A Machine Learning Based Approach for Prediction of Breast Cancer Patient Prognosis through Clinical Analysis

Adithya Nair[1] and Sharifa Sahai[#]

[1]Branham High School, San Jose, CA, USA
[#]Advisor

ABSTRACT

Patient prognosis for cancer patients is a crucial aspect in the healthcare industry with researchers providing novel insights to doctors evaluating treatment options that have significant implications on patient lifestyle choices. By analyzing correlations of the genetic and clinical attributes for breast cancer patients, previous studies have utilized machine learning algorithms to predict the probability of patient survival based on a five-year timeframe. However, our project focuses on predicting a more specific label (overall survival months), utilizing the extensive dataset of an international breast cancer study, considering only the clinical attributes in scope. The application of the Multivariate Regression and Random Forest models was used to assess the relative importance of each clinical variable. The project results present the Random Forest model to be a better fit, accounting for 44% of the variance in the testing dataset. Further analysis with the expansion of other datasets would help improve the model accuracy.

Introduction

Breast Cancer is the most common form of cancer for women across the world, impacting over 2.3 million lives annually. It remains the second leading cancer-related cause of death for women, closely following skin cancer. In 2020 alone, close to 700,000 women died from breast cancer (Breast Cancer Research Foundation [BCRF], 2021.) Such staggering numbers have prompted collaboration in the healthcare industry between doctors and data scientists to help detect patients with a high risk of breast cancer. Many of these studies have fostered the need for medical datasets with clinical and diagnostic attributes for oncology, including factors such as patient demography (Demo, 2022); (HealthITAnalytics, 2022); (Mucaki et al., 2016.) Combined with greater emphasis on research through scientific mechanisms, machine learning techniques are being increasingly adopted in this industry (Li J;Zhou Z;Dong J;Fu Y;Li Y;Luan Z;Peng X, n.d.) Often open-sourced for public access, these projects can be used to reveal associations between clinical variables and patient response to help those at risk receive appropriate treatment (Howlader; Cronin; Kurian; Andridge; 2018.); (Mucaki et al., 2016); (UCSF Health, 2022.)

The dataset used for the project was taken from the METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) cohort that involved an international study of over 1900 breast tumors between 1977 and 2005 (Alharbi, 2020.) Collected based on primary pathology reports, the METABRIC database contains patient case studies with histology slides available for central review. One of the prominent projects referenced was the collaboration between data scientists and histopathologists, in which ten multivariable logistic regression models were applied to each IntClust group (classification of a large collection of independent samples) for the study of gene expression signatures to predict survival times on patients receiving hormone treatments (HT) and chemotherapy (CT) (Benz, 2008.)

To date, the majority of research conducted has used machine learning algorithms to identify at-risk patients and predict the probability of overall survival, often taking death from cancer within a five-year timeframe as their label (Humphries & Gill, 2003); (Li J;Zhou Z;Dong J;Fu Y;Li Y;Luan Z;Peng X, n.d.) (Mucaki et al., 2016.) However, our project seeks to determine a more interpretable prognosis indicator: patient survival months, and the feature importance of the variables with their correlation coefficient. The implementation of two supervised learning algorithms - Multivariate Linear Regression and Random Forest - was used to predict overall patient survival months using selected variables from the METABRIC database. The project findings from such study and all future works can be instrumental in helping doctors and healthcare providers understand and make adjustments to determine the best course of treatment.

Methods

In this segment, the dataset used in the project study and the justification behind each chosen variable is described. The data preparation steps, namely the cleaning and the feature engineering applied prior to the model implementation, are also noted.

METABRIC Database

The METABRIC database contains clinical and genetic attributes of close to 2000 breast cancer patients. For our project, the scope was limited to only considering clinical variables (28 from 31 clinical attributes in the database). These 28 attributes served as our model input for determining patient survival time. For the model input parameters, the analysis of the clinical variables and 8 key clinical variables were selected as part of the hypothesis as the significant variables.

Analysis of Clinical Variables

In consultation with specialized breast cancer researchers and referencing related studies in the field, this section contains our hypothesis of the 8 key clinical variables as the most influential attributes for our model in order of relevance.

