Breast Cancer Survival Prediction

Advanced Machine Learning Project Report

This project is organized within the specified Github Repository [1], where all scripts pertinent to the sections outlined below are available. Additionally, the applications have been included in the provided ZIP file.

Contents

Abstract

This project explains the creation of a Machine Learning model to forecast breast cancer survival using real data from breast cancer patients.
The first part of the report describes the preprocessing steps, including analyzing outliers, feature extraction and data exploration. It then details the feature selection techniques used.
Finally, the modeling section is addressed, implementing multiple models. Each model is evaluated based on precision, recall, and F1-scores, with Random Forest Classifier obtaining the best results.

1 Introduction

This project aims to forecast breast cancer survival using real data from breast cancer patients. These kinds of studies typically monitor patients for a certain period until the cut-off date. The cut-off date is the point when the researchers stop collecting data or consider the data "final" for analysis. With our model, we will try to predict if the patient will be alive on the cut-off date. The `Status` variable records whether the patient was dead or alive on the cut-off date. This approach ensures that the analysis only considers deaths that occurred during the study period.
Our model utilizes the Breast Cancer dataset on Kaggle [2]. This dataset of breast cancer patients was obtained from the 2017 November update of the SEER Program of the NCI, which provides information on population-based cancer statistics. The dataset involved female patients with infiltrating duct and lobular carcinoma breast cancer diagnosed in 2006-2010. Patients with unknown tumour size and patients whose survival months were less than 1 month were excluded.
The dataset contains **4024 patients** each with **16 features** described in the following tables 3 and 4.

2 Data exploration

In this section, we will perform a univariate and bivariate exploratory analysis to gain a deeper understanding of the data, uncovering patterns and detecting anomalies allowing us to create additional features for a possible enhancement in the modeling phase as well as a deeper understanding of the behaviors of the different models.

Before starting, we have removed the `Grade` column because as explained in Table 4 it is the grade associated with the level of differentiation of the tumor, meaning it contains the same information as the `Differentiate` variable. Additionally, it's important to note that our dataset does not include any missing values so no imputations will be necessary.

2.1 Univariate Exploratory Analysis

We start by obtaining a comprehensive overview of all the features in our dataset studying the distribution of both numerical and categorical variables. Figure 1 displays the different distributions for each variable representing the categorical variables with a barplot and the numerical with a histogram and a boxplot. The boxplot contains a horizontal red line separating the extreme outliers from the rest of the patients.

Observing the last plot of the Figure 1, we can see that the target variable `Status` is very unbalanced, having almost 85% of the patients classified as Alive. This imbalance can lead to biased predictions and learning difficulties. To solve this, resampling techniques will be applied (See Section 2.5) as well as specific metrics such as F1-score with the minority class serving as

the positive class.

Moving on to the rest of the variables and focusing on the categorical variables, we can extract the following information for each variable:

- `T Stage`: T1 and T2 stages dominate the majority of the data, while T3 and T4 stages have minimal representation.
- `N stage`: N1 observations stand out with 67.9%, followed by a progressively decreasing number of observations in N2 and N3.
- `6th Stage`: a progressive decreasing pattern also appears going from IIA to IIIC with the anomaly of the category IIIB, which has very few appearances.
- `differentiate`: most of the tumors are classified as moderately differentiated with both extremes representing a small fraction of the dataset.
- `A Stage` is very uneven, with only 2.3% of the observations having a Distant value.
- `Estrogen Status` and `Progesterone Status` variables have a high percentage of positive observations, with 93.3% and 82.65% respectively, indicating an imbalance.
- `Status`: nearly 85% of the observations are labeled as Alive.

Moving on to the numerical variables:

- `Age` variable histogram reveals a notable increase in patients aged 50 compared to 30, remaining steady until 60 before declining. The boxplot shows no outliers.
- `Tumor Size` variable is left-skewed, suggesting mostly small tumors, but with extreme outliers (>105 mm).
- `Regional Node Examined` feature exhibits left-skewness, with some extreme outliers where patients had more than 50 regional lymph nodes examined.
- `Regional Node Positive` variable is heavily left-skewed, with 38% of patients having no regional positive nodes, and extreme outliers observed in patients with more than 18 positive examined nodes.
- `Survival Months` variable displays two distinct behaviors, with a notable increase in counts around the 50-month mark. This suggests a transition in the survival journey.

The variables with extreme outliers are `Tumor Size`, `Regional Node Examined` and `Regional Node Positive`. In order to take it into account a new column counting the number of extreme outliers has been added, but for now, no further actions have been taken since the outliers are logical and do not seem to be erroneous. This is because both tumor sizes and positive regional nodes often exhibit extreme upper outliers due to exponential tumor growth, late

detection, variability in tumor aggressiveness, and delays in treatment, leading to a few cases with significantly larger tumors and more extensive lymph node involvement.

