_positive: To take samples of the lymph node during the surgery and see if there were involved by the cancer

mutation_count: Number of gene that has relevant mutations

nottingham_prognostic_index: It is used to determine prognosis following surgery for breast cancer

oncotree_code: 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: Duration from the time of the intervention to death

overall_survival: Target variable wether the patient is alive of dead

pr_status: Cancer cells are positive or negative for progesterone receptors

radio_therapy: Whether or not the patient had radio as a treatment

3-gene_classifier_subtype: Three Gene classifier subtype

tumor_size: Tumor size measured by imaging techniques

tumor_stage: Stage of the cancer based on the involvement of surrounding structures, lymph nodes and distant spread

death_from_cancer: Wether the patient's death was due to cancer or not

Genetic Attributes:

The genetics part of the dataset contains m-RNA levels z-score for 331 genes, and mutation for 175 genes

m-RNA: The DNA molecules attached to each slide act as probes to detect gene expression, which is also known as the transcriptome or the set of messenger RNA (mRNA) transcripts expressed by a group of genes. To perform a microarray analysis, mRNA molecules are typically collected from both an experimental sample and a reference sample.

m-RNA z-score: For mRNA expression data, The calculations of the relative expression of an individual gene and tumor to the gene's expression distribution in a reference population is done. That reference population is all samples in the study . The returned value indicates the number of standard deviations away from the mean of expression in the reference population (Z-score). This measure is useful to determine whether a gene is up- or down-regulated relative to the normal samples or all other tumor samples.

Formula:

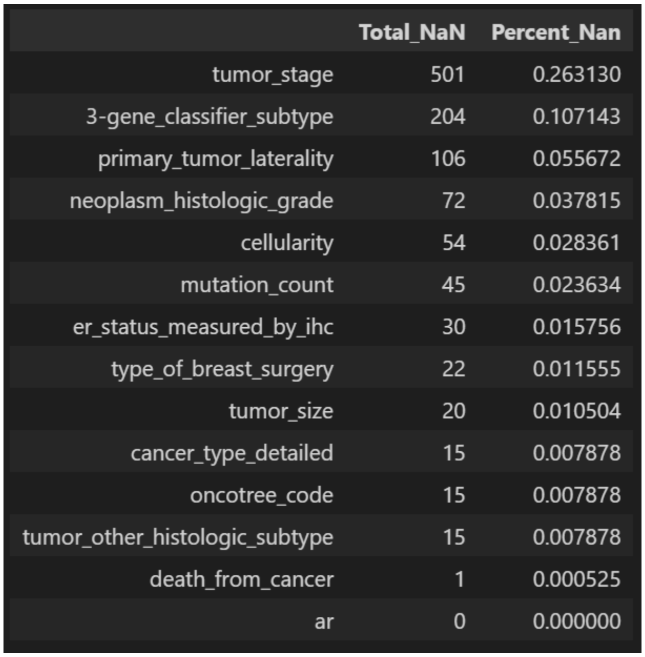
z = (expression in tumor sample - mean expression in reference sample) / standard deviation of expression in reference sample

III. Workflows

1. EDA & engineer features:

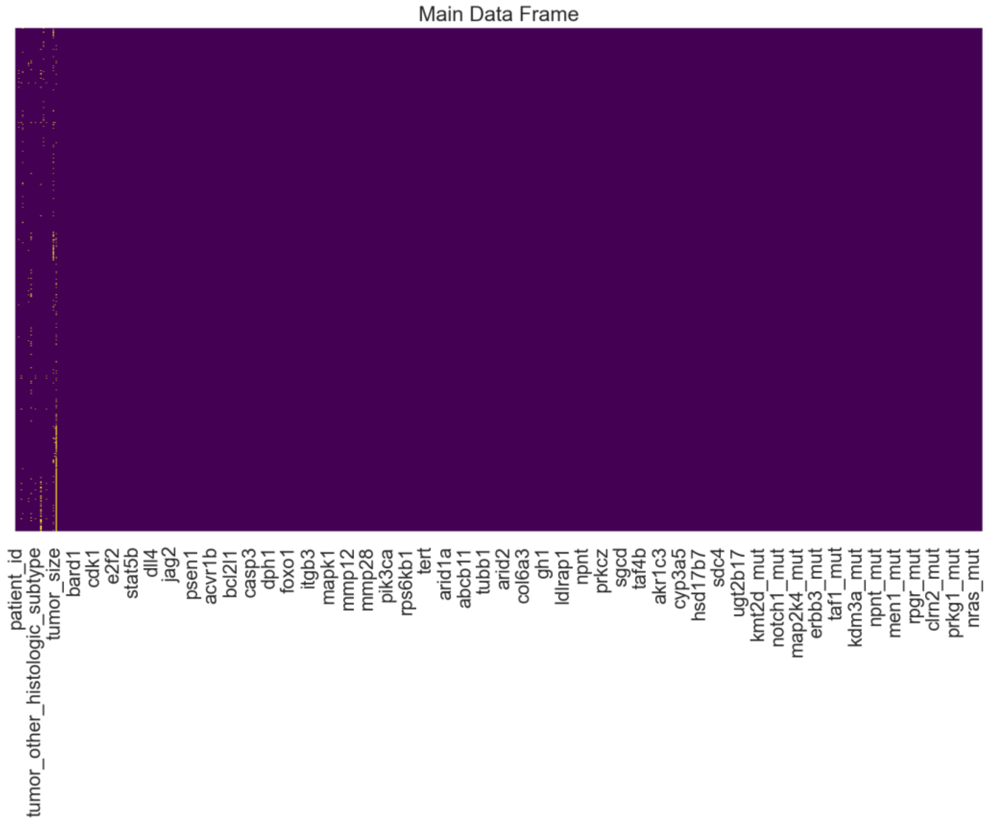
a) Preliminary cleaning of the dataset:

+ Finding missing data and the percentage of it in each column:



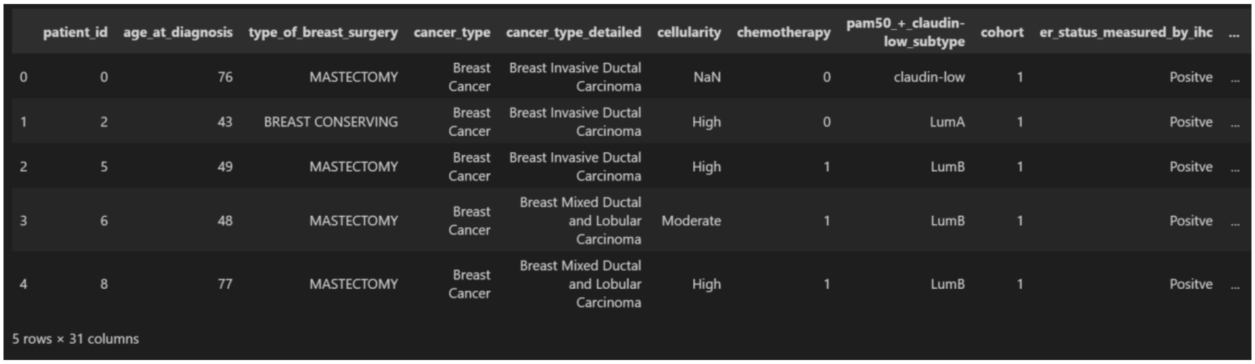
	Total_NaN	Percent_Nan
tumor_stage	501	0.263130
3-gene_classifier_subtype	204	0.107143
primary_tumor_laterality	106	0.055672
neoplasm_histologic_grade	72	0.037815
cellularity	54	0.028361
mutation_count	45	0.023634
er_status_measured_by_ihc	30	0.015756
type_of_breast_surgery	22	0.011555
tumor_size	20	0.010504
cancer_type_detailed	15	0.007878
oncotree_code	15	0.007878
tumor_other_histologic_subtype	15	0.007878
death_from_cancer	1	0.000525
ar	0	0.000000

+ Visualize of missing data with heatmap:



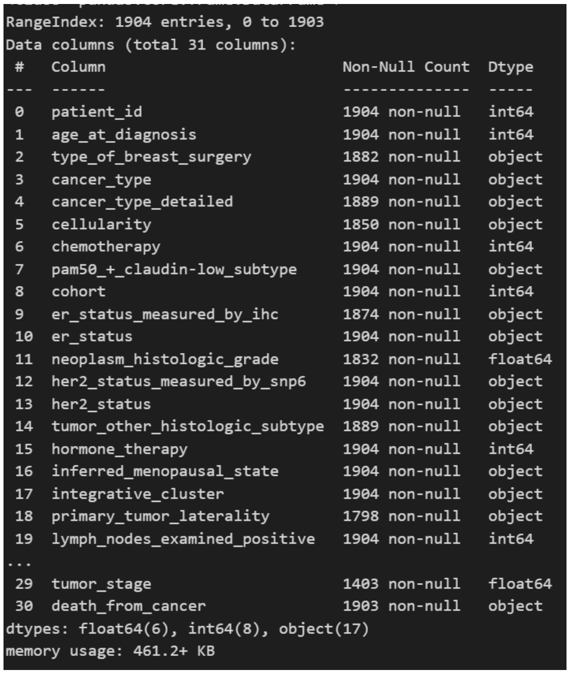
b) Relationship between clinical attributes and outcomes

+ Create a new data frame for clinical attributes only:



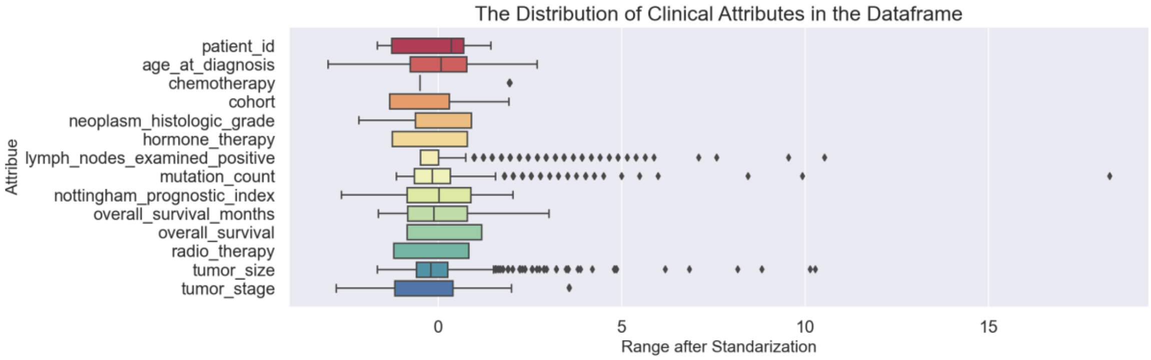
	patient_id	age_at_diagnosis	type_of_breast_surgery	cancer_type	cancer_type_detailed	cellularity	chemotherapy	pam50_+_claudin-low_subtype	cohort	er_status_measured_by_ihc	...
0	0	76	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	NaN	0	claudin-low	1	Positve	...
1	2	43	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	High	0	LumA	1	Positve	...
2	5	49	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	High	1	LumB	1	Positve	...
3	6	48	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Moderate	1	LumB	1	Positve	...
4	8	77	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	High	1	LumB	1	Positve	...