Age at Diagnosis

Women over the age of 50 account for approximately 80% of all breast cancer cases (UCSF Health, 2022.) In addition, these older patients are often treated with less intensive treatment after adjustment for a multitude of factors: medication allergies/restrictions, impacts on lifestyle changes, weaker immune systems, psychological motivational factors, etc. Therefore, many studies have reported age as one of the most significant factors in prognosis, often based on its direct relation to whether cancer has metastasized (UCSF Health, 2022); (Howlader; Cronin; Kurian; Andridge; 2018.)

Tumor Size

The TNM staging system is a widely used classification system for determining tumor stage through the measurement of tumor size, the spread of cells to lymph nodes, and metastasis of cancer to other areas. Based on a study conducted by ACS (American Cancer Society), the 5-year survival rate for patients with stage 4 breast cancer is 28%, drastically lower in comparison to the average mean for all stages at 90% (Koehrsen, 2018.) In the dataset, the "tumor_stage" attribute had over 25% of the attributes marked as missing values, so our model focused on the related "tumor_size" attribute.

Mutation Count

Inherited mutations in BRCA1 or BRCA2 genes can often lead to abnormal cell growth. Based on a PubMed Abstract study, approximately 55 to 72% of women with BRCA1 variant and 45 to 69% of women with BRCA2 variant will develop breast cancer between the ages of 60 and 70 (Madell, 2021.) However, there is also an association between the inheritance of particular genes and the earlier development of breast cancer. So, crucial preventative measures for those with BRCA1 and BRCA2 variants include earlier screenings for their children to detect possible inheritance of malign tumors.

Cellularity

In the database, cellularity, the measured proportion of tumor and normal cells, is denoted by 3 intervals (low, medium, high). While there was no observed direct relationship to overall survival months, there is a clear positive correlation with other corresponding factors such as tumor size and stage. As expected, a significant decrease in cellularity of the tumor was noted in patients undergoing chemotherapy treatment. Further analysis by pathological review can serve to complement quantitative image analysis of HER2 count.

Lymph Nodes Examined Positive

Based on a Seer dataset from 2012 to 2018, the five-year survival rate for a regional spread of cancer cells is 86%, meaning that cancer has spread to nearby lymph nodes (Petrucelli; Daly; Pal; 2022.) For many of these patients, radiation therapy proves as a viable solution following surgical treatment. A direct relationship between lymph nodes present and diagnosis of stage of breast cancer was observed in the database. However, further analysis needs to be conducted on the relationship between lymph node dissection and the cancer diagnosis stage.

Hormone Therapy

Hormone therapy is only applied where the presence of receptors for estrogen or progesterone hormones is seen. Based on the Breast Cancer Index, an analysis of a sample of blood cells would indicate the best course of action for hormone therapy treatment options. Past studies have noted the following associated benefits - reducing chances of cancer relapse, preventing the growth of malignant tumors, helping aid preparation for surgical treatment, etc... (Cancer.Net, 2022); (Humphries & Gill, 2003.) Other possible options include the usage of LHRH (Luteinizing Hormone-Releasing Hormone) drugs in conjunction with other hormone drugs such as FDA-approved tamoxifen in treatment plans.

HER2 Status

Based on a study published in the NIH, the four-year survival rate for women with HR+/HER2+is estimated to be 90.3% as compared to a survival rate for patients with HR - / HER2- at 77% (Howlader; Cronin; Kurian; Andridge; 2018.) Some therapeutic options including chemotherapy and hormone therapy can help assess which patients are susceptible to a higher risk of relapse. There is also a linked association between younger age and the likelihood of HER2+ cancer.

Tumor Histologic Subtype

The histological differences of breast carcinomas highlight different prognosis implications. The classification of molecular subtypes has been made possible by gene expression profiling. One popular treatment option for certain responsive subtypes includes endocrine therapy. Further studies on specific histologic subtypes are necessary to identify associations with other clinical variables.

Feature Engineering

A review of the data revealed the missing values and the need for feature engineering across categorical classification and clustering. Table below outlines the various feature engineering steps for the 28 chosen attributes out of 31 in the database.