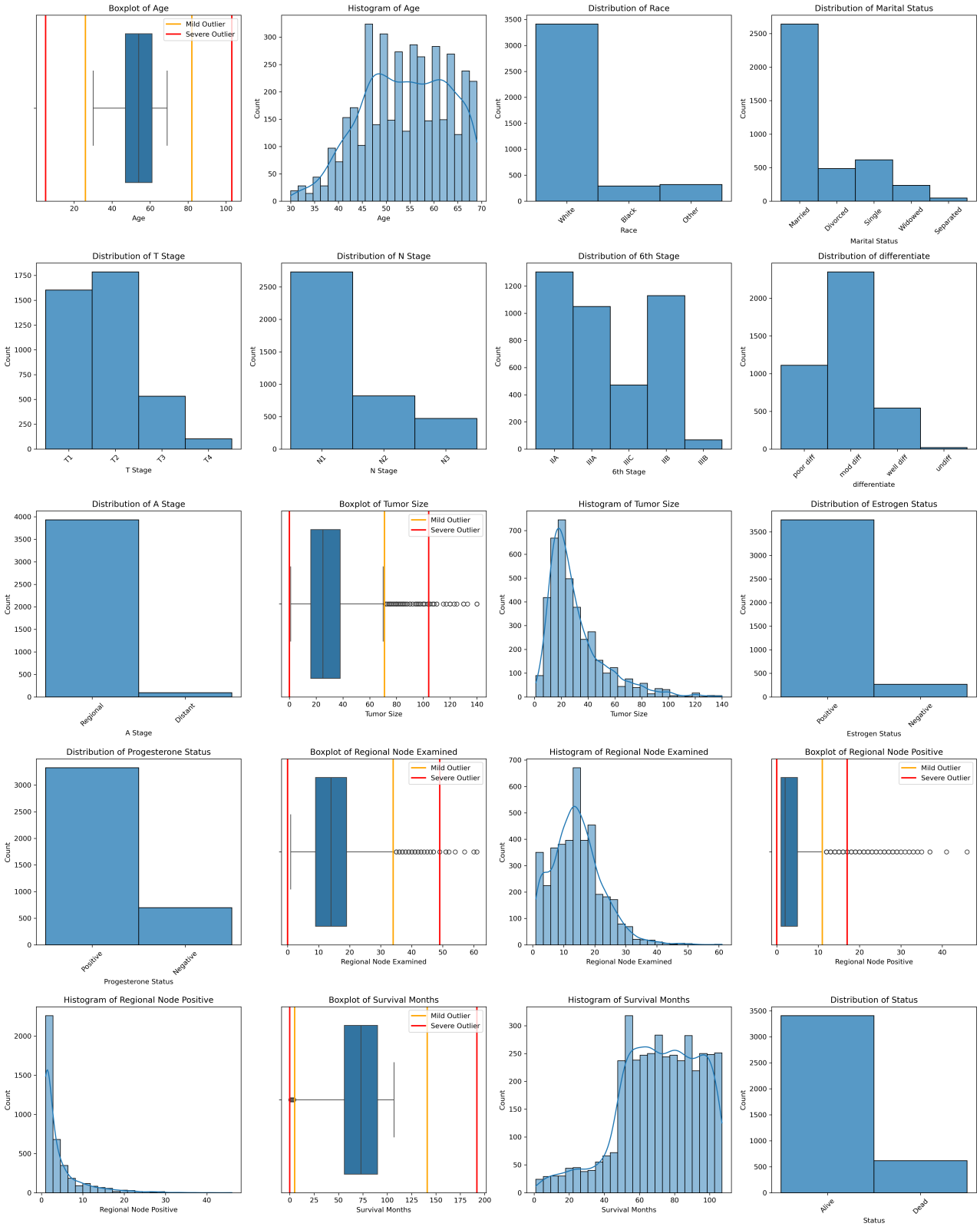

Figure 1: Univariate Analysis

2.2 Bivariate Exploratory Analysis

Following the univariate analysis, we will examine the relationship between pairs of variables concerning the target variable `Status`. Starting with the numerical variables, in Figure 2 we can see a matrix of scatter plots containing all combinations of numerical variables categorized by `Status`, and on the diagonal two density plots overlapping showing the distribution of that variable with alive and dead patients.

