BREAST CANCER PROGNOSIS: WHAT FACTORS CONTRIBUTE TO PROGRESSION?

Problem Statement:

Predicting breast cancer stage, and outcomes utilizing the prevalence of 4 breast cancer-related proteins, and hormone receptor status.

Central Questions

1. What factors (protein levels, receptor status, etc.) influence survival of the individual?
2. Can protein levels and receptor status be used as an accurate indicator of the stage of breast cancer?
3. What demographic data trends are seen within the dataset of breast cancer patients?

Goals:

Finding significant correlations between protein or receptor status with the stage of cancer, and prognosis. Analyzing the mortality rates associated with protein levels, receptor status, type of cancer, and type of surgery. A slidedeck and a report will be submitted regarding the findings gathered during the study. Additionally, a program will be created in which users can input data regarding their protein levels, and receptor status to determine prognosis (if proven to be a significant measure).

Datasets

Link: https://www.kaggle.com/datasets/kellistephenson/increasing-breast-cancer-awareness

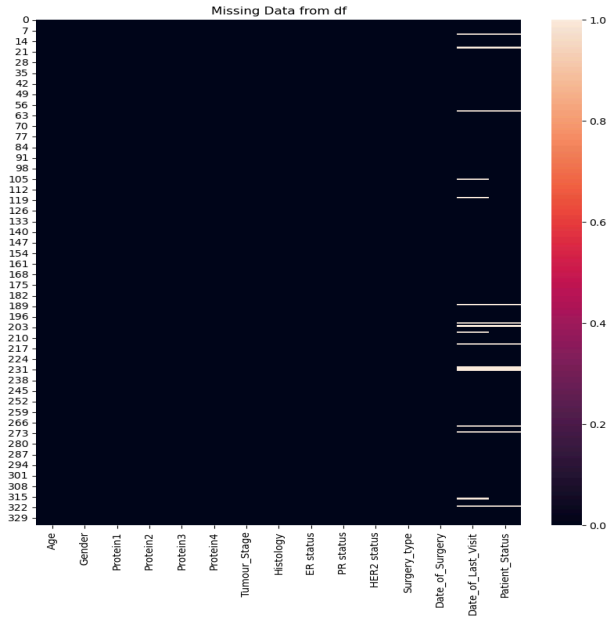
Description: The Data source provides information on patients with breast cancer including, the levels of four proteins, their sex, their age, their receptor status, the cancer stage, the type of cancer, scheduled surgeries, their mortality status, and their last visit to the clinic.

Data Cleaning and Data Wrangling

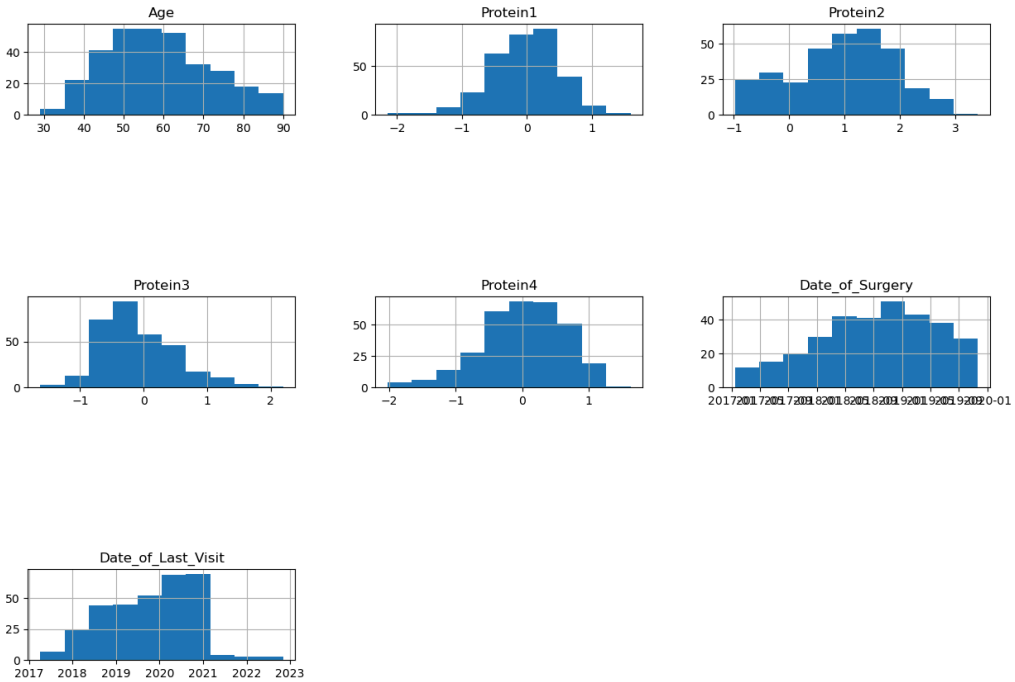
Initially, the dataframe (df) had **15 columns and 334 entries**. The columns include the age, gender, protein1,protein2,protein3,protein4,tumour_stage,histology,ER status,PR status, HER2 Status, Surgery_type, Date_of_surgery, Date_of_Last_Visit, Pateint_Status.