5 rows × 31 columns



```
RangeIndex: 1904 entries, 0 to 1903
Data columns (total 31 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   patient_id                      1904 non-null   int64
 1   age_at_diagnosis                1904 non-null   int64
 2   type_of_breast_surgery          1882 non-null   object
 3   cancer_type                     1904 non-null   object
 4   cancer_type_detailed            1889 non-null   object
 5   cellularity                     1850 non-null   object
 6   chemotherapy                    1904 non-null   int64
 7   pam50_+_claudin-low_subtype     1904 non-null   object
 8   cohort                          1904 non-null   int64
 9   er_status_measured_by_ihc       1874 non-null   object
 10  er_status                       1904 non-null   object
 11  neoplasm_histologic_grade       1832 non-null   float64
 12  her2_status_measured_by_snp6    1904 non-null   object
 13  her2_status                     1904 non-null   object
 14  tumor_other_histologic_subtype  1889 non-null   object
 15  hormone_therapy                 1904 non-null   int64
 16  inferred_menopausal_state       1904 non-null   object
 17  integrative_cluster             1904 non-null   object
 18  primary_tumor_laterality        1798 non-null   object
 19  lymph_nodes_examined_positive   1904 non-null   int64
...
 29  tumor_stage                     1403 non-null   float64
 30  death_from_cancer               1903 non-null   object
dtypes: float64(6), int64(8), object(17)
memory usage: 461.2+ KB
```

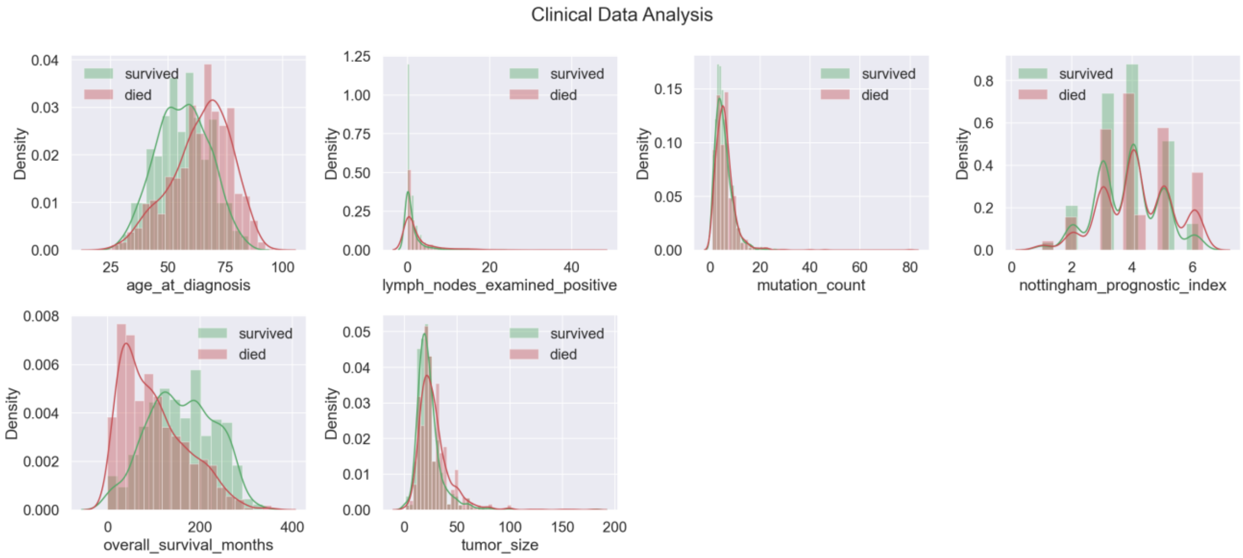
+ Plot a boxplot to show the distribution of clinical attributes in the data frame:

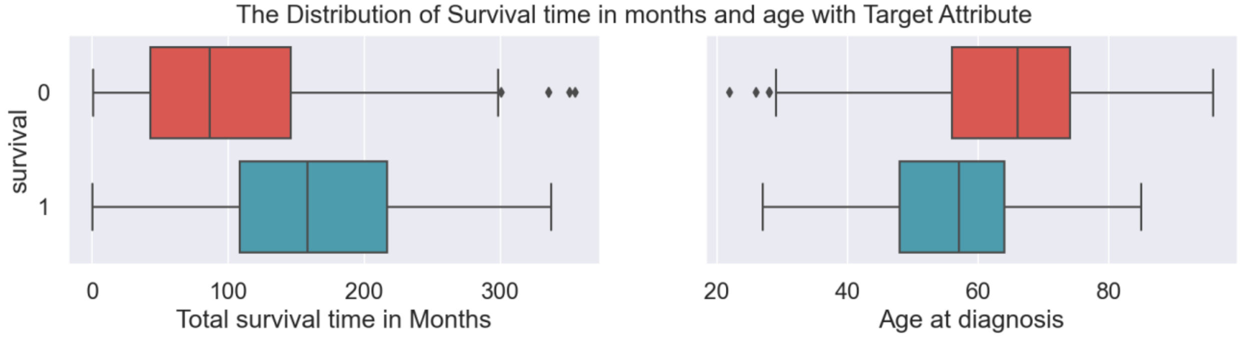


Observations/ Conclusions:

- For the distribution of all numerical data, 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 like lymph_nodes_examined_positive, mutation_count, and tumor_size.
- Should keep the outliers, as they are very important in healthcare data.

+ Plots to show the Distribution of the Two Target Classes in Numerical Clinical Columns in the Data frame:

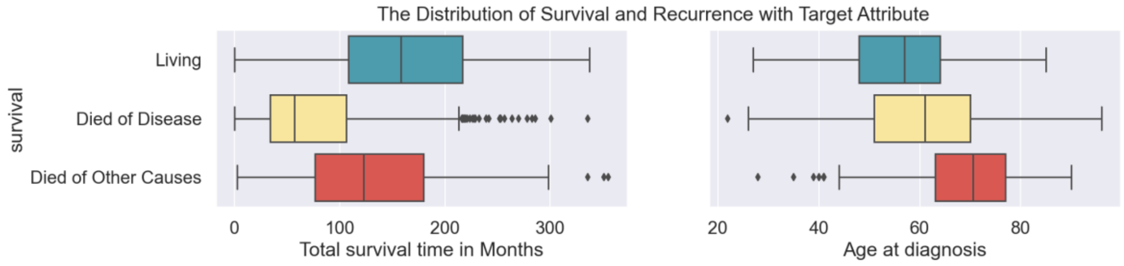




Observations/ Conclusions:

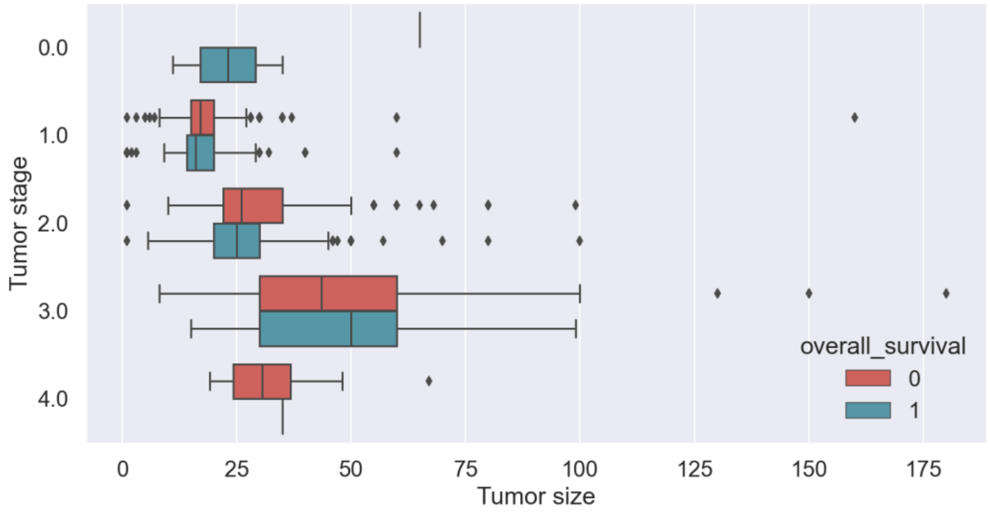
- To compare between the two classes of patients who survived and patients who did not, we can see the difference between the two distributions in age_at_diagnosis column, as patients who were younger when diagnosed with breast cancer were more likely to survive.

- Also, the duration from the time of the intervention to death or to current time is longer in the patients who survive. This means that patients are either dying early from breast cancer or surviving.



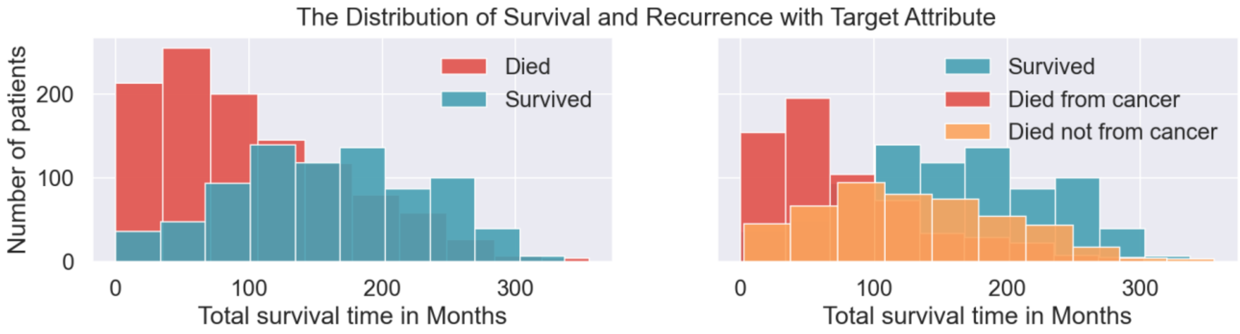
Observations/ Conclusions:

- The variable 'death_from_cancer' shows us if the patient is alive or died from cancer or its complications or died of other causes.
- From the distribution of the three classes, we can see that the median of the survival time in months of patients who died from breast cancer is low compared to the other two classes, and its distribution os right-skewed with a lot of outliers.
- Also, patients who died from other causes than cancer tend to be older than the other two classes. The distribution of it is left-skewed with some younger outliers.



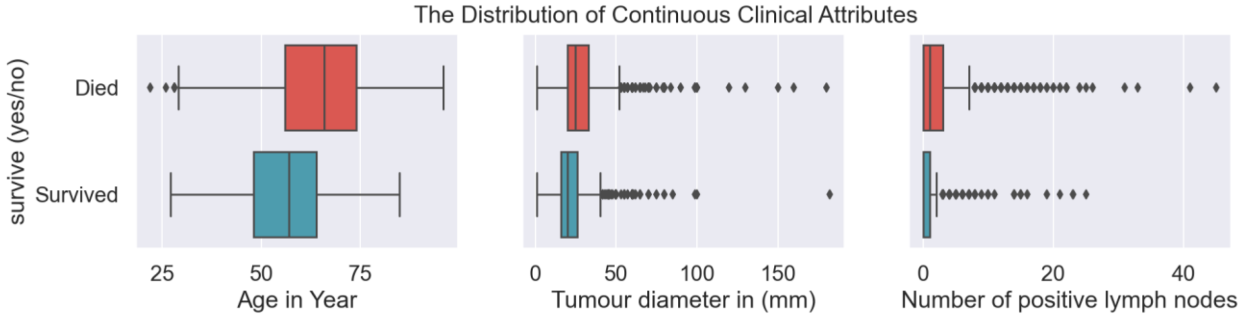
Observations/ Conclusions:

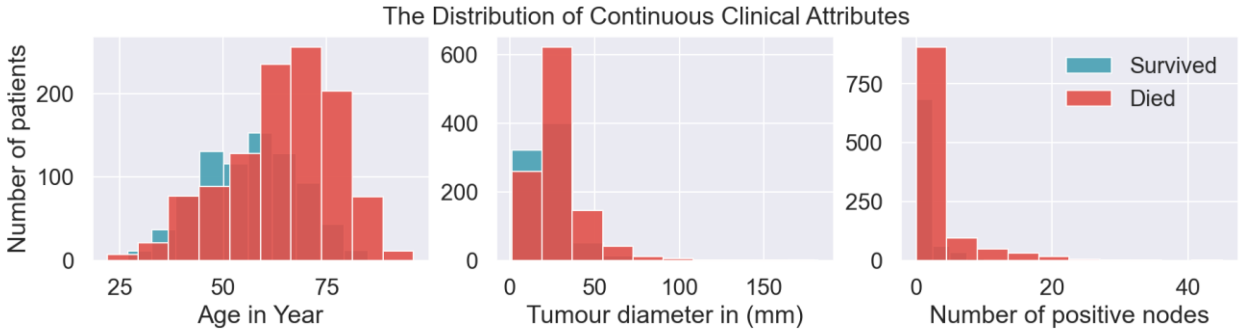
- As the Tumor stage increases the tumor size increases as well. Also, if lower tumor stages the probability of survival is higher than when the patient reaches the fourth stage.