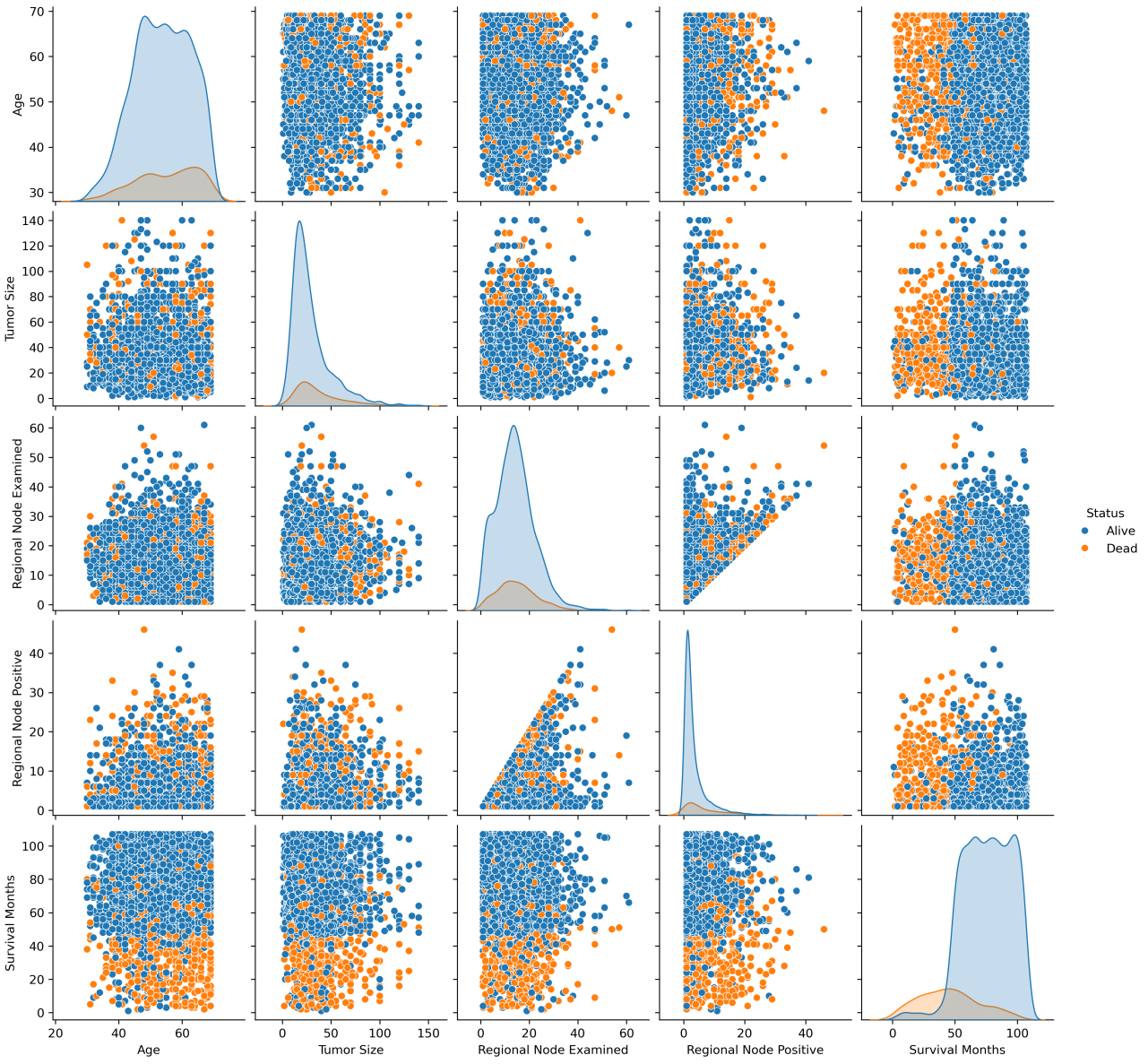

Figure 2: Bivariate Analysis Numerical

In this type of analysis, it is crucial to pay special attention to variables that clearly separate these two classes, indicating it has a strong discriminative power making it a valuable feature for classification. A clear example is the `Survival Months` feature with a clear discriminative behavior in all the scatter plots. The scatter plot from `Regional Nodes Examined` and `Regional Node Positive` reinforces the quality of the data and the lack of erroneous data. This is because there are no patients with more positive regional nodes than examined nodes

shown by the triangular shape on the plot, with the diagonal having a positive slope of 1. Additionally, it can be seen that there is a higher concentration of deceased patients closer to the diagonal meaning that the ratio of examined nodes divided by positive nodes could have a high discriminative power.

In Figure 3 we can see two box plots for each numerical variable separated by the `Status` indicating the tendencies of each feature with respect to the target variable. We see that age surprisingly does not have a clear impact on the outcome of the patient as well as examined regional nodes. On the other hand, tumor size and regional positive nodes exhibit an inverse correlation with the outcome. The survival months box plots support what was seen in the previous figure with a clear discriminative power on the target variable.