The sub plot indicates that in a list of instances when Date_of_Last_Visit is missing, so is the Patient_Status. It is hard to determine prognosis, if patient status and date of last

visit is missing, so I will drop those columns from the dataset. Additionally, if patient status is dead, and the last clinic visit date for the patient is missing, it is safe to assume that last clinic visit date is surgery date.



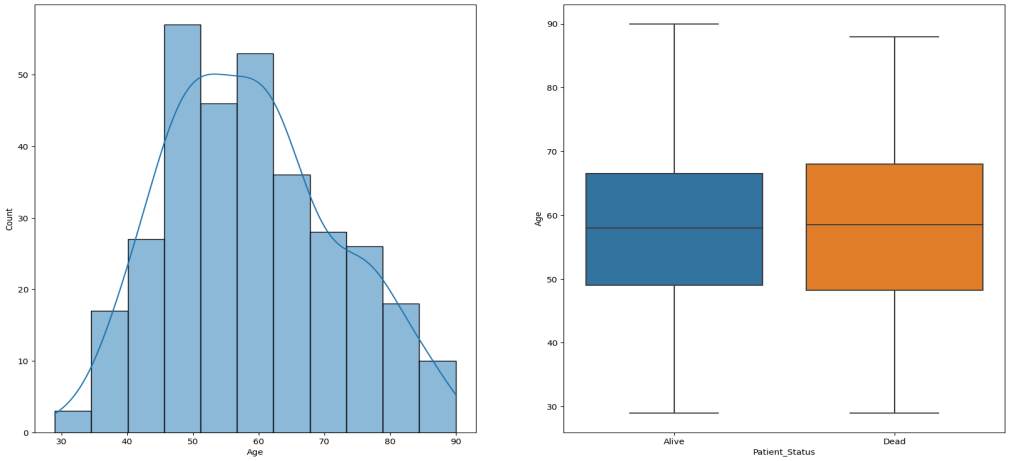
The distributions of some of the numerical variables displayed some outliers such as the date of last visit, with some of them being past the current date. I chose to assume that the date of their surgery will be the date of their last visit to the clinic.After adjusting those outliers the distributions appear to me normal as indicated by the plot below.



With the removal of rows with missing data, and the imputation of outliers, the cleaned dataset had a total of 321 entries, showing that 13 entries were removed during the data cleaning process.