Table 1. Feature Engineering Approach with Range of Values for 31 Clinical Variables

#	Attribute	Existing Value(s)	Feature Eng. Approach
1	patient_id	Unique Numeric Values	No modification needed
2	Age_at_Diagnosis	Numeric Values (21.83 to 96.23)	Binning (8 Bins)
3	type_of_breast_surgery	2 Categorical Values	Replaced by 0 & 1
4	cancer_type	1 Categorical Value (except 1 record)	Exclude the Attribute
5	cancer_type_detailed	5 Categorical Values	Classification (0 to 4)
6	Cellularity	3 Categorical Values (High to Low)	Classification (0 to 2)
7	chemotherapy	2 Numeric Variables	Retained AS-IS
8	pam50_+_claudin-low_subtype	6 Categorical Values	Classification (0 to 6)
9	Cohort	5 Numeric Values	Retained AS-IS
10	er_status_measured_by_ihc	2 Categorical Values	Classification (0 to 1)
11	er_status	2 Categorical Values	Classification (0 to 1)
12	neoplasm_histologic_grade	3 Categorical Values (1 to 3)	Retained AS-IS
13	her2_status_measured_by_snp6	4 Categorical Values	Classification (0 to 3)
14	her2_status	2 Categorical Values	Classification (0 to 1)
15	tumor_other_histologic_subtype	8 Categorical Values	Classification (0 to 7)
16	hormone_therapy	2 Numeric Values	Classification (0 to 1)
17	inferred_menopausal_state	2 Categorical Values	Classification (0 to 1)

18	integrative_cluster	10 Categorical Values & 4 ER+ 4ER-	No Change to most of the classification except 4 ER+ & 4 ER-
19	primary_tumor_laterality	2 Categorical Values	Classification (0 to 1)
20	lymph_nodes_examined_positive	Numeric Values (0 to 45)	Retained AS-IS
21	mutation_count	Numerical Values (1 to 80)	Retained AS-IS
22	nottingham_prognostic_index	Range observed from 1 to 6.36	Classification (0 to 3) Patients were grouped into four categories according to the NPI score: I (excellent) ≤2.4; II (good) >2.4 but ≤3.4; III (moderate) >3.4 but ≤5.4; and IV (poor) >5.4.
23	oncotree_code	6 Categorical Values	Classification (0 to 5)
24	overall_survival_months	Numeric Values (0 to 355.2)	Classification (1 to 10) based on mean frequency and distribution of data: <12 months, Each year till 60, then 3-4 year ranges till 240 months and >240 months
25	overall_survival	2 Numeric Values	Retained AS-IS
26	pr_status	2 Categorical Values	Classification (0 to 1)
27	3-gene_classifier_subtype	4 Categorical Values	Exclude the Attribute
28	radio_therapy	2 Numeric Values	Retained AS-IS
29	tumor_size	Numeric Values (1 to 182)	Retained AS-IS
30	tumor_stage	Numeric Values (0 to 5)	Exclude the Attribute
31	death_from_cancer	3 Categorical Values	Classification (0 to 2)

After the numerical classification of the categorical attributes, the second part of the data cleaning was to handle missing values with suitable dropout methods or imputation techniques. There were about 400 rows of data that were deleted as the imputation methods would have introduced noise/biases in the data set.

Finally, within the dataset, rows where the attribute "death from other causes" had values (n<5 years) were removed as this data was ambiguous to interpret and was not showing any correlation with other variables.

Model Construction

As part of the supervised machine learning regression models, 2 primary algorithms were chosen - Linear (Multivariate Regression) & Non-Linear (Random Forest). The implementation steps for each model construction are described below.

Multivariate Regression

The first algorithm implemented was the multivariate regression model. This model is used to display the relationship between 2 or more independent variables (model input) and a dependent variable (model output). Primarily used in larger datasets due to its mathematical complexity, this model is used in assessing the relative importance of each variable for predicting a given outcome. This algorithm, similar to a linear regression model, is based on multiple key assumptions: the data must be continuous in nature, the residual error remains consistent throughout the dataset, the spread is normally distributed, and the input variables have some correlation with the output variables.