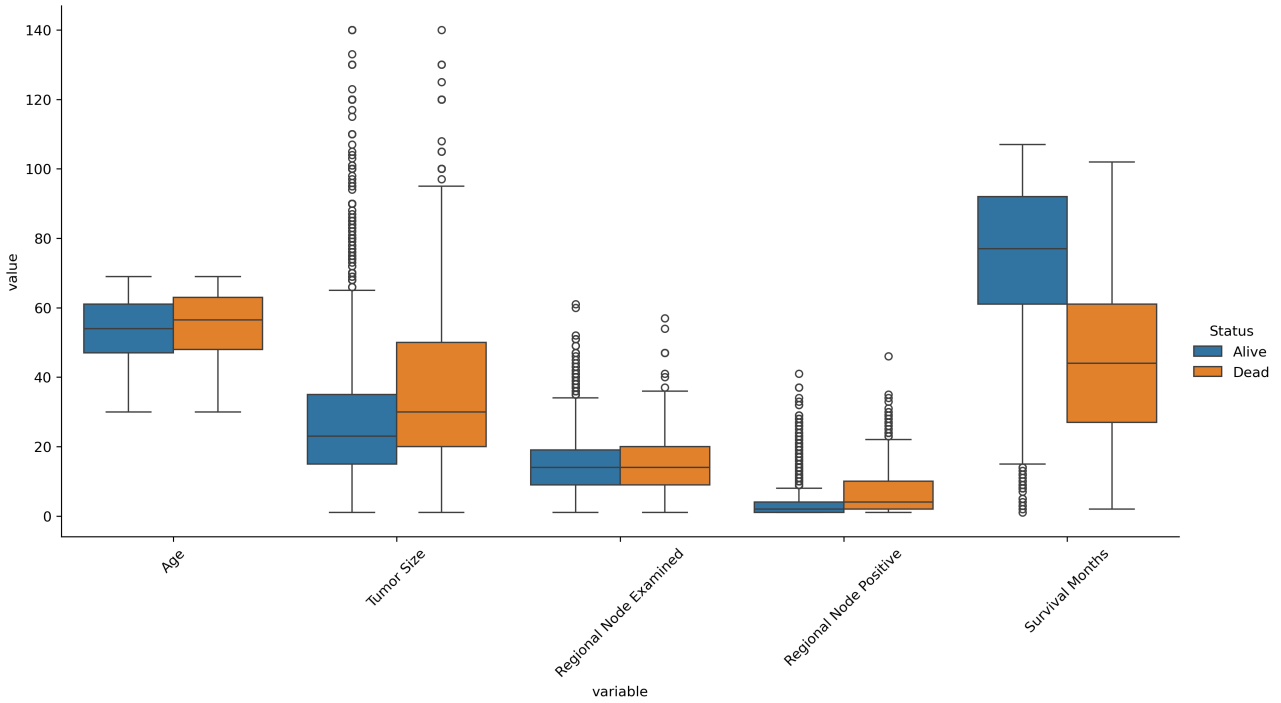

Figure 3: Bivariate Analysis Numerical Box Plots

To continue with the analysis, we will focus on the categorical variables. Figure 6 displays a bar chart of each feature comparing the proportions of alive and dead patients. The charts account for the imbalanced nature of the dataset by normalizing the proportions within each category of the target variable, allowing for a direct comparison. In marital status, there is an interesting tendency with patients who are married to have a slight survival advantage over those who are not. Among the T Stage and N Stage variables, we observe a clear pattern: as the stage increases, the proportion of deceased patients rises. These stages reflect disease progression, making them strongly indicative of survival outcomes. The A Stage variable is also discriminative as patients classified in the "Distant" category show a higher proportion of deaths compared to those in the "Regional" category. The hormone receptor status variables, including Estrogen Status and Progesterone Status, also show strong discriminative power. Patients with Negative Estrogen Status or Negative Progesterone Status have a higher percentage of deaths compared to those with positive statuses.

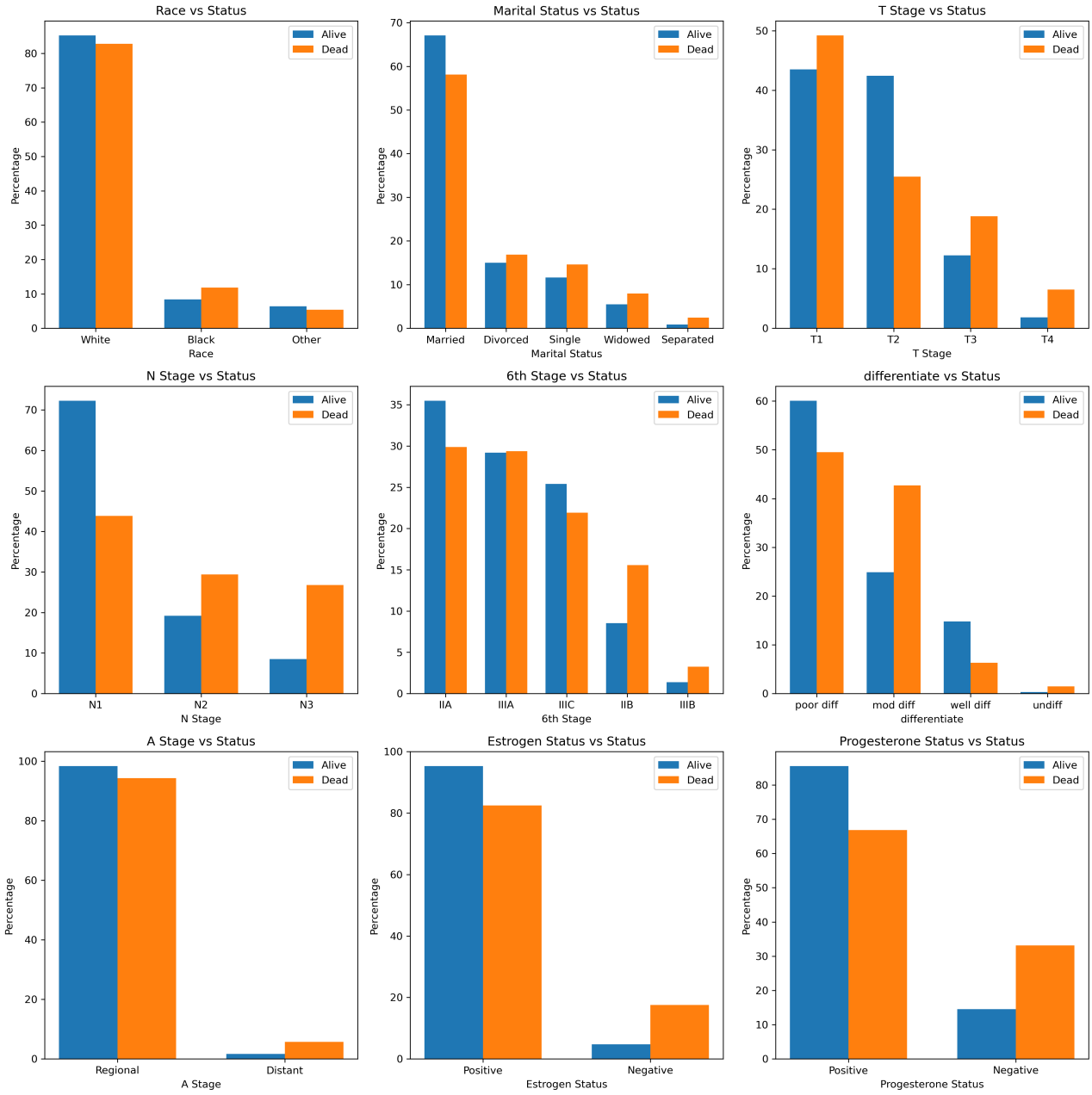

Figure 4: Bivariate Analysis Categorical

2.3 Feature Extraction

As mentioned before in the bivariate exploratory analysis, the bivariate scatter plot seen in Figure 2 shows a higher concentration of deceased patients closer to the diagonal. Taking this into account, we have created a new variable called `Node Ratio` with the following formula, where γ represents a small positive value to prevent computational errors and ensure the variable is meaningful and stable across all observations in the dataset.

$$\text{Node Ratio} = \frac{\text{Regional Positive Nodes} + \gamma}{\text{Regional Examined Nodes} + \gamma}$$

Figure 5 shows a box plot categorized by status, illustrating the discriminative ability of the variable. The *Dead* distribution exhibits much higher values compared to the *Alive* class.