Observations/ Conclusions:

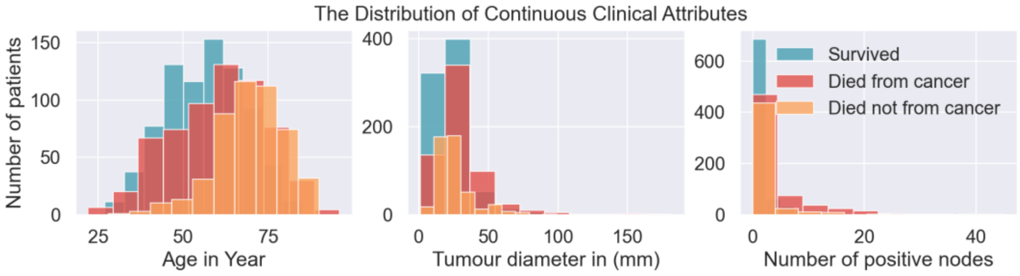
- When the total survival time in months increases, the probability of survival increases as well, and the probability of dying from reasons other than cancer decrease with time slightly.

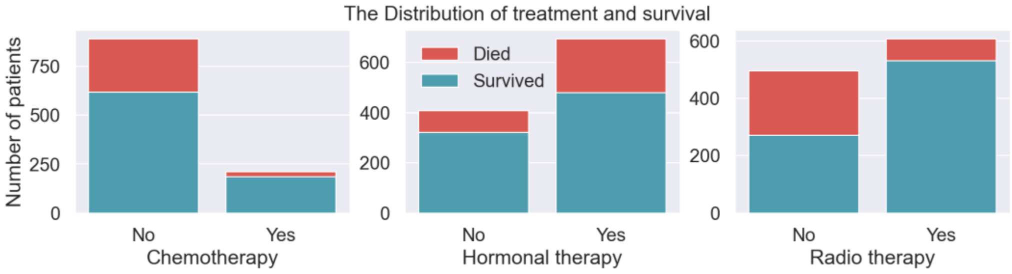


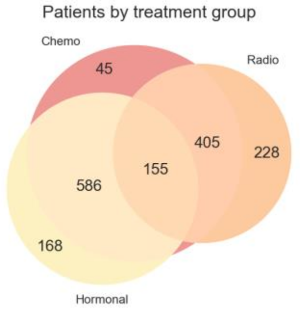


Observations/ Conclusions:

- The median of tumor size and the number of positive lymph nodes is lower in the survived class than the died class.

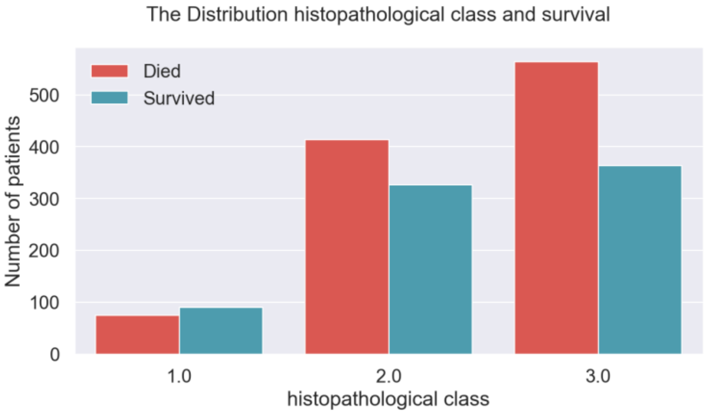




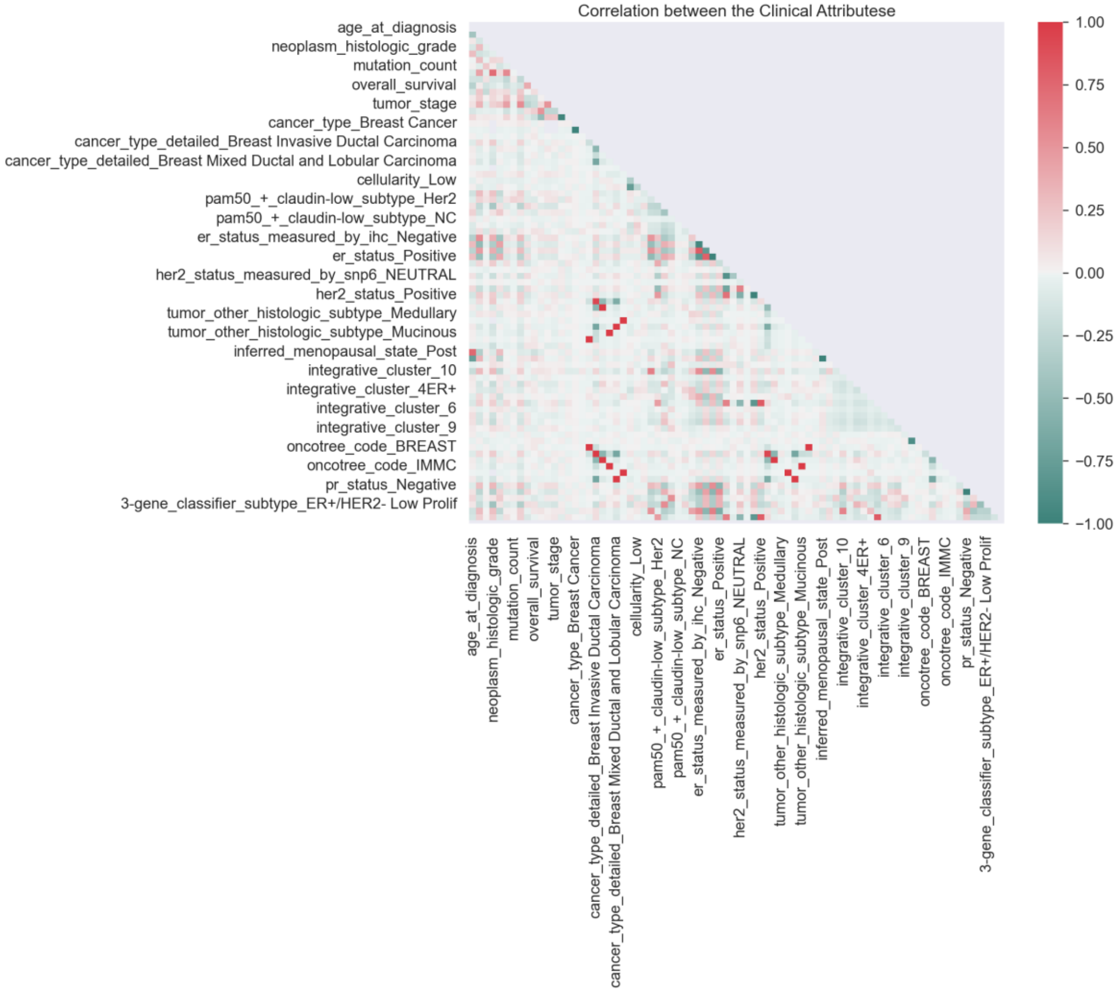


Observations/ Conclusions:

- Venn diagram for the three different treatments for breast cancer and the distribution of patients amongst them.
- We can see that most patients either have chemo and hormonal therapy or chemo and radio therapy.
- 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.
- There were 289 patients and their survival rate was slightly lower than the rest of patients.



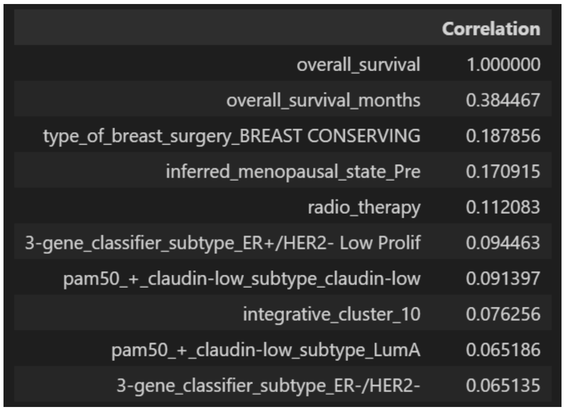
+ Plot to show the correlation between the Clinical Attributes:



Observations/ Conclusions:

- It can see that there is high correlation between some of the columns.

+ Correlation between the Clinical Attributes and survival:



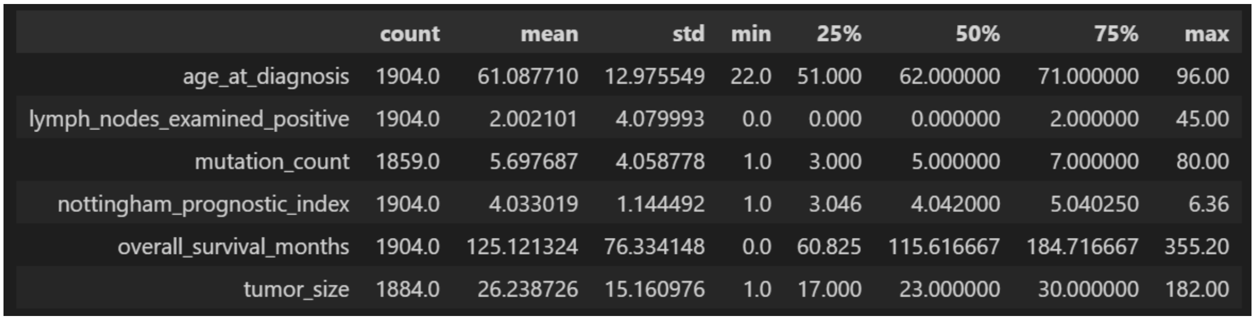
	Correlation
overall_survival	1.000000
overall_survival_months	0.384467
type_of_breast_surgery_BREAST CONSERVING	0.187856
inferred_menopausal_state_Pre	0.170915
radio_therapy	0.112083
3-gene_classifier_subtype_ER+/HER2- Low Prolif	0.094463
pam50_+_claudin-low_subtype_claudin-low	0.091397
integrative_cluster_10	0.076256
pam50_+_claudin-low_subtype_LumA	0.065186
3-gene_classifier_subtype_ER-/HER2-	0.065135

Observations/ Conclusions:

- There is a positive correlation between survival and overall survival in months, conserving surgery type, pre menopaus status.
- But there is a negative correlation between survival and lymph nodes examined positive, mastectomy surgery type, tumor stage, and age at diagnosis.

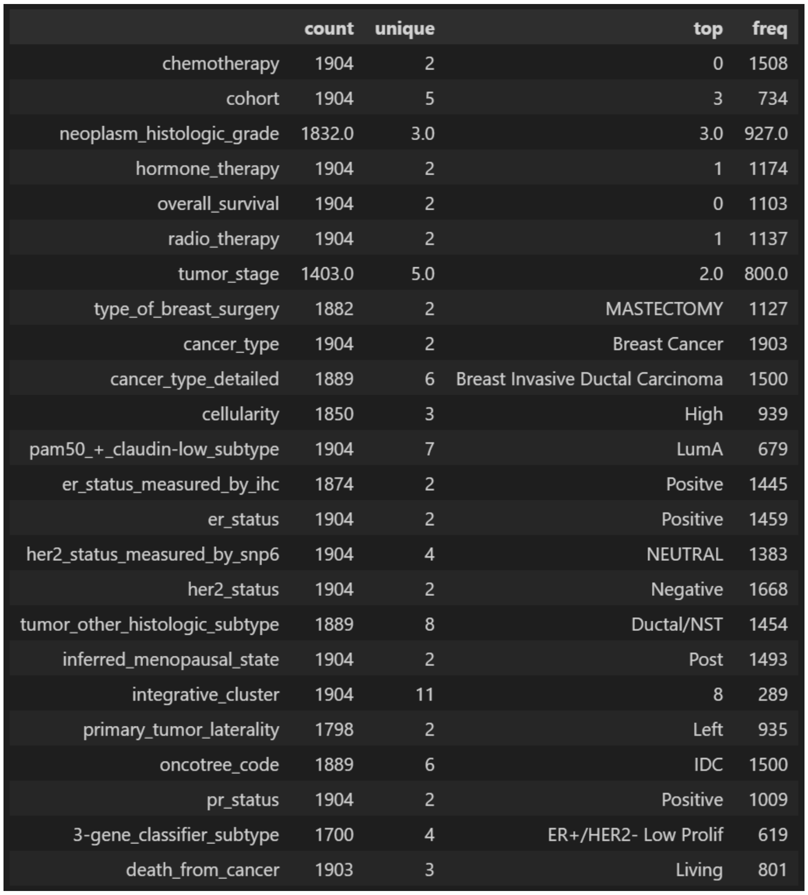
+ Statistical Summaries of Clinical Columns in the Data frame:

- Statistical summary for numerical clinical attributes:



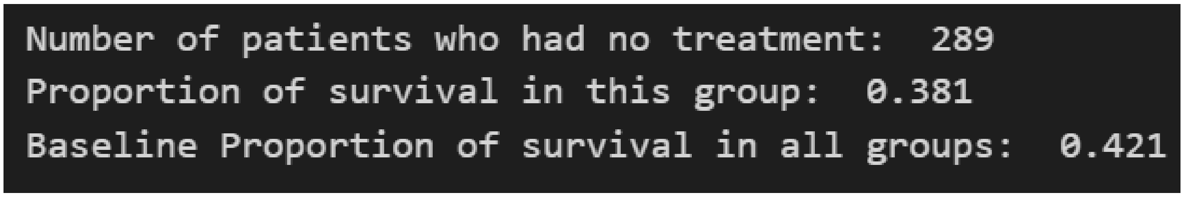
	count	mean	std	min	25%	50%	75%	max
age_at_diagnosis	1904.0	61.087710	12.975549	22.0	51.000	62.000000	71.000000	96.00
lymph_nodes_examined_positive	1904.0	2.002101	4.079993	0.0	0.000	0.000000	2.000000	45.00
mutation_count	1859.0	5.697687	4.058778	1.0	3.000	5.000000	7.000000	80.00
nottingham_prognostic_index	1904.0	4.033019	1.144492	1.0	3.046	4.042000	5.040250	6.36
overall_survival_months	1904.0	125.121324	76.334148	0.0	60.825	115.616667	184.716667	355.20
tumor_size	1884.0	26.238726	15.160976	1.0	17.000	23.000000	30.000000	182.00

- Statistical summary for categorical clinical attributes:



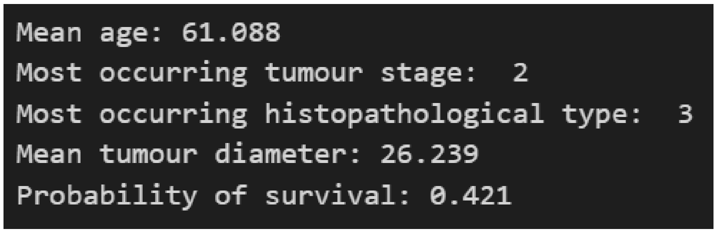
	count	unique	top	freq
chemotherapy	1904	2	0	1508
cohort	1904	5	3	734
neoplasm_histologic_grade	1832.0	3.0	3.0	927.0
hormone_therapy	1904	2	1	1174
overall_survival	1904	2	0	1103
radio_therapy	1904	2	1	1137
tumor_stage	1403.0	5.0	2.0	800.0
type_of_breast_surgery	1882	2	MASTECTOMY	1127
cancer_type	1904	2	Breast Cancer	1903
cancer_type_detailed	1889	6	Breast Invasive Ductal Carcinoma	1500
cellularity	1850	3	High	939
pam50_+_claudin-low_subtype	1904	7	LumA	679
er_status_measured_by_ihc	1874	2	Positve	1445
er_status	1904	2	Positive	1459
her2_status_measured_by_snp6	1904	4	NEUTRAL	1383
her2_status	1904	2	Negative	1668
tumor_other_histologic_subtype	1889	8	Ductal/NST	1454
inferred_menopausal_state	1904	2	Post	1493
integrative_cluster	1904	11	8	289
primary_tumor_laterality	1798	2	Left	935
oncotree_code	1889	6	IDC	1500
pr_status	1904	2	Positive	1009
3-gene_classifier_subtype	1700	4	ER+/HER2- Low Prolif	619
death_from_cancer	1903	3	Living	801

- Statistics for the no treatment group and comparison with the baseline:



```
Number of patients who had no treatment:  289
Proportion of survival in this group:  0.381
Baseline Proportion of survival in all groups:  0.421
```

+ Characteristics of the average member of the population:

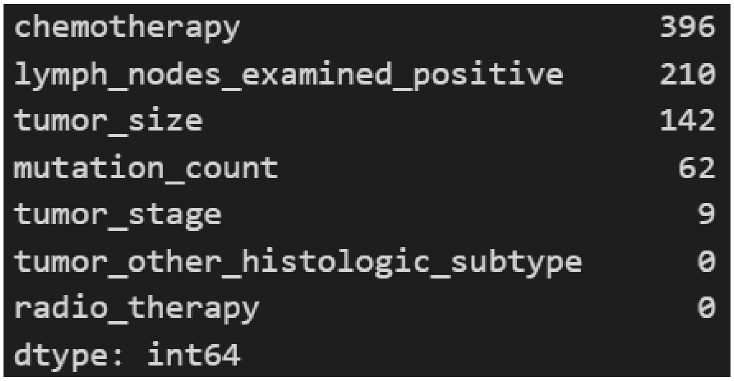


```
Mean age: 61.088
Most occurring tumour stage:  2
Most occurring histopathological type:  3
Mean tumour diameter: 26.239
Probability of survival: 0.421
```

Observations/ Conclusions:

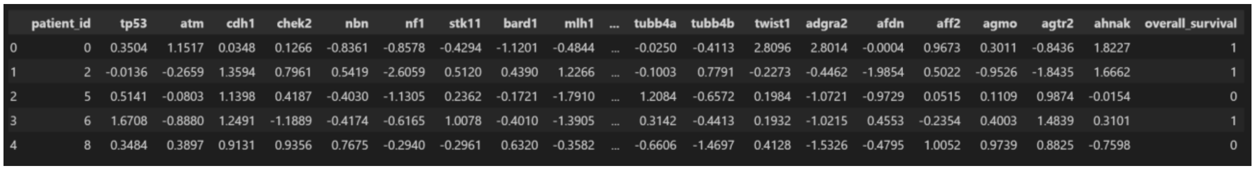
- The average breast cancer patient in the dataset is a 61-year-old women with a stage 2 tumor with 2 lymph nodes examined positive, with a mean tumor size of 26 mm.
- The patient has a probability of 76% of not having chemotherapy as a treatment, but only hormonal and radiotherapy with surgery.

+ Number of outliers in each clinical feature:



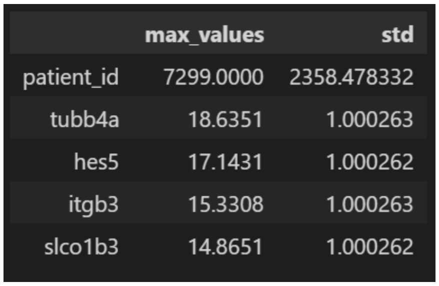
```
chemotherapy                      396
lymph_nodes_examined_positive     210
tumor_size                        142
mutation_count                     62
tumor_stage                         9
tumor_other_histologic_subtype      0
radio_therapy                       0
dtype: int64
```

c) Relationship between genetic attributes and outcomes



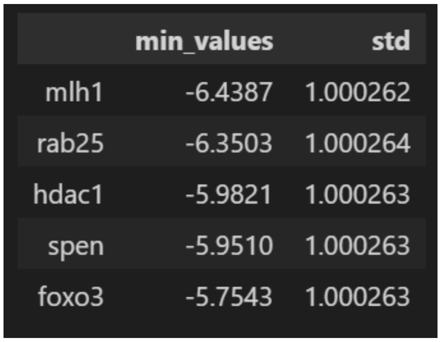
	patient_id	tp53	atm	cdh1	chek2	nbn	nf1	stk11	bard1	mlh1	...	tubb4a	tubb4b	twist1	adgra2	afdn	aff2	agmo	agtr2	ahnak	overall_survival
0	0	0.3504	1.1517	0.0348	0.1266	-0.8361	-0.8578	-0.4294	-1.1201	-0.4844	...	-0.0250	-0.4113	2.8096	2.8014	-0.0004	0.9673	0.3011	-0.8436	1.8227	1
1	2	-0.0136	-0.2659	1.3594	0.7961	0.5419	-2.6059	0.5120	0.4390	1.2266	...	-0.1003	0.7791	-0.2273	-0.4462	-1.9854	0.5022	-0.9526	-1.8435	1.6662	1
2	5	0.5141	-0.0803	1.1398	0.4187	-0.4030	-1.1305	0.2362	-0.1721	-1.7910	...	1.2084	-0.6572	0.1984	-1.0721	-0.9729	0.0515	0.1109	0.9874	-0.0154	0
3	6	1.6708	-0.8880	1.2491	-1.1889	-0.4174	-0.6165	1.0078	-0.4010	-1.3905	...	0.3142	-0.4413	0.1932	-1.0215	0.4553	-0.2354	0.4003	1.4839	0.3101	1
4	8	0.3484	0.3897	0.9131	0.9356	0.7675	-0.2940	-0.2961	0.6320	-0.3582	...	-0.6606	-1.4697	0.4128	-1.5326	-0.4795	1.0052	0.9739	0.8825	-0.7598	0

+ Find maximum values and standard deviation in each column, standard deviation is always 1 because the datapoints are z-scores:



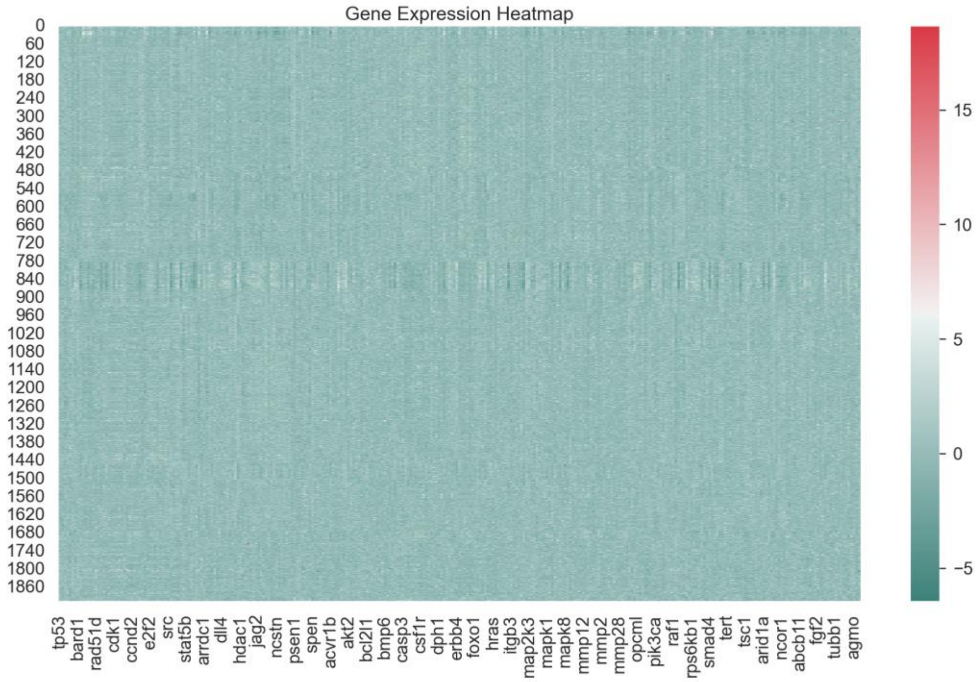
	max_values	std
patient_id	7299.0000	2358.478332
tubb4a	18.6351	1.000263
hes5	17.1431	1.000262
itgb3	15.3308	1.000263
slco1b3	14.8651	1.000262

+ Find minimum values and standard deviation in each column, standard deviation is always 1 because the datapoints are z-scores:

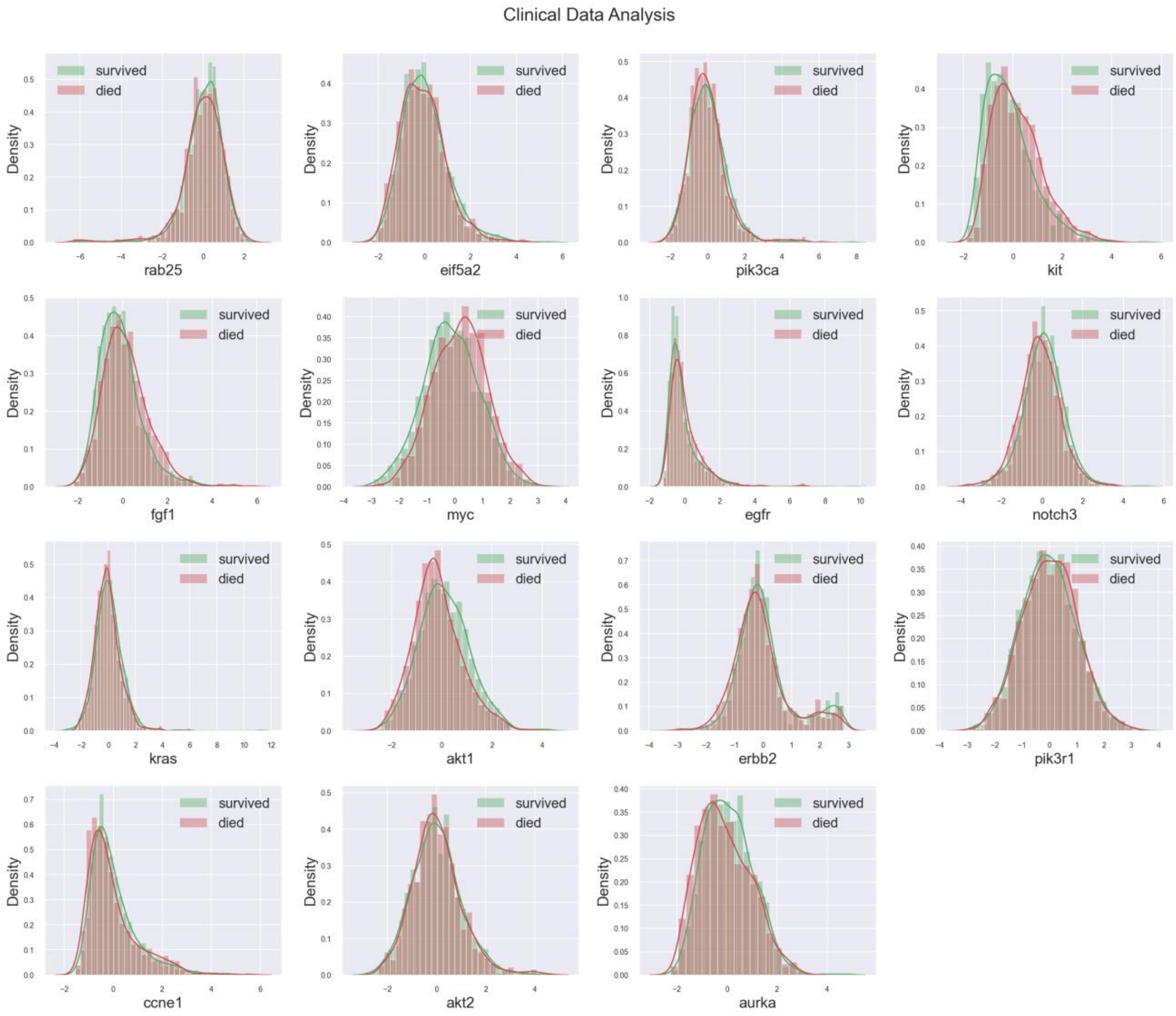


	min_values	std
mlh1	-6.4387	1.000262
rab25	-6.3503	1.000264
hdac1	-5.9821	1.000263
spen	-5.9510	1.000263
foxo3	-5.7543	1.000263

+ Plot a heatmap for visualizing the mRNA values:



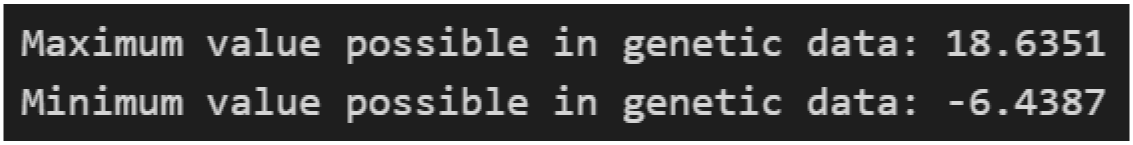
+ Plot to show the distribution of the “overall_survival” column:



Observations/ Conclusions:

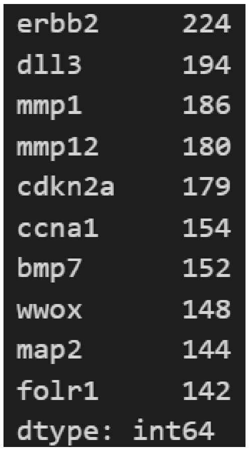
- The distribution of data in the two classes of survival are very similar with few outliers in some genes.

+ Find the maximum and minimum value possible in the genetic data:



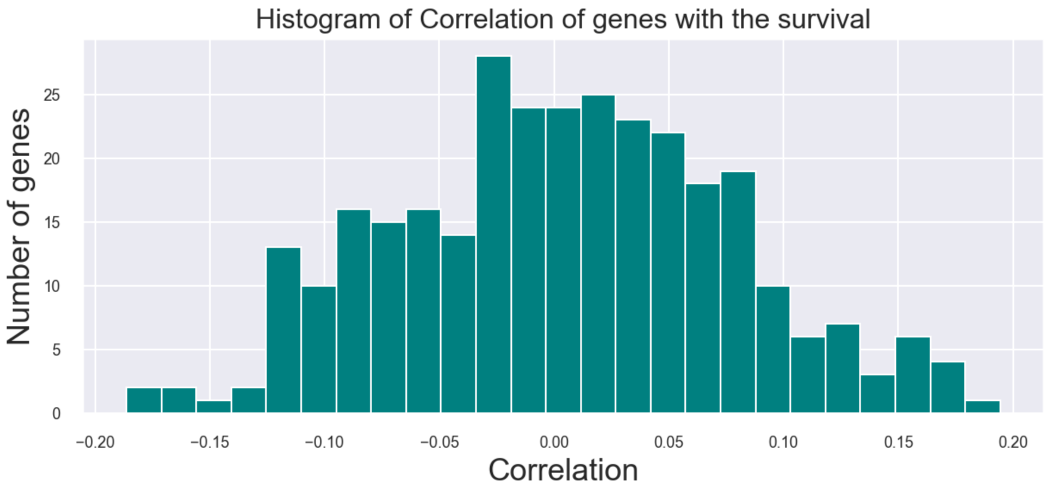
```
Maximum value possible in genetic data: 18.6351
Minimum value possible in genetic data: -6.4387
```

+ Number of outliers in the top 10 genetic features:

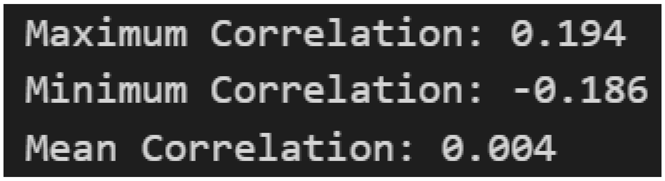


```
erbb2      224
dll3       194
mmp1       186
mmp12      180
cdkn2a     179
ccna1      154
bmp7       152
wwox       148
map2       144
folr1      142
dtype: int64
```

+ Plot to show the correlation of between the genetic Attributes and outcome:



+ Find the Maximum Correlation, Minimum Correlation and Mean Correlation:

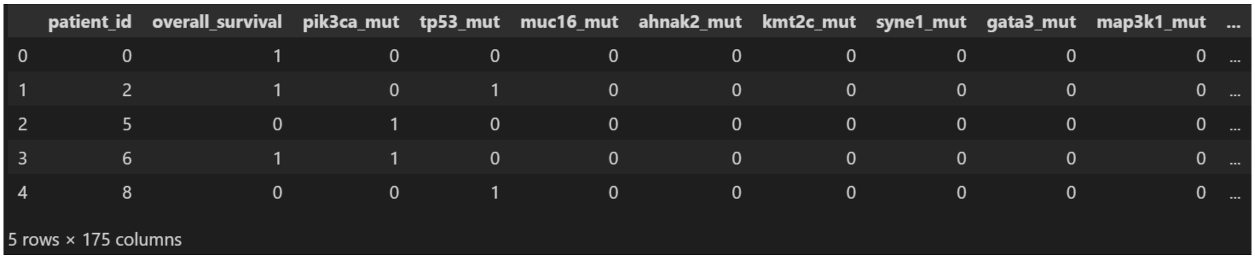


```
Maximum Correlation: 0.194
Minimum Correlation: -0.186
Mean Correlation: 0.004
```

Observations/ Conclusions:

- The correlation between our target and the genetic features shows that most features do not actually correlate.

d) Relationship between genetic mutation attributes and outcomes



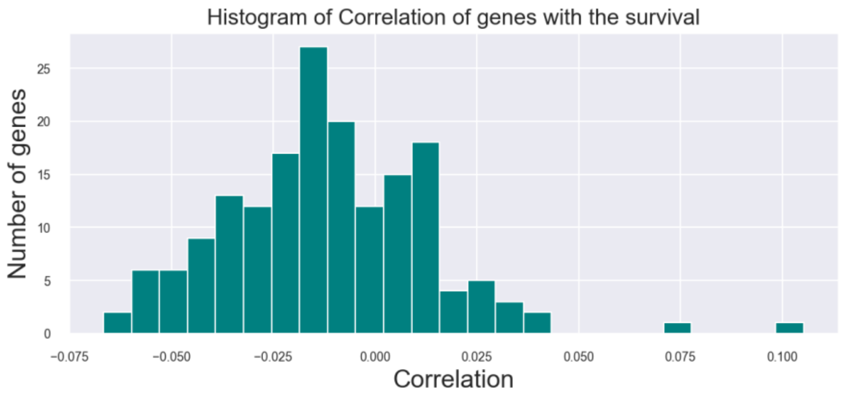
	patient_id	overall_survival	pik3ca_mut	tp53_mut	muc16_mut	ahnak2_mut	kmt2c_mut	syne1_mut	gata3_mut	map3k1_mut	...
0	0	1	0	0	0	0	0	0	0	0	...
1	2	1	0	1	0	0	0	0	0	0	...
2	5	0	1	0	0	0	0	0	0	0	...
3	6	1	1	0	0	0	0	0	0	0	...
4	8	0	0	1	0	0	0	0	0	0	...

5 rows × 175 columns

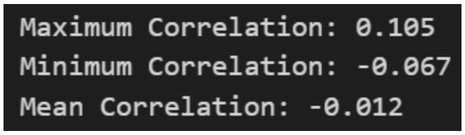
Observations/ Conclusions:

- Some genes had much more mutations than other genes. For example: PIK3CA (coding mutations in 40.1% of the samples) and TP53 (35.4%) dominated the mutation landscape.
- Only five other genes harbored coding mutations in at least 10% of the samples: MUC16 (16.8%); AHNAK2 (16.2%); SYNE1 (12.0%); KMT2C (also known as MLL3; 11.4%) and GATA3 (11.1%).

+ Plot histogram of variation using standard deviation as a measure to show the correlation of genes with the survival:



+ Find the Maximum Correlation, Minimum Correlation and Mean Correlation:



```
Maximum Correlation: 0.105
Minimum Correlation: -0.067
Mean Correlation: -0.012
```

Observations/ Conclusions:

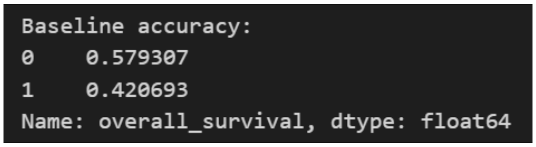
- No correlation at all between survival and mutations, as we changed the mutation to 0s and 1s instead of 0s if there is no mutations and the kind of mutation if there is a mutation.

- Decided to exclude the mutations from the modeling part for now, and maybe include it later when analyze them in more detail.

2. Data preprocessing:

+ Use a Stratified K fold because we need the distribution of the two classes in all of the folds to be the same.