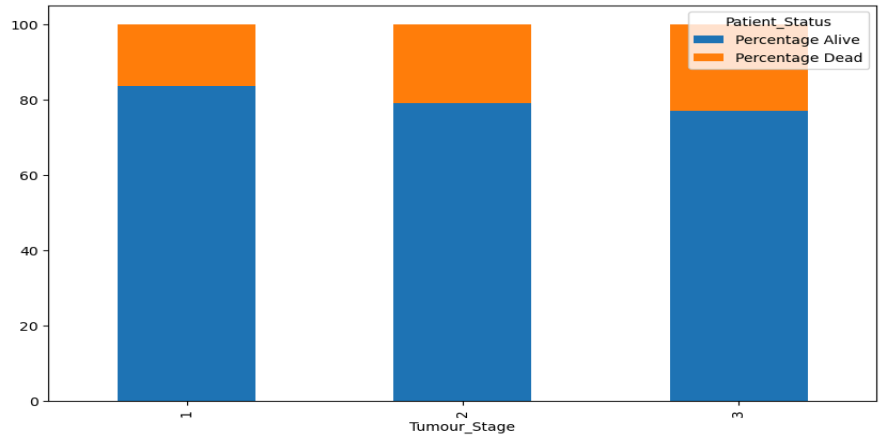
Exploratory Data Analysis

Age vs. Patient Status



From comparing the boxplot and histogram, age does not seem to be highly correlated with patient status. The boxplots for both alive, and dead status are similar with minor differences. Further examination and exploration will be needed to confirm this lack of relationship