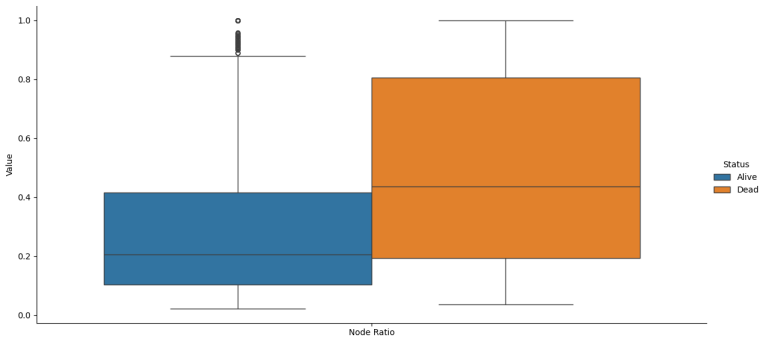

Figure 5: Node Ratio Box Plot

2.4 Multivariate Outliers Analysis

We will conduct multivariate outlier detection on the numerical variables of the dataset using Mahalanobis Distance. To achieve this, we need to calculate the covariance matrix and its inverse to measure the distances between dataset observations. We've set the cutoff value for outlier detection at 1%, corresponding to a desired significance level of 0.99 based on the Chi-Square distribution.

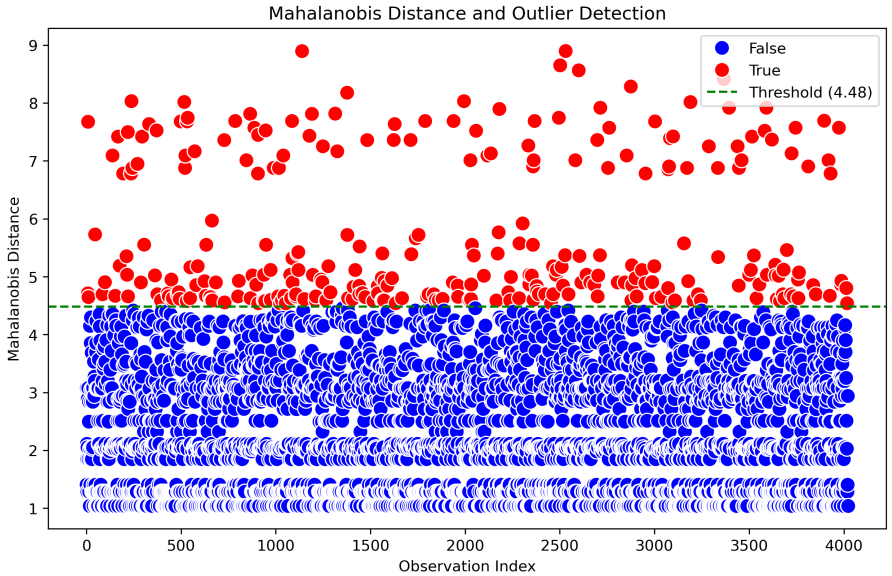

Figure 6: Mahalanobis Distance and Outliers

As shown in the figure above, there are 276 multivariate outliers, but after observation, we have considered keeping them because as mentioned before the quality of the data is high and therefore the multivariate outliers don't seem erroneous.

2.5 Splitting and Resampling

We chose to split the data into 80% training and 20% test data. This approach allows us to use a larger portion of the dataset for training, which helps in building a more generalized model. We decided not to include a separate validation set due to the limited size of the dataset and the class imbalance. A dedicated validation set might not adequately represent all classes, leading to unreliable performance evaluation. Instead, we opted for a larger training set, enabling the use of cross-validation techniques for internal model validation during training, while maintaining a reliable test set to evaluate the model's performance on a representative sample of the overall data distribution.

To address the imbalance in the target variable, we will explore two techniques listed below.

- **Oversampling Method:** We will use SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous) to increase the number of samples in the minority class. Undersampling techniques were not considered because, in medium-sized datasets like this one, they can discard valuable information and potentially distort class relationships.
- **Set Class Weights:** We will adjust the model's loss function to penalize misclassification of minority class samples (set as the positive class) more heavily without altering the dataset.

Therefore, we will have two distinct datasets for the analysis: one oversampled and one not. The feature selection steps described in the next section were performed on both datasets; however, as the results were very similar, we will only present the plots for the dataset without oversampling.

2.6 Feature Selection

Once we have cleaned and explored the dataset, we will start our feature selection analysis.

2.6.1 Features Correlation

The ordinal categorical variables (`T Stage`, `N Stage`, `6th Stage`, and `differentiate`) have been encoded and converted to numerical in order to be included in the correlation matrix.
If we observe the correlation matrix found in Figure 7, we can see that there is a high correlation between the variables `6th Stage`, `N Stage` and `Regional Node Positive`. In addition, the variables `T Stage` and `Tumor Size` also exhibit a high correlation as expected. Highly correlated variables in a model can inflate the variance of coefficient estimates, reduce model interpretability, and lead to unreliable and unstable predictions. To mitigate this issue we will use the dimensionality reduction technique (PCA) applied in section 2.6.2, combining correlated variables into a set of uncorrelated dimensions which in our case will be the principal components.