Equation 1: Multivariate Regression Algorithm:

$$\mathrm{Y_{ik}} = \mathrm{b_{0k}} + \sum_{\mathrm{j=1}}^{\mathrm{p}} \mathrm{b_{jk}}\,\mathrm{x_{ij}} + \mathrm{e_{jk}}$$

"Y_{ik} defines the predicted variable for the i^{th} observation, b_{0k} is the intercept for the k^{th} response, b_{jk} is the j^{th} predictor variable slope for the k^{th} response, x_{ij} is the j^{th} predictor variable for the i^{th} observation, and e_{jk} is the error vector" (Jain, 2022.)The model's individual coefficients can be determined using the Ordinary Least Squares (OLS) equation.

Equation 2: Ordinary Least Squares Equation:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}\,)(y_i - \overline{y}\,)}{\sum_{i=1}^{n}(x_i - \overline{x}\,)^2}$$

B_1 is the model intercept, x_i represents the explanatory variable, and y_1 represents the predicted variable (Alto, 2019.) By calculating the minimal sum of squared residuals (SSR), the equation helps us find the regression coefficients of the model.

Finally, the R-Squared coefficient (value from 0 to 1) equation, as shown below, was used to measure how close the line of best fit is to the original data points.

Equation 3: R-Squared Coefficient Equation

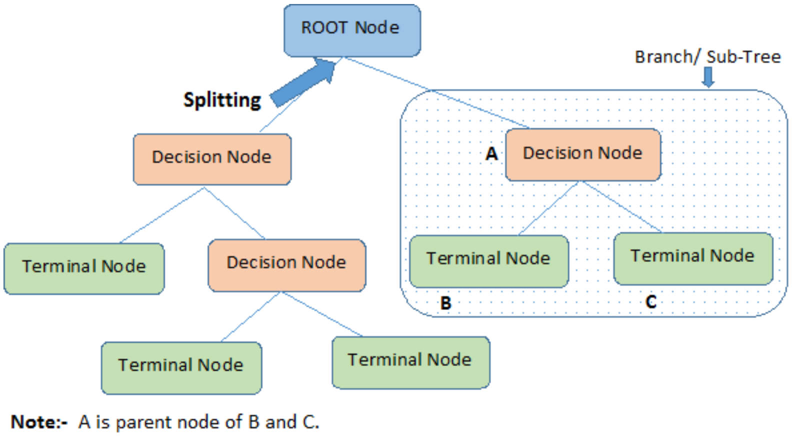
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$$

While high R-Squared coefficients reflect greater correlations between the explanatory and measured variables, it is important to note that low R-Squared coefficients can also reflect key insights into the database as well. This coefficient represents our final model accuracy score.

Random Forest

The second algorithm chosen was the Random Forest (RF) model. Random Forest is an ensemble learning algorithm that is frequently used in classification or regression problems. Due to its effectiveness in handling complex datasets through the formation of numerous decision trees, the random forest model is used in the healthcare industry for its ability to identify associations between patients with similar characteristics and group them accordingly (Demo, 2022.) Unlike Multivariate Regression models, Random Forest can handle skewed data points and can assess many explanatory variables. However, one of the drawbacks is that models with hundreds of decision trees are often computationally inefficient, leading to longer runtimes and difficulties in training the dataset.

One of the key aspects of the Random Forest algorithm is choosing the right hyperparameters to optimize model performance. Random Forest, which is based on the formation of multiple decision trees as seen in Figure 1, has 5 parameters that need to be determined: ‘n_estimators’, ‘min_samples_split’, ‘max_samples_split’, ‘max_features’, and ‘max_depth’ (Kurama, 2021.) An explanation of each hyperparameter alongside its significance in model performance is described in the next paragraph.



Note:- A is parent node of B and C.

Figure 1. Different Branches in Decision Tree Diagram

N_estimators, or the number of estimators, refer to the number of decision trees in the model. The higher the number of estimators, the easier it is for the model to evaluate trends in the data. However, a high number of estimators significantly increases the runtime for training processes.

Min_samples_split and max_samples_split define the minimum or the maximum number of samples necessary for an internal node to split. Denoted as either a percentage of samples or as an integer, the optimal value for this parameter would mean the model is able to accurately analyze correlations in the data without issues of model underfitting or overfitting.

Max_features represents the number of features to evaluate when searching for the optimal split. In regression problems, square root or sqrt (n_features) is frequently used as a measure.