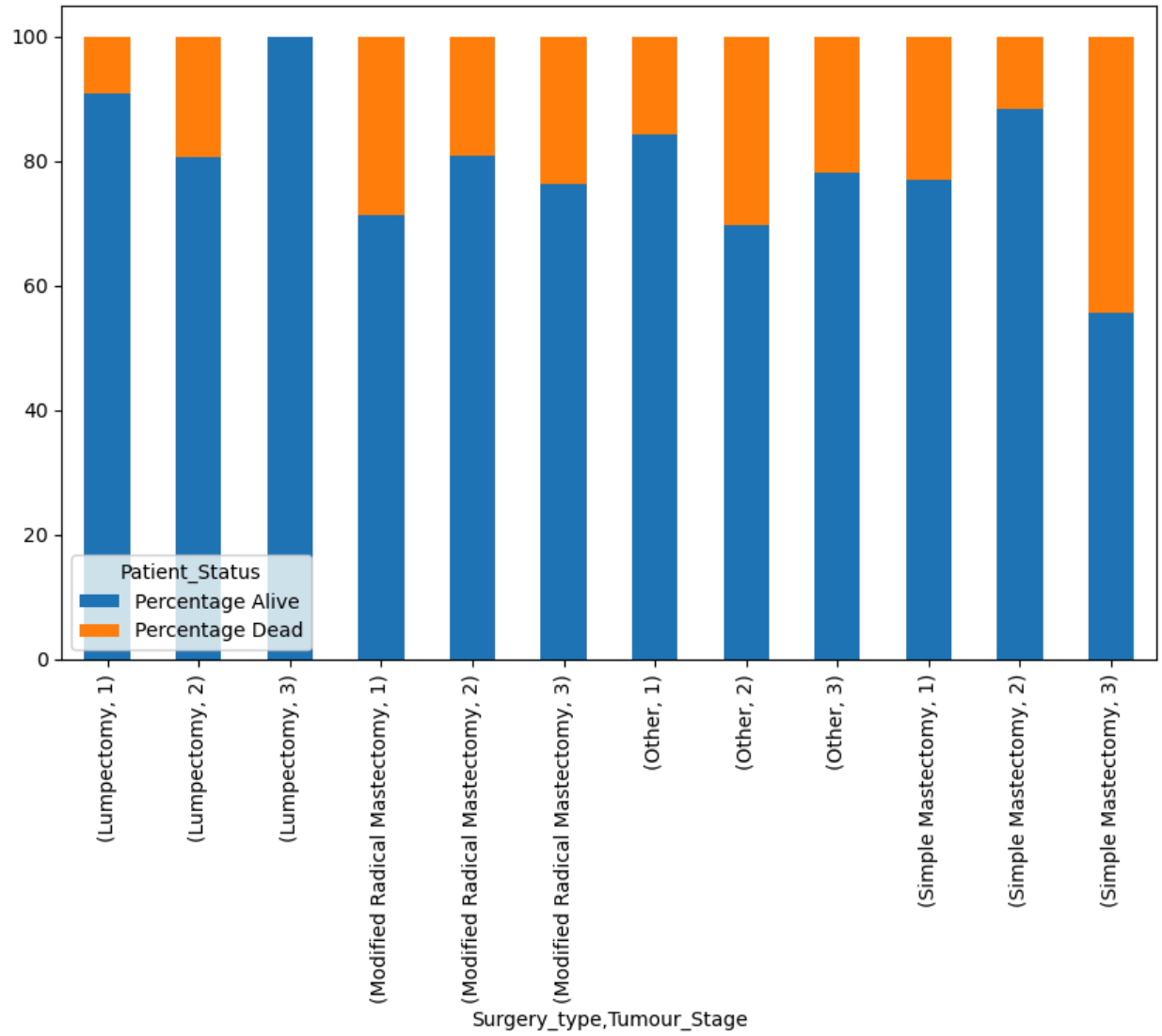
Tumour Stage vs. Patient Status



A clear relationship can be seen with patient status and tumor stage. As the tumor stage increases the prognosis for the patient decreases, indicated by the increase in the percentage dead. I will

explore this correlation further in the modeling stage. Tumor Stage could be used as an indicator for both mortality of the patient and the prognosis of the condition.

Type of Surgery vs. Patient Status

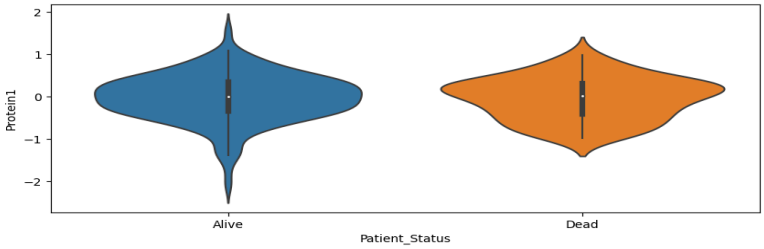


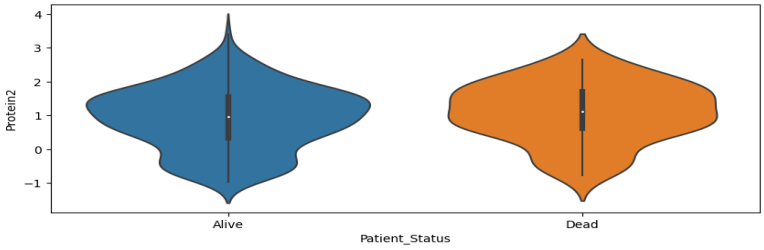
The graph above shows the percentage of patients alive and dead for each kind of surgical procedure, for each stage of cancer. The types of procedures include: Lumpectomy, modified radical mastectomy,other, and simple mastectomy. The aspect of stage plays an interesting role in this relationship, because within some procedures there's an increase in mortality as tumor stage increases, while others show a more sporadic pattern. There may be various other factors that go into determining the success of a surgery, including age, receptor status, and etc. Due to the sporadic nature of the patient statuses, type of surgery may not play an important role in indicating prognosis.