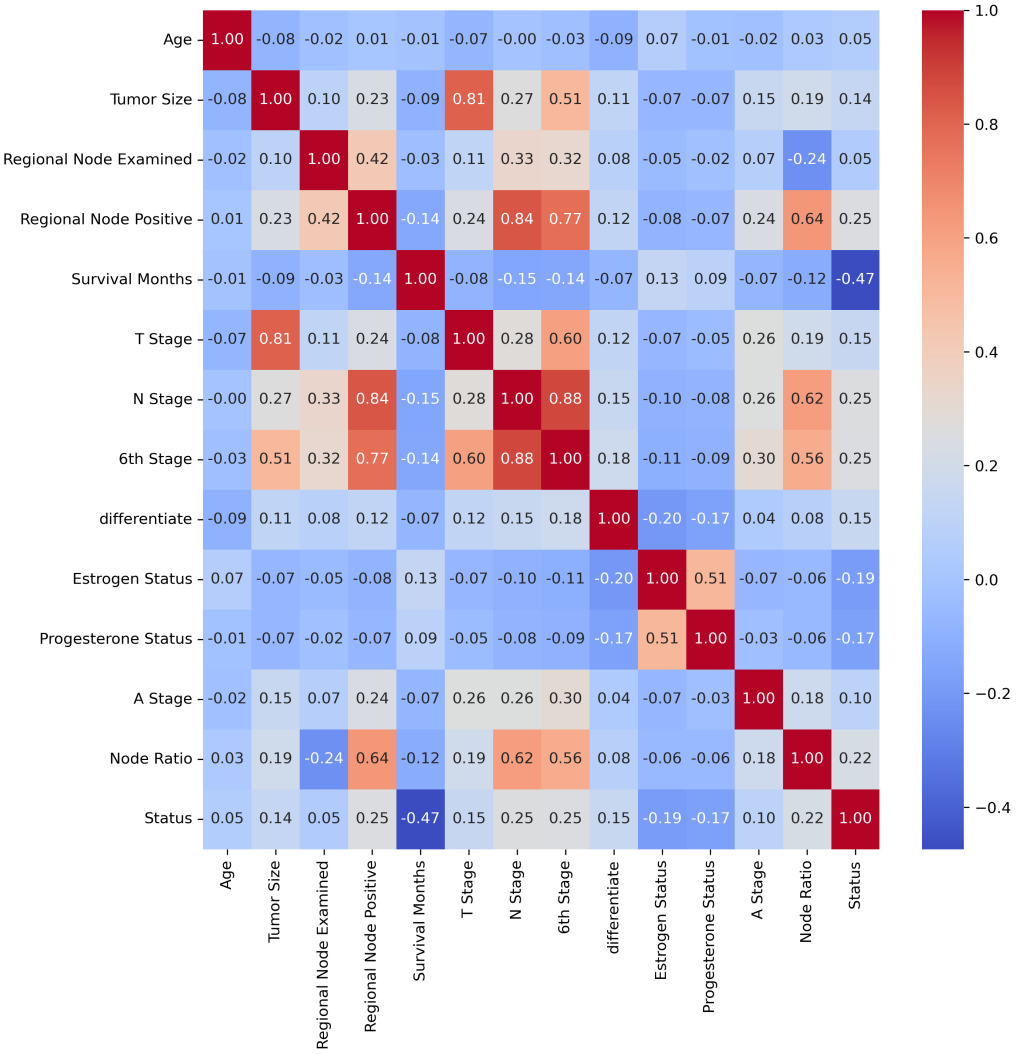

Figure 7: Correlation Matrix

When evaluating the correlation between categorical variables, we selected the Cramér's V measure due to its suitability for assessing the strength of association between two categorical variables in a contingency table. In Figure 8, we present a heatmap displaying the results of the Cramér's V measures to identify variables with significant correlations. The only significant association is between `Estrogen Status` and `Progesterone Status`, with a Cramér's V value of 0.51, indicating a moderate correlation.

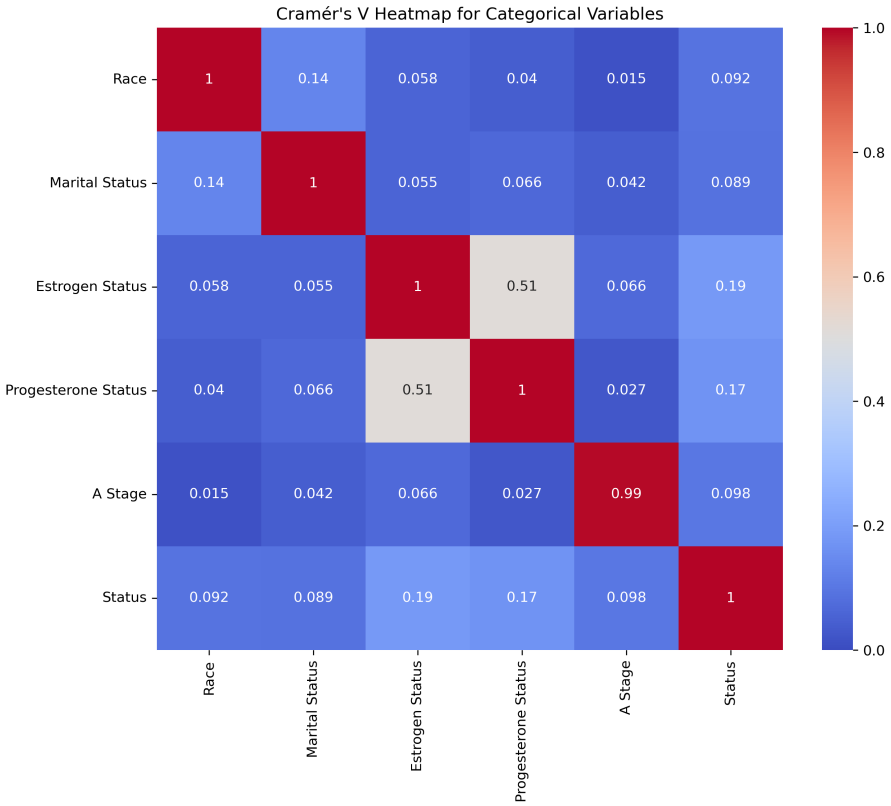

Figure 8: Cramér's V Scores

After that, we use the Mutual Information (MI) measure to determine how much information each categorical feature contributes to predicting the target. Observing the results in Table 1, `Progesterone Status` has a higher feature importance compared to `Estrogen Status`. Furthermore, `Maritial Status` and `Race` showed very low MI scores. Therefore, we decided to remove these features to improve model performance and reduce dimensionality.

