**About the Dataset –**

The **METABRIC project** focuses on understanding the molecular and genetic profiles of breast cancer, which helps in categorizing breast cancer into more precise subtypes. This categorization is crucial for developing more effective, personalized treatment strategies. The data typically includes information like gene expression profiles, clinical features, patient outcomes, and possibly genomic alterations.

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) is a Canada-UK project that contains a database of breast cancer data. Dataset are available on: https://www.cbioportal.org/datasets

Here are the first 26 columns from the **METABRIC RNA dataset**, along with a brief explanation of each:

1. **patient\_id**: Identifier for each patient.

2. **age\_at\_diagnosis**: Age of the patient at the time of cancer diagnosis.

3. **type\_of\_breast\_surgery**: Type of breast surgery undergone (Mastectomy, Breast Conserving, etc.).

4. **cancer\_type**: General type of cancer diagnosed.

5. **cancer\_type\_detailed**: More detailed description of the breast cancer type.

6. **cellularity:** Measure of tumor cell density.

7. **chemotherapy**: Indicator of whether the patient received chemotherapy (0 = No, 1 = Yes).

8. **pam50\_+\_claudin-low\_subtype**: Classification of breast cancer based on gene expression profile.

9. **cohort**: Group or cohort the patient belongs to within the study.

10. **er\_status\_measured\_by\_ihc**: Estrogen receptor status measured by immunohistochemistry.

11. **er\_status**: Status of estrogen receptors in cancer cells (Positive/Negative).

12**. neoplasm\_histologic\_grade**: Histologic grade of the neoplasm.

13. **her2\_status\_measured\_by\_snp6**: HER2 status measured by SNP6 (a type of genomic analysis).

14. **her2\_status**: Status of HER2, a protein that can affect cancer cell growth.

15. **tumor\_other\_histologic\_subtype**: Other histologic subtypes of the tumor.

16. **hormone\_therapy:** Indicator of whether the patient received hormone therapy (0 = No, 1 = Yes).

17. **inferred\_menopausal\_state:** Menopausal state of the patient (Pre/Post).

18. **integrative\_cluster:** Classification based on integrated analysis of genomic and transcriptomic data.

19. **primary\_tumor\_laterality**: Laterality of the primary tumor (Left/Right).

20. **lymph\_nodes\_examined\_positive**: Number of lymph nodes examined that were found to be positive for cancer.

21. **mutation\_count**: Count of mutations found in the tumor.

22. **nottingham\_prognostic\_index**: A prognostic index used for breast cancer.

23. **oncotree\_code**: Oncotree codes representing different types of tumors.

24. **overall\_survival\_months**: Duration of overall survival in months since diagnosis.

25. **overall\_survival**: Indicator of overall survival (0 = No, 1 = Yes).

26. **pr\_status:** Status of progesterone receptors in cancer cells (Positive/Negative).

**Data Preprocessing –**

1. Handling missing value
2. Scale the numerical value with Z score
3. Convert categorical values to numerical values with one hot encoding

**Data Exploratory and Visualization –**

After visualizing, we find out the dataset was divided into three parts and each part were trained particularly to find out the importance of each part. All the visualization is shown in the [**code**](https://github.com/kazimahathir73/Breast-Cancer-Survival-Analysis-and-Prediction/blob/main/Breast%20Cancer%20Survival%20Analysis.ipynb).

1. Detecting outliner with IQR and visualization
2. Reduce the effect of outliner in model, we used cube transform

**Model and Training –**

1. **Support Vector Machine (SVM**): Used for its effectiveness in classification tasks, particularly in high-dimensional spaces.

2. **Logistic Regression**: Chosen for its simplicity and efficiency in binary classification problems.

3. **K-Nearest Neighbors (KNN):** Utilized for its intuitiveness and ability to classify data based on feature similarity.

4. **Random Forest:** Selected for its robustness and accuracy, especially in handling large datasets with multiple features.

**Result and Conclusion –**

As the data was trained three times and the charts shows the F1 score, Accuracy and ROC\_AUC:

**For first\_features only:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | KNN | Random Forest | SVM | Logistic Regression |
| Accuracy | 0.7192 | 0.7612 | 0.7585 | 0.7402 |
| F1 score | 0.6515 | 0.6873 | 0.6892 | 0.6689 |
| ROC\_AUC | 0.7582 | 0.8409 | 0.8382 | 0.8157 |

**For first\_features + second\_features:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | KNN | Random Forest | SVM | Logistic Regression |
| Accuracy | 0.6299 | 0.7218 | 0.7690 | 0.7060 |
| F1 score | 0.5087 | 0.5985 | 0.6879 | 0.6387 |
| ROC\_AUC | 0.6408 | 0.7547 | 0.8028 | 0.7165 |

**For first\_features + second\_features + third\_features (whole dataset):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | KNN | Random Forest | SVM | Logistic Regression |
| Accuracy | 0.6404 | 0.6955 | 0.7769 | 0.6772 |
| F1 score | 0.5324 | 0.5323 | 0.6996 | 0.6019 |
| ROC\_AUC | 0.6640 | 0.7145 | 0.8094 | 0.7254 |