+ Calculate baseline accuracy dividing the unique value count to the value count of the “overall_survival” feature:



```
Baseline accuracy:
0    0.579307
1    0.420693
Name: overall_survival, dtype: float64
```

+ Drop the “patient_id” column because not needed.

+ Drop the “death_from_cancer” and “overall_survival_months” columns because we only need the “overall_survival” column.

+ Get dummies for all categorical columns by Pandas get_dummies.

+ Split data into 67% for training, 33% for testing.

+ Using Stratify for y because we need the distribution of the two classes to be equal in train and test sets.

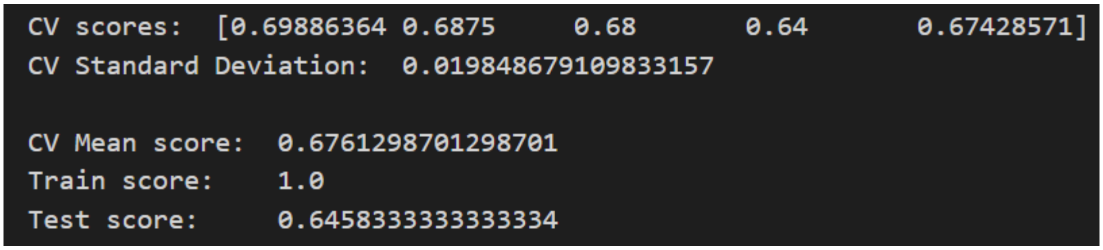
3. Classification Models and Evaluation:

a) Classification with only clinical attributes:

+ K Nearest Neighbors Classifier:

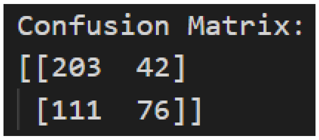
- Set the parameters range for hyperparameter tunning to identify the best parameters for estimation and create an instance of Grid Search CV model and fit the classification model to find the best estimator.

- Cross Validation Score:



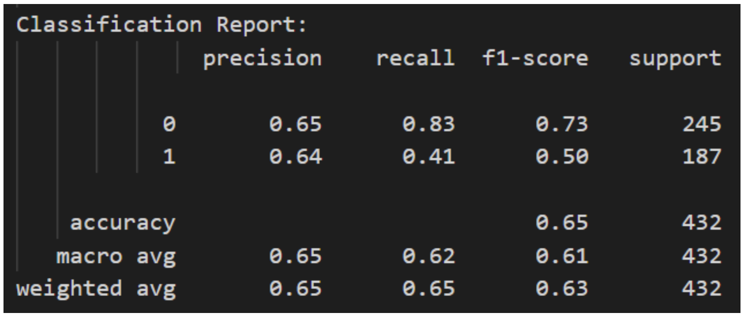
```
CV scores:  [0.69886364 0.6875     0.68       0.64       0.67428571]
CV Standard Deviation:  0.019848679109833157

CV Mean score:  0.6761298701298701
Train score:    1.0
Test score:     0.6458333333333334
```

- Confusion Matrix:



```
Confusion Matrix:
[[203  42]
 [111  76]]
```

- Classification Report:

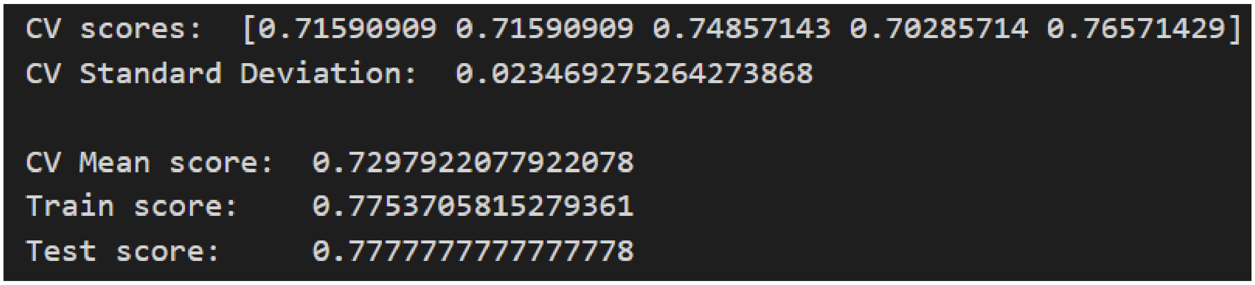


```
Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.83      0.73       245
           1       0.64      0.41      0.50       187

    accuracy                           0.65       432
   macro avg       0.65      0.62      0.61       432
weighted avg       0.65      0.65      0.63       432
```

+ Logistic Regression:

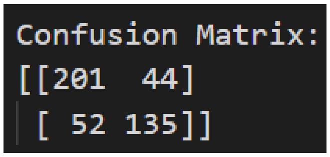
- Set the parameters range for hyperparameter tunning to identify the best parameters for estimation and create an instance of Grid Search CV model and fit the classification model to find the best estimator.

- Cross Validation Score:



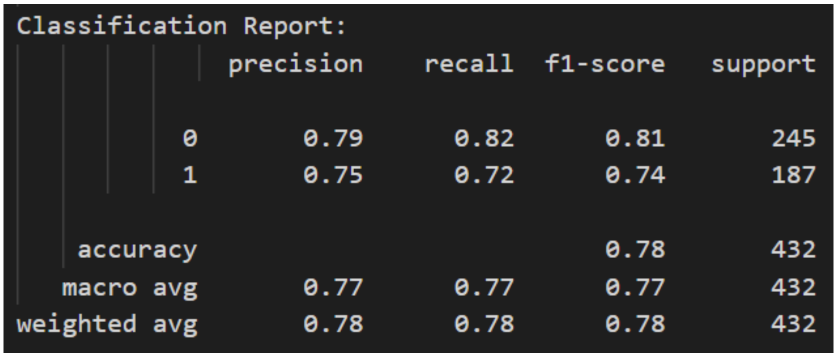
```
CV scores:  [0.71590909 0.71590909 0.74857143 0.70285714 0.76571429]
CV Standard Deviation:  0.023469275264273868

CV Mean score:  0.7297922077922078
Train score:    0.7753705815279361
Test score:     0.7777777777777778
```

- Confusion Matrix:



```
Confusion Matrix:
[[201  44]
 [ 52 135]]
```

- Classification Report:

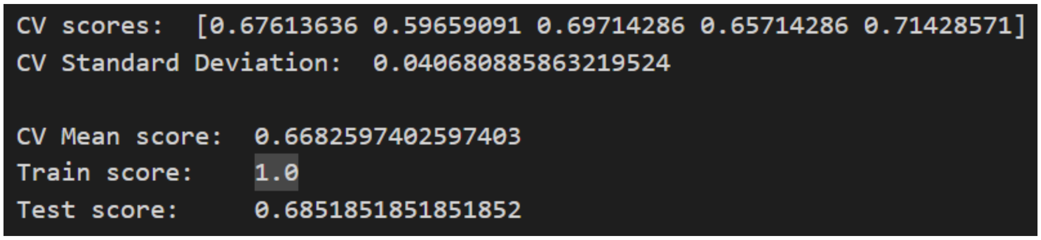


```
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.82      0.81       245
           1       0.75      0.72      0.74       187

    accuracy                           0.78       432
   macro avg       0.77      0.77      0.77       432
weighted avg       0.78      0.78      0.78       432
```

+ Decision Tree Classifier

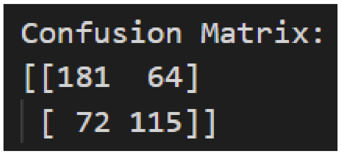
- Set the parameters range for hyperparameter tunning to identify the best parameters for estimation and create an instance of Grid Search CV model and fit the classification model to find the best estimator.

- Cross Validation Score:



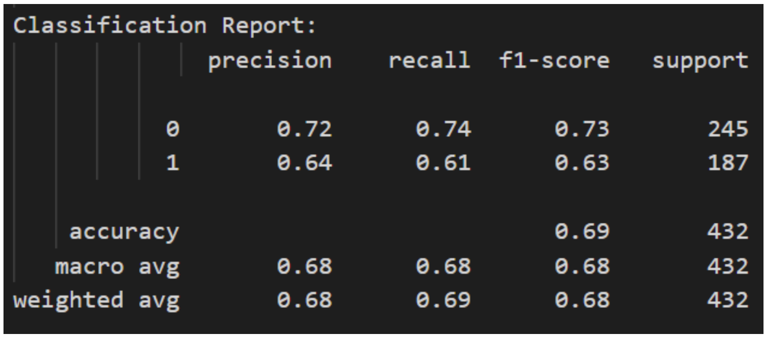
```
CV scores:  [0.67613636 0.59659091 0.69714286 0.65714286 0.71428571]
CV Standard Deviation:  0.040680885863219524

CV Mean score:  0.6682597402597403
Train score:    1.0
Test score:     0.6851851851851852
```

- Confusion Matrix:



```
Confusion Matrix:
[[181  64]
 [ 72 115]]
```

- Classification Report:

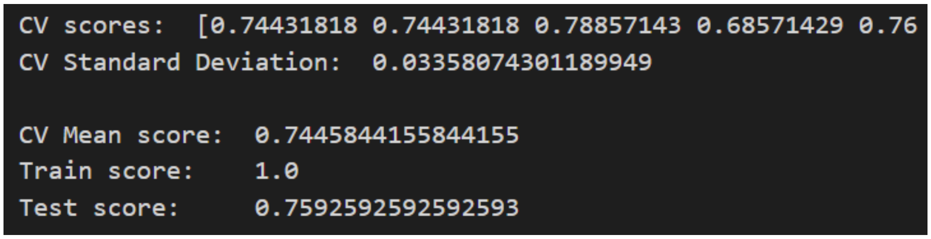


```
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.74      0.73       245
           1       0.64      0.61      0.63       187

    accuracy                           0.69       432
   macro avg       0.68      0.68      0.68       432
weighted avg       0.68      0.69      0.68       432
```

+ Random Forest Classifier

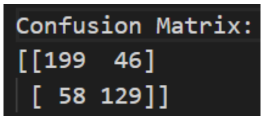
- Set the parameters range for hyperparameter tunning to identify the best parameters for estimation and create an instance of Grid Search CV model and fit the classification model to find the best estimator.

- Cross Validation Score:



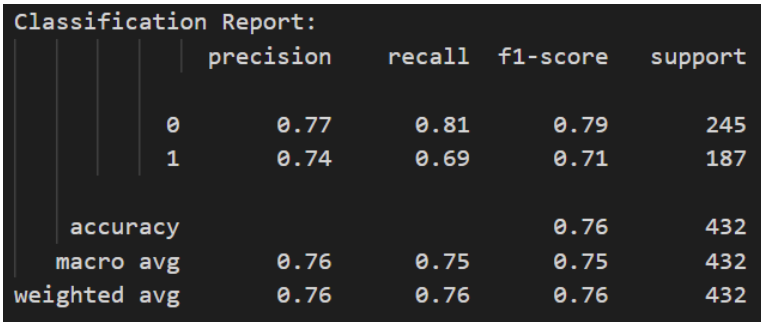
```
CV scores:  [0.74431818 0.74431818 0.78857143 0.68571429 0.76
CV Standard Deviation:  0.03358074301189949

CV Mean score:  0.7445844155844155
Train score:    1.0
Test score:     0.7592592592592593
```

- Confusion Matrix:



```
Confusion Matrix:
[[199  46]
 [ 58 129]]
```

- Classification Report:

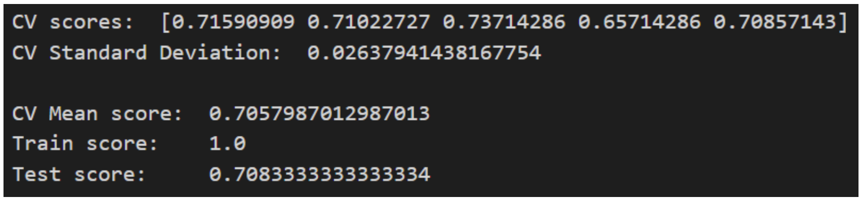


```
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.81      0.79       245
           1       0.74      0.69      0.71       187

    accuracy                           0.76       432
   macro avg       0.76      0.75      0.75       432
weighted avg       0.76      0.76      0.76       432
```

+ Extra Trees Classifier

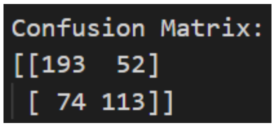
- Set the parameters range for hyperparameter tunning to identify the best parameters for estimation and create an instance of Grid Search CV model and fit the classification model to find the best estimator.