Feature	**MI Score**
Estrogen Status	0.013212
Race_Other	0.008955
Progesterone Status	0.007202
Marital Status_Separated	0.004133
A Stage	0.003703
Marital Status_Single	0.002988
Marital Status_Married	0.000387
Race_White	0.000000
Marital Status_Widowed	0.000000

Table 1: MI Scores for Categorical Features

2.6.2 Principal Component Analysis (PCA)

We will performed a PCA on the numerical variables with the aim of reducing dimensionality while preserving most of the variance in the data.

To ensure that all features contribute equally to the PCA, we scaled each variable. Variables without severe outliers were scaled using the StandardScaler, while variables with severe outliers (i.e. `Tumor Size`, `Regional Node Examined`, `Regional Node Positive`) were scaled using the RobustScaler, which is more robust to outliers.

Figure 9 depicts the cumulative explained variance as a function of the number of principal components. It shows that with the first six principal components, we are able to capture more than 90% of the total variance in the data.

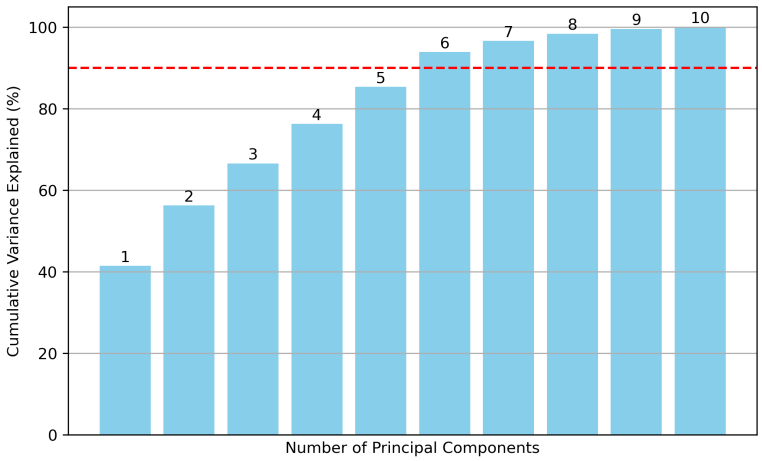

Figure 9: PCA Cumulative Explained Variance

In the Factorial Map of Figure 10, each point represents a patient. The color of the point indicates whether the patient is alive (red) or deceased (blue). From this scatter plot, it is not possible to observe a clear separation between the two types of patients. We also attempted to apply Kernel PCA to capture more complex relationships and plotted the results in a 3D space with the first three principal components. However, the results were not satisfactory.

Analyzing the centroids of each class in Figure 11, we can see that deceased patients tend to have more positive regional nodes and a higher N Stage. Other variables, such as Tumor Size, 6th Stage, or Age, do not show clear discriminative patterns in this map.

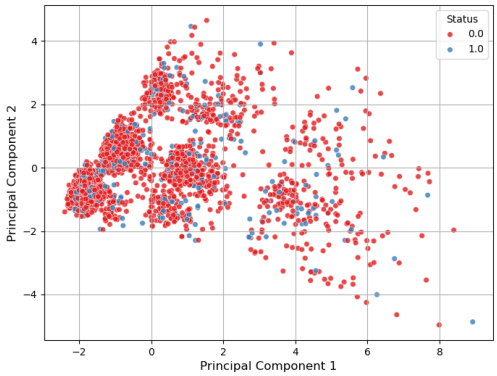

Figure 10: Factorial Map with Target Variable distribution

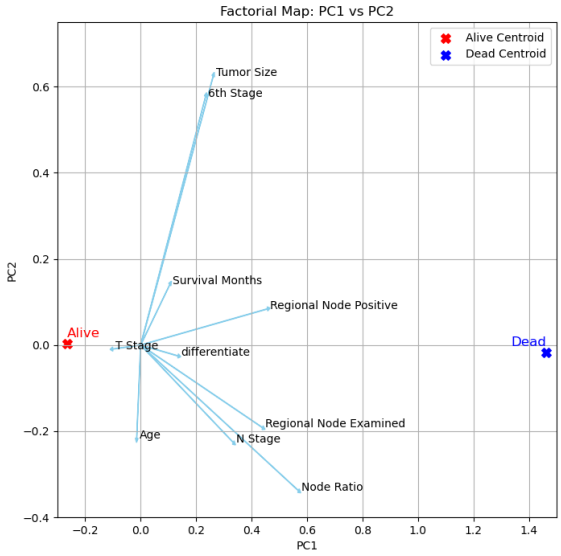

Figure 11: Factorial Map PC1 vs PC2

2.6.3 Best Features Choice

We will now apply a final step of Recursive Feature Elimination (RFE) to extract the best features from each dataset. The considered features include all ten Principal Components, the binary variables (`Estrogen Status`, `Progesterone Status`, and `A Stage`), and the two most discriminative variables (`Survival Months` and `Node Ratio`).

This process begins by fitting a Random Forest Classifier using all the variables and extracting a list of features sorted by importance. Subsequently, the model is iteratively trained using subsets of features, removing the least important feature at each iteration. At each step, the model's performance is evaluated using the Out-of-Bag (OOB) error. Finally, the subset of features that yields the best performance is selected.