Max_depth refers to the maximum depth of each tree in the model. Similar to n_estimators and min/max_samples_split, the optimal number is necessary to help avoid model under/overfitting.

Hyperparameter Tuning

To determine the optimal split for each parameter, hyperparameter tuning was implemented. In general, this approach works better based on experimental data through trial and error, rather than mathematical calculations (Koehrsen, 2018.) In addition, K-Fold Cross Validation was used to split the training data into further subsets to help provide further insights into model performance.

Performance Metrics

The calculation of the mean absolute/mean square error, which compares the accuracy of the line of the best fit to the original data points, as well as the R2 coefficient, was used to assess model accuracy.

Results

Our results highlight that the Random Forest model was able to better assess patient prognosis with the following results:

Table 2: Three Calculations used to measure Model Accuracy

Mean Absolute Error	1.34
Mean Squared Error	3.13
Final R2 Coefficient	0.44

This table signifies that our model's input variables were able to explain 44% of the variance in the predicted values for patient prognosis time. Further, a plot of the influence of n_estimators (hyper parameter for number of decision trees) is depicted below.

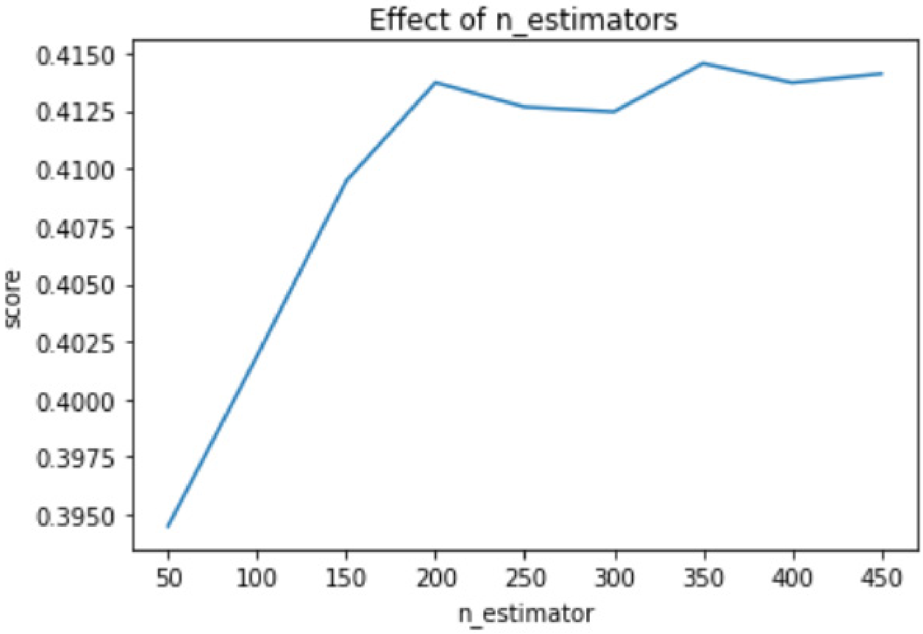


Figure 2: Plot of Influence of N_Estimators on Accuracy Score

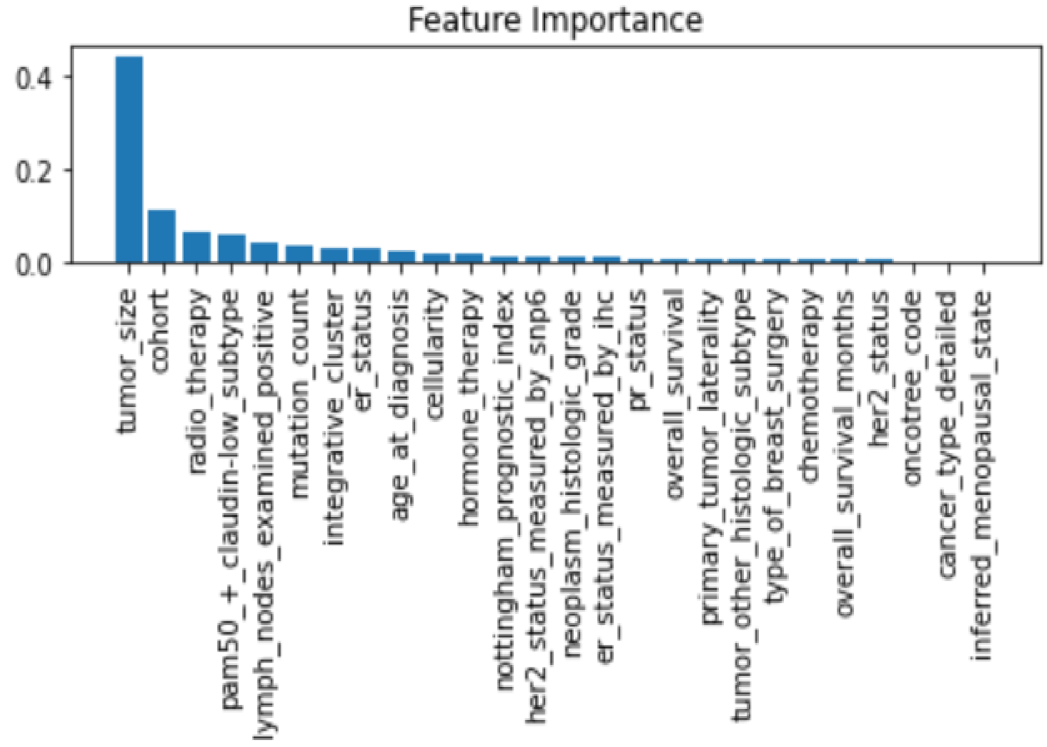


Figure 3: Feature importance chart for the 26 clinical variables excluding "death_from_cancer" and "overall_survival" attributes

Discussion

In general, it was observed that our model was able to predict a more accurate prognosis for patient demographics within the following 3 clinical variables: tumor_size, lymph_nodes_present, and mutation count, in accordance with our hypothesis. Further analysis of trends that explain the correlation between these input variables and "overall_survival_months" attribute needs to be conducted for deeper insight into the reasoning behind these relations.

Contrary to our clinical analysis, our model observed limited associations to patient prognosis for the following 3 clinical variables: age_at_diagnosis, cellularity, and hormone_therapy. For "age_at_diagnosis" attribute, this can partially be explained by the dataset curated as the distribution was limited in its scope of including enough patients (n>70 years). For the "cellularity" and "hormone_therapy" attributes, more studies need to be conducted to understand why these factors were not as influential in predicting the model output.