- Cross Validation Score:



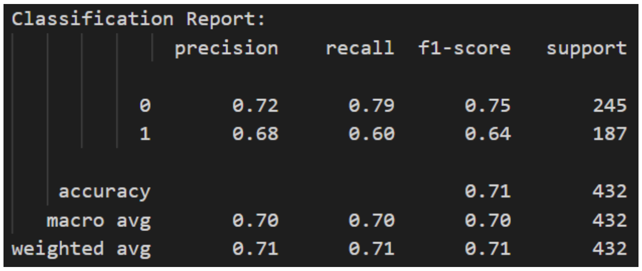
```
CV scores:  [0.71590909 0.71022727 0.73714286 0.65714286 0.70857143]
CV Standard Deviation:  0.02637941438167754

CV Mean score:  0.7057987012987013
Train score:    1.0
Test score:     0.7083333333333334
```

- Confusion Matrix:



```
Confusion Matrix:
[[193  52]
 [ 74 113]]
```

- Classification Report:

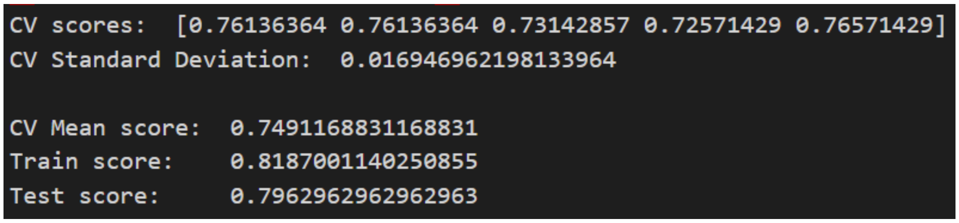


```
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.79      0.75       245
           1       0.68      0.60      0.64       187

    accuracy                           0.71       432
   macro avg       0.70      0.70      0.70       432
weighted avg       0.71      0.71      0.71       432
```

+ AdaBoost Classifier

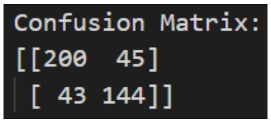
- Set the parameters range for hyperparameter tunning to identify the best parameters for estimation and create an instance of Grid Search CV model and fit the classification model to find the best estimator.

- Cross Validation Score:



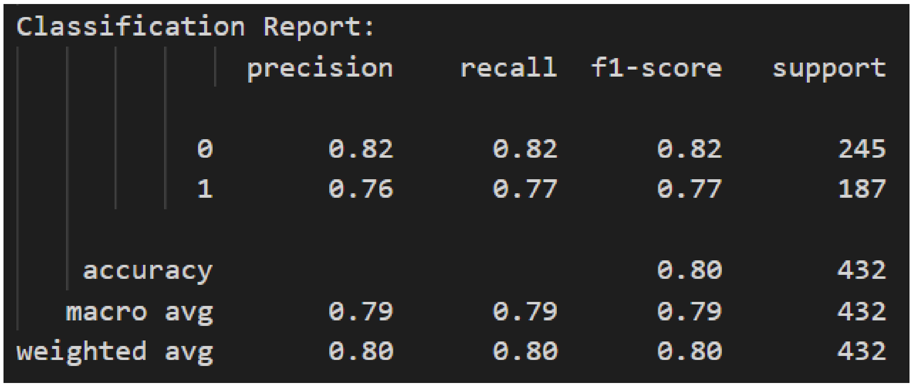
```
CV scores:  [0.76136364 0.76136364 0.73142857 0.72571429 0.76571429]
CV Standard Deviation:  0.016946962198133964

CV Mean score:  0.7491168831168831
Train score:    0.8187001140250855
Test score:     0.7962962962962963
```

- Confusion Matrix:



```
Confusion Matrix:
[[200  45]
 [ 43 144]]
```

- Classification Report:

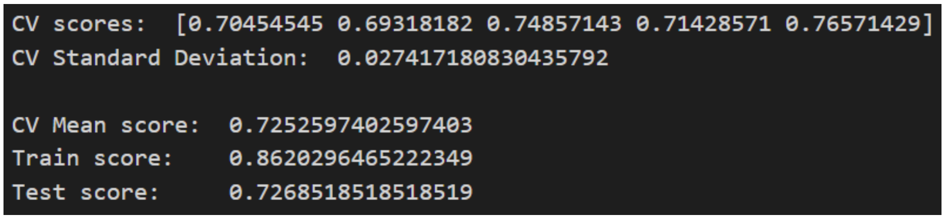


```
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.82      0.82       245
           1       0.76      0.77      0.77       187

    accuracy                           0.80       432
   macro avg       0.79      0.79      0.79       432
weighted avg       0.80      0.80      0.80       432
```

+ SVC Classifier

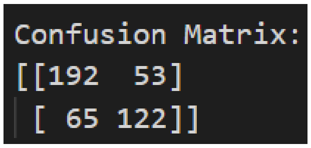
- Set the parameters range for hyperparameter tunning to identify the best parameters for estimation and create an instance of Grid Search CV model and fit the classification model to find the best estimator.

- Cross Validation Score:



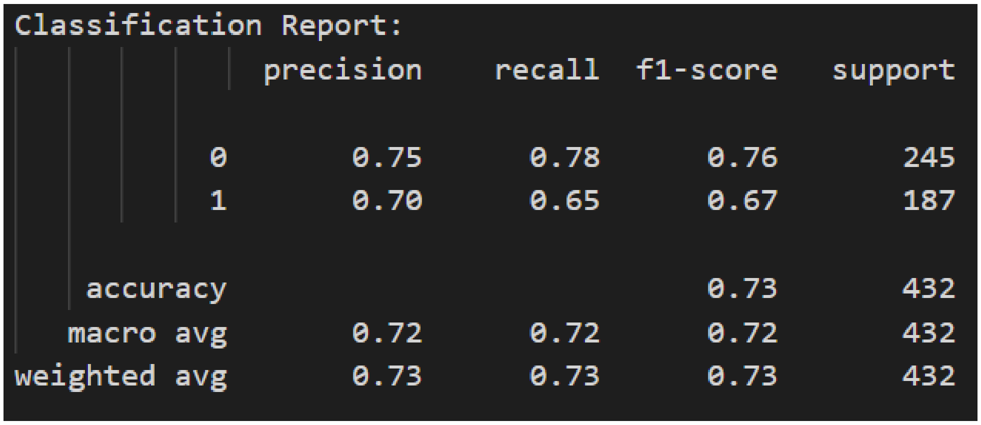
```
CV scores:  [0.70454545 0.69318182 0.74857143 0.71428571 0.76571429]
CV Standard Deviation:  0.027417180830435792

CV Mean score:  0.7252597402597403
Train score:    0.8620296465222349
Test score:     0.7268518518518519
```

- Confusion Matrix:



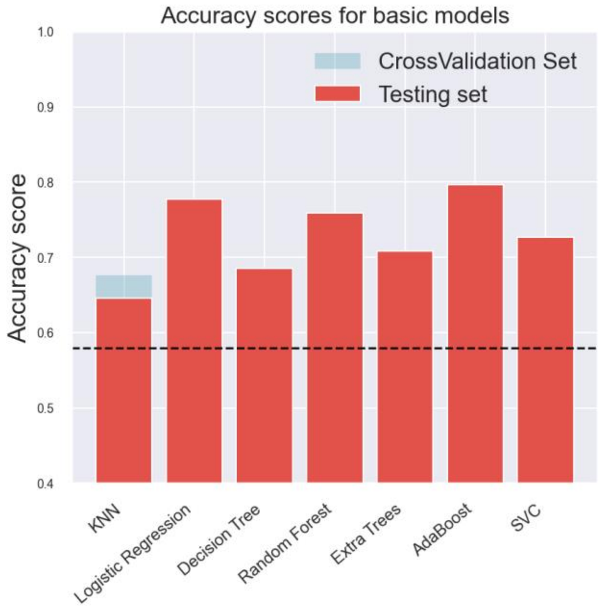
```
Confusion Matrix:
[[192  53]
 [ 65 122]]
```

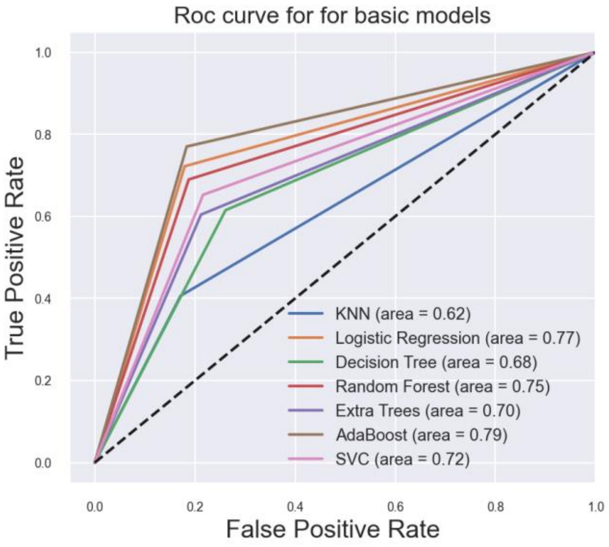
- Classification Report:



```
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.78      0.76       245
           1       0.70      0.65      0.67       187

    accuracy                           0.73       432
   macro avg       0.72      0.72      0.72       432
weighted avg       0.73      0.73      0.73       432
```

+ Compare performance of all listed models:





Observations/ Conclusions:

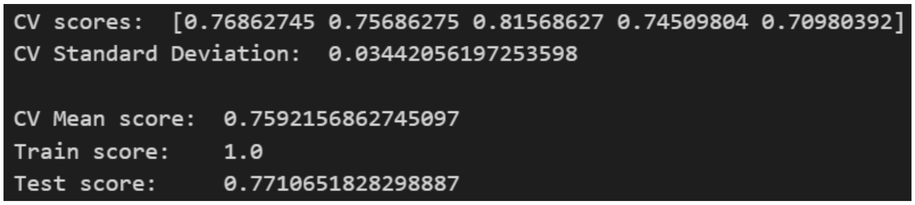
- Logistic regression model preformed the best with accuracy of 0.777 and AUC of 0.777, KNN having the lowest accuracy of 0.64, and AUC of 0.62

b) XGBoost Classifier for clinical attributes only

Final test to see if it is possible to increase the predictive score:

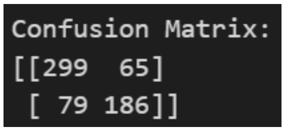
+ Set the parameters range for hyperparameter tunning to identify the best parameters for estimation and create an instance of Grid Search CV model and fit the classification model to find the best estimator.

+ Cross Validation Score:



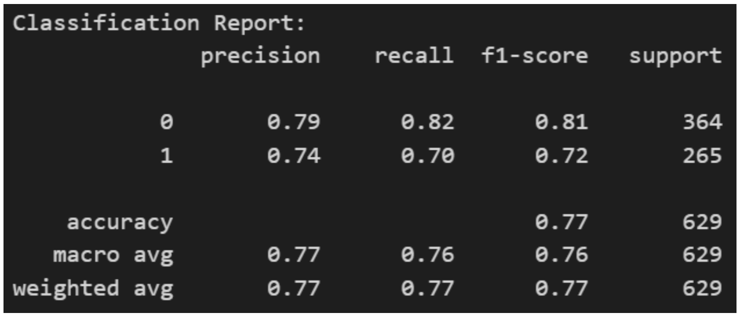
```
CV scores:  [0.76862745 0.75686275 0.81568627 0.74509804 0.70980392]
CV Standard Deviation:  0.03442056197253598

CV Mean score:  0.7592156862745097
Train score:    1.0
Test score:     0.7710651828298887
```

+ Confusion Matrix:



```
Confusion Matrix:
[[299  65]
 [ 79 186]]
```

+ Classification Report:



```
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.82      0.81       364
           1       0.74      0.70      0.72       265

    accuracy                           0.77       629
   macro avg       0.77      0.76      0.76       629
weighted avg       0.77      0.77      0.77       629
```

Observations/ Conclusions:

- XGBoost Classifier preformed very well compared to other traditional basic models with accuracy of 0.779.