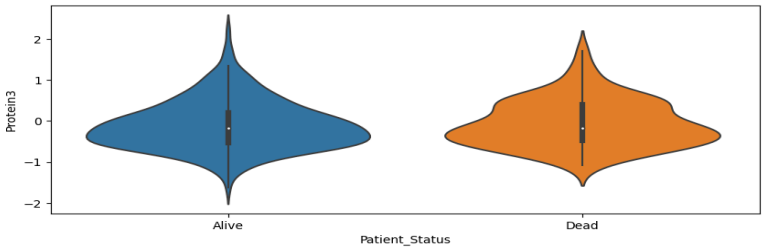
Protein Levels

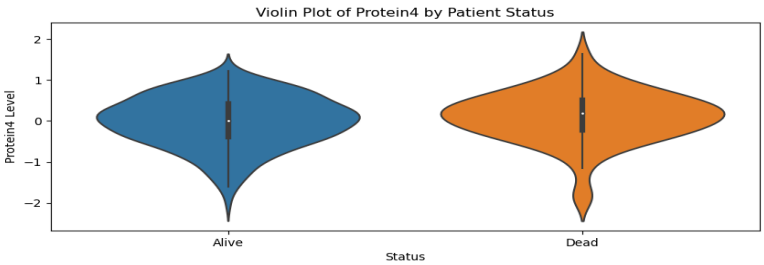
One of the major aspects of my problem statement included the exploration of protein levels among the mortality of patients, their diagnosis types, and tumor stages. I wanted to see if protein levels can be indicative of a mutating cancer, or indicative of specific cancer types. During the eda stage, I pinned protein levels against other factors to see if any significant correlations can be noticed. The dataset included levels of 4 important proteins that are known to be associated with breast cancer.

Protein Levels vs. Patient Status









The violin plots above are separated by status of patient (alive/dead) and the levels of protein in their system. I used a violin plot since the shape of the two plots can be easily compared to discern any

major difference. From an initial glance, there are no significant differences among protein levels for any of the proteins, between patients who are still alive and those that have passed. These results were supported by the correlation anova tests, which did not output p scores that indicated a significant correlation.

Model Implementation

During the model implementation, after encoding my variables such as tumor stages, histologies, surgery types, and patient status, and scaling of numerical variables. I ran a pycaret test that gave me an overall idea of which models may perform better and which ones I should avoid for the problem at hand.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.7951	0.4927	0.9889	0.8010	0.8846	0.0516	0.0611	0.0430
lr	Logistic Regression	0.7949	0.4806	1.0000	0.7949	0.8856	0.0000	0.0000	0.0250
ridge	Ridge Classifier	0.7949	0.0000	1.0000	0.7949	0.8856	0.0000	0.0000	0.0170
dummy	Dummy Classifier	0.7949	0.5000	1.0000	0.7949	0.8856	0.0000	0.0000	0.2560
lda	Linear Discriminant Analysis	0.7862	0.4650	0.9889	0.7928	0.8798	-0.0142	-0.0163	0.0150
gbc	Gradient Boosting Classifier	0.7723	0.4850	0.9608	0.7964	0.8700	0.0130	0.0136	0.0260
et	Extra Trees Classifier	0.7547	0.5570	0.9386	0.7908	0.8571	-0.0066	0.0070	0.0400
knn	K Neighbors Classifier	0.7379	0.4856	0.9222	0.7838	0.8464	-0.0567	-0.0555	0.0180
dt	Decision Tree Classifier	0.6615	0.4908	0.7817	0.7922	0.7844	-0.0329	-0.0324	0.0170
ada	Ada Boost Classifier	0.6611	0.4324	0.8092	0.7752	0.7901	-0.1041	-0.1108	0.0260
svm	SVM - Linear Kernel	0.6395	0.0000	0.7203	0.8159	0.7437	0.0104	0.0245	0.0160
qda	Quadratic Discriminant Analysis	0.6223	0.5760	0.6761	0.7880	0.7053	0.0951	0.0784	0.0160
nb	Naive Bayes	0.5941	0.4951	0.6987	0.7459	0.7023	-0.0743	-0.0960	0.0170

The results of this test gave me direction to pursue both random forest and logistic regression based models.

Two models were finalized after testing and tuning. Although the logistic regression provided a good initial baseline, the random forest models proved to be the best at predicting both tumor stage, and patient status. Below I will detail the results of those models.

Random Forest: Tumor Stage Prediction

```
Accuracy: 0.7732
Classification Report:
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        18
           2       0.75      0.93      0.83        58
           3       0.43      0.14      0.21        21

    accuracy                           0.77        97
   macro avg       0.73      0.69      0.68        97
weighted avg       0.73      0.77      0.73        97

Confusion Matrix:
[[18  0  0]
 [ 0 54  4]
 [ 0 18  3]]
```

The initial steps involve data preparation, where the dataset is divided into features (`X_stage`) and the target variable (`y_stage`). Subsequently, the data is split into training and testing sets. Hyperparameter tuning is performed using grid search, exploring different values for the number of estimators (`n_estimators`) and the maximum depth (`max_depth`). The best hyperparameters are determined through cross-validation. A Random Forest model is then instantiated with these optimal parameters, trained on the training set, and evaluated on the test set.