Limitations

Within the METABRIC dataset, there is a bias towards higher age population, which skews the distribution of IntClust prognostic and pathological variables. In addition, the data also had significant missing rows on two of the clinical attributes "3-gene_classifier_subtype" and "tumor_stage". So, these two attributes had to be dropped as the imputation methods would have added noise in our dataset. Further, only selected clinical variables within the METABRIC database were analyzed to predict prognosis. Our study could potentially be expanded to analyze more clinical variables as well as the addition of genetic traits.

Future Works

In addition to the caveats listed above, the project could broaden its scope to analyze genomic aberrations and gene expression-defined subtypes, which would be useful to make more accurate predictions. This would involve including more datasets to include histopathology review.

The implementation of the Cox9 Regression analysis could also be explored to help assess the relative importance of each variable given its ability to effectively handle additional input variables such as genetic attributes. The key attributes from this analysis could then be incorporated into a more advanced model such as XGBoost Regression, and the results could be compared to see if there was scope for improving model performance.

Conclusion

The purpose of this study was to utilize machine learning models to understand correlations between clinical variables, their feature importance, and patient survival prognosis time. Our study results provided useful insights into the significant variables and were in conformance with the initial hypothesis assumed about the eight clinically significant variables that can be used in future studies to help determine a strong correlation to patient survival tenure and treatment options. However, the model could not provide a high degree of confidence in predicting the survival time tenures, as was expected given the limited range of data values in the data set. But the model with future enhancements through other tuning parameters (regularization, drop-out features, etc.), combined with augmented datasets on genetic attributes could lead to better insights and a means to propel further model studies in this domain. Nevertheless, our study is a considerable advancement in the way clinical and genetic attributes can be studied further to gain insights and help the research and healthcare providers in enhancing the patient treatment options and quality of life significantly.