Actions:

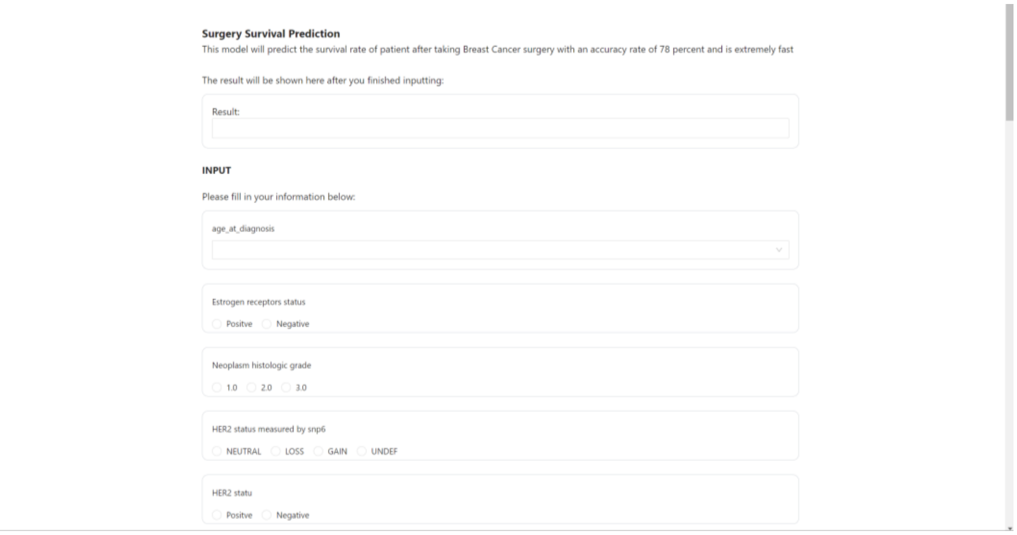
- Save the trained XGBoost Classifier model to deploy later on.

4. Web Deploy:

+ Write a function to output a conclusion for the survival rate of patients.

+ Build and design web layout for the Breast Cancer Surgery Survival Prediction.

IV. Web Interface



Surgery Survival Prediction

This model will predict the survival rate of patient after taking Breast Cancer surgery with an accuracy rate of 78 percent and is extremely fast

The result will be shown here after you finished inputting:

Result:

INPUT

Please fill in your information below:

age_at_diagnosis

Estrogen receptors status

○ Positve ○ Negative

Neoplasm histologic grade

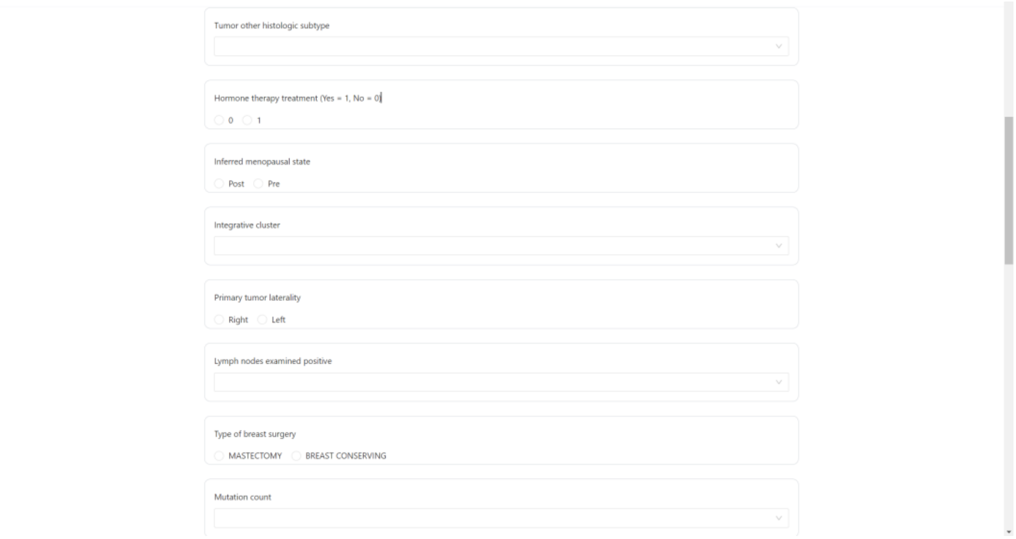
○ 1.0 ○ 2.0 ○ 3.0

HER2 status measured by snp6

○ NEUTRAL ○ LOSS ○ GAIN ○ UNDEF

HER2 statu

○ Positve ○ Negative



Tumor other histologic subtype

Hormone therapy treatment (Yes = 1, No = 0)

○ 0 ○ 1

Inferred menopausal state

○ Post ○ Pre

Integrative cluster

Primary tumor laterality

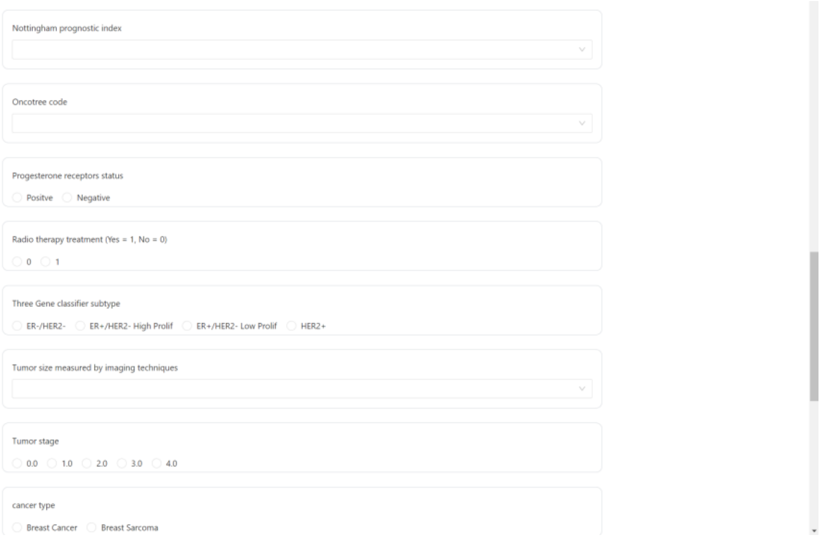
○ Right ○ Left

Lymph nodes examined positive

Type of breast surgery

○ MASTECTOMY ○ BREAST CONSERVING

Mutation count



Nottingham prognostic index

Oncotree code

Progesterone receptors status

○ Positve ○ Negative

Radio therapy treatment (Yes = 1, No = 0)

○ 0 ○ 1

Three Gene classifier subtype

○ ER-/HER2- ○ ER+/HER2- High Prolif ○ ER+/HER2- Low Prolif ○ HER2+

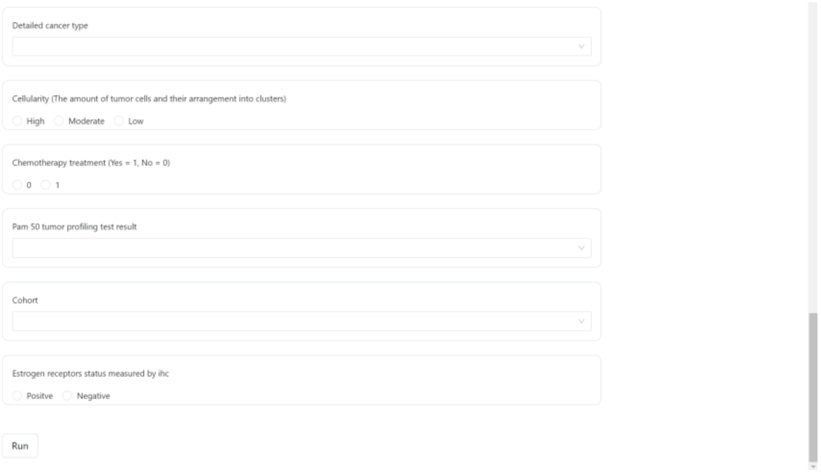
Tumor size measured by imaging techniques

Tumor stage

○ 0.0 ○ 1.0 ○ 2.0 ○ 3.0 ○ 4.0

cancer type

○ Breast Cancer ○ Breast Sarcoma



Detailed cancer type

Cellularity (The amount of tumor cells and their arrangement into clusters)

○ High ○ Moderate ○ Low

Chemotherapy treatment (Yes = 1, No = 0)

○ 0 ○ 1

Pam 50 tumor profiling test result

Cohort

Estrogen receptors status measured by ihc

○ Positve ○ Negative

- Input for users:

+ A dropdown for users to choose their age when diagnosed.

+ Buttons to choose whether their Estrogen receptors status is Positive or Negative.

+ Buttons to choose their Neoplasm histologic grade.

+ Buttons to choose their HER2 status measured by snp6.

+ Buttons to choose whether their HER2 status is Positive or Negative.

+ A dropdown for users to choose their Tumor other histologic subtype.

+ Buttons to choose whether they had Hormone therapy treatment or not.

+ Buttons to choose whether their Inferred menopausal state is post or pre.

+ A dropdown for users to choose their Integrative cluster.

+ Buttons to choose whether their Primary tumor laterality is right or left.

+ A dropdown for users to choose their Lymph nodes examined positive.

+ Buttons to choose their Type of breast surgery.

+ A dropdown for users to choose their Mutation count.

+ A dropdown for users to choose their Nottingham prognostic index.

+ A dropdown for users to choose their Oncotree code.

+ Buttons to choose whether their Progesterone receptors status is Positive or Negative.

+ Buttons to choose whether they had Radio therapy treatment or not.

+ Buttons to choose their Three Gene classifier subtype.

+ A dropdown for users to choose their Tumor size measured by imaging techniques.

+ Buttons to choose their Tumor stage.

+ Buttons to choose their cancer type.

+ A dropdown for users to choose their Detailed cancer type.

+ Buttons to choose their Cellularity (The amount of tumor cells and their arrangement into clusters).

+ Buttons to choose whether they had Chemotherapy treatment or not.

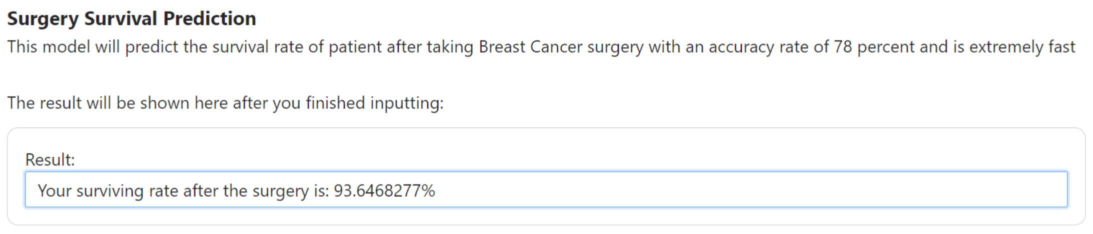
+ A dropdown for users to choose their Pam 50 tumor profiling test result.

+ A dropdown for users to choose their Cohort.

+ Buttons to choose whether their Estrogen receptors status measured by ihc is Positive or Negative.

- Output:

+ A conclusion text showing the prediction result which indicates the survival rate of the patient if they would take the surgery for breast cancer.



Surgery Survival Prediction

This model will predict the survival rate of patient after taking Breast Cancer surgery with an accuracy rate of 78 percent and is extremely fast

The result will be shown here after you finished inputting:

Result:

Your surviving rate after the surgery is: 93.6468277%

V. Conclusion

In conclusion, by offering accurate survival rate predictions and personalized insights, this Breast Cancer Surgery Survival Prediction web service aims to also support patients in making informed decisions about their treatment, reducing anxiety, and fostering trust with their healthcare providers. This innovative tool can help patients approach surgery with a more stable mindset, ultimately improving their overall experience and outcomes.

VI. Presentation

Link to the demo video presenting use cases and step-by-step usage of this Breast Cancer Surgery Survival Prediction:

https://youtu.be/NO7jPyVKsNs