The results of the Random Forest model reveal an overall accuracy of approximately 77.32%. The classification report provides a detailed breakdown of precision, recall, and F1-score for each tumor stage class. Notably, the model demonstrates outstanding performance for the first class, achieving perfect precision, recall, and F1-score. It also performs reasonably well for the second class, showing a balance between precision and recall. However, challenges arise in accurately predicting instances of the third class, as indicated by lower precision, recall, and F1-score values. These results offer insights into the model's strengths and weaknesses, guiding potential improvements or further investigations, particularly for enhancing predictions in the challenging third class.

If stage 3 has significantly fewer instances, it might contribute to the model's difficulty in predicting this class. Techniques like oversampling or undersampling could be explored. Stage 1 and Stage 3 have significantly lower instances than stage 2. Stage 1 may have been correctly predicted due to other indicating factors of an initial condition. With stage 2 and Stage 3, those indicative features could be less clear and apparent, as the condition is now in advanced stages.

Random Forest: Patient Status

```
Accuracy: 0.7538
Classification Report:
              precision    recall  f1-score   support

           0       0.33      0.07      0.11        15
           1       0.77      0.96      0.86        50

    accuracy                           0.75        65
   macro avg       0.55      0.51      0.48        65
weighted avg       0.67      0.75      0.68        65

Confusion Matrix:
[[ 1 14]
 [ 2 48]]
```

I implemented another Random Forest classifier for predicting patient status (alive or dead) based on the breast cancer dataset. Initially, the data is prepared, with features denoted as `X_stat` and the target variable as `y_stat`. The dataset is then split into training and testing sets, and a Random Forest model is trained using 200 estimators. The model achieves an overall accuracy of 75.38%. However, an in-depth analysis reveals that while the model performs well in predicting the "alive" class (1), it struggles with the "dead" class (0), as evident from the low precision, recall, and F1-score for Class 0.

To address this, hyperparameter tuning is performed using grid search to find the optimal parameters for the Random Forest model. The best parameters obtained are a maximum depth of None and 200 estimators. Despite the parameter tuning, the final model's performance on the test set remains consistent, indicating that the chosen hyperparameters might not significantly impact the model's ability to predict Class 0. The detailed classification report sheds light on the strengths and weaknesses of the model for each class, providing a nuanced understanding of its predictive capabilities.

Future Considerations

Future Research Questions

1. **Molecular Subtypes**: Can we delve deeper into the dataset and categorize breast cancer cases into distinct molecular subtypes? How do these subtypes correlate with the identified protein levels and receptor statuses? This exploration aims to unravel more nuanced insights into the intricate characteristics of the disease.

2. **Long-Term Follow-up**: What unfolds in the long-term for patients based on their initial protein levels and receptor status? A longitudinal study could offer a more comprehensive understanding of how these factors unfold over time, providing valuable insights into the lasting impact on patient outcomes.

4. **Genetic Factors**: Is there a genetic underpinning to the observed variations in protein levels and receptor statuses? Exploring the genetic landscape of breast cancer patients might uncover additional layers of complexity in the progression of the disease.

5. **Incorporating Lifestyle Factors**: How do lifestyle factors, such as diet, exercise, and stress, intersect with protein levels and receptor status in shaping breast cancer outcomes? Integrating lifestyle data could enrich our understanding of the holistic dynamics of the disease.

Future Considerations:

1. **Data Diversity**: How representative is our current dataset, and what steps can we take to enhance diversity, considering variables like ethnicity, socioeconomic status, and geographical location? A more diverse dataset would undoubtedly enhance the applicability of our findings.

2. **External Validation**: How do the correlations we've identified hold up when applied to external datasets or real-world clinical settings? External validation is paramount for affirming the robustness and applicability of our findings beyond the confines of our dataset.

3. **Ethical Implications**: What ethical considerations should guide the development and implementation of prognostic tools based on protein levels and receptor status? Addressing potential

biases and navigating ethical concerns is pivotal for the responsible application of our insights in the realm of healthcare.