Acknowledgments

The first author would like to thank the research mentor for all their help and guidance without which the project could not have been completed. The author would also like to express their gratitude for CBIO Portal's curated dataset of the METABRIC database. The work done in the project is solely representative of the author's knowledge and is not endorsed by either the CBIO company or METABRIC study.

References

Alto, V. (2019, August 17). *Understanding the OLS method for simple linear regression*. Medium. Retrieved June 29, 2022, from https://towardsdatascience.com/understanding-the-ols-method-for-simple-linear-regression-e0a4e8f692cc

Benz, C. C. (2008, April). *Impact of aging on the biology of breast cancer*. Critical reviews inoncology/hematology. Retrieved July 27, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2626623/

BRCA1- and BRCA2 -associated Hereditary Breast and ... - NCBI bookshelf. (n.d.). Retrieved August 25, 2022, from https://www.ncbi.nlm.nih.gov/books/NBK1247/

Breast cancer - statistics. Cancer.Net. (2022, May 24). Retrieved June 29, 2022, from https://www.cancer.net/cancer-types/breast-cancer/statistics

Breast cancer gene expression profiles (METABRIC). Kaggle. (n.d.). Retrieved June 8, 2022, from https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric

Breast cancer statistics and resources: Breast Cancer Research Foundation: BCRF. Breast Cancer Research Foundation. (2021, August 31). Retrieved July 21, 2022, from https://www.bcrf.org/breast-cancer-statistics-and-resources/

Demo. (2022, August 9). *Working together, data scientists and cancer researchers can transform cancer treatment*. Susan G. Komen®. Retrieved July 24, 2022, from https://blog.komen.org/blog/data-scientists-and-cancer-researchers/

HealthITAnalytics. (2022, January 19). *Machine learning supports breast cancer diagnosis predictions*. HealthITAnalytics. Retrieved July 2, 2022, from https://healthitanalytics.com/news/machine-learning-supports-breast-cancer-diagnosis-predictions

Hormone therapy for breast cancer fact sheet. National Cancer Institute. (n.d.). Retrieved July 23, 2022, from https://www.cancer.gov/types/breast/breast-hormone-therapy-fact-sheet

Humphries, K. H., & Gill, S. (2003, April 15). *Risks and benefits of hormone replacement therapy: The evidence speaks*. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. Retrieved June 19, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC152685/

Jain, R. (2022, April 20). *Application of multivariate regression analysis*. Knowledge Tank. Retrieved June 29, 2022, from https://www.projectguru.in/application-of-multivariate-regression-analysis/

Koehrsen, W. (2018, January 10). *Hyperparameter tuning the random forest in python*. Medium. Retrieved June 18, 2022, from https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Kurama, V. (2021, April 9). *A complete guide to decision trees*. Paperspace Blog. Retrieved August 27, 2022, from https://blog.paperspace.com/decision-trees/

Li J;Zhou Z;Dong J;Fu Y;Li Y;Luan Z;Peng X; (n.d.). *Predicting breast cancer 5-year survival using Machine Learning: A Systematic Review*. PloS one. Retrieved August 23, 2022, from https://pubmed.ncbi.nlm.nih.gov/33861809/

Madell, R. (2021, June 23). *Metastatic breast cancer prognosis*. Healthline. Retrieved July 19, 2022, from https://www.healthline.com/health/breast-cancer/metastatic-prognosis

Mucaki, E. J., Baranova, K., Pham, H. Q., Rezaeian, I., Angelov, D., Ngom, A., Rueda, L., & Rogan, P. K. (2016, August 31). *Predicting outcomes of hormone and chemotherapy in the molecular taxonomy of breast cancer international consortium (METABRIC) study by biochemically-inspired machine learning*. F1000Research. Retrieved August 4, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5461908/

R;, H. N. C. K. A. K. A. W. A. (n.d.). *Differences in breast cancer survival by molecular subtypes in the United States*. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. Retrieved August 8, 2022, from https://pubmed.ncbi.nlm.nih.gov/29593010/

UCSF Health. (2022, June 24). *Breast cancer risk factors*. ucsfhealth.org. Retrieved July 16, 2022, from https://www.ucsfhealth.org/education/breast-cancer-risk-factors