As we are working with medical data, minimizing the number of false negatives (i.e., a dead patient classified as alive) is crucial. To address this, we will use the weighted Recall Score metric for model evaluation. Additionally, we have decided to consider a score better only if it is at least 0.001 higher than the previous best score. By setting a minimum improvement criterion, we avoid selecting subsets of features that may not offer substantial enhancements in model performance. The following diagram provides a more illustrative explanation of the process.

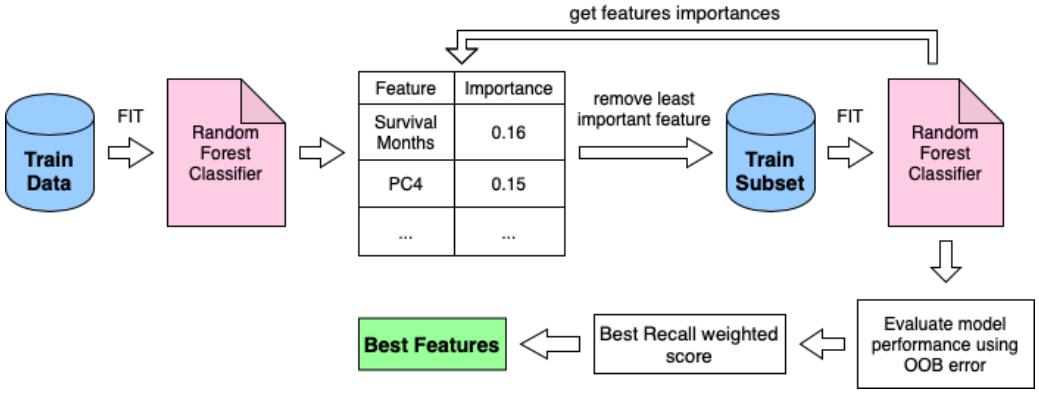

Figure 12: Recursive Feature Elimination Diagram

Pros:

- Random Forest inherently ranks features based on importance, making it a natural choice for RFE.
- The ensemble nature of Random Forest helps mitigate overfitting, especially for datasets with a mix of relevant and irrelevant features. It can also
- Random Forest works well with binary, continuous, and categorical variables, which is ideal for medical datasets.
- The OOB error provides an internal performance estimate without needing a separate validation set, reducing data wastage.

Cons:

- Iteratively fitting the model and evaluating subsets of features can be computationally expensive. While we considered using Decision Trees instead of Random Forests due to their computational efficiency, they are more prone to overfitting. As time was not a significant constraint for us, we prioritized obtaining a more accurate model over computational efficiency.
- If the Random Forest struggles with the dataset, the feature importance rankings might not be reliable.
- Random Forest classifiers are not gradient-based, which could make RFE less precise in identifying optimal feature subsets compared to methods like gradient boosting.

Results:

For the original dataset we obtained the following subset of 14 features:
`All 10 PCs`, `Survival Months`, `Node Ratio`, `Estrogen Status`, `Progesterone Status`

For the oversampled dataset we obtained the following subset of 10 features:
`PC1`, `PC3`, `PC4`, `PC5`, `PC6`, `PC8`, `PC9`, `PC10` `Survival Months`, `Node Ratio`

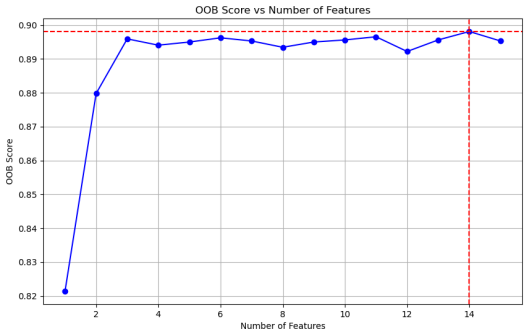

Figure 13: OOB Recall Score vs Number of Features (Original Dataset)

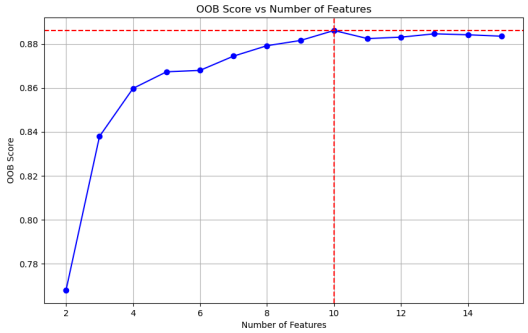

Figure 14: OOB Recall Score vs Number of Features (Oversampled Dataset)

3 Modeling and Results

In this section, we present the results of five classification models applied to the dataset: Ridge Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Neural Networks, and Random Forest. Each model was tuned using optimal hyperparameters, and the performance was evaluated on both the original and oversampled datasets.

3.1 Ridge Logistic Regression

For Ridge Logistic Regression, the best regularization parameter C was determined to be 2.78. The model performed well on the original dataset, achieving an accuracy of 90% with a precision of 0.91 and a recall of 0.97 for class 0. However, the model struggled with class 1, where the recall was only 0.49, resulting in a relatively low F1-score of 0.59. This indicates that the model was not effective in detecting the minority class.

When oversampling was applied to the dataset, the performance for class 1 improved significantly, with a recall of 0.77. However, this came at the cost of a reduced precision of 0.49 for class 1, and the overall accuracy dropped to 84%.

3.2 Support Vector Machine

For the Support Vector Machine, the best regularization parameter C was found to be 100. The SVM achieved an accuracy of 90% on the original dataset, with excellent precision and recall for class 0 (precision = 0.92, recall = 0.97). However, the recall for class 1 was lower, at 0.50, which led to an F1-score of 0.61 for this class.

After oversampling, the SVM showed a notable improvement in recall for class 1, reaching 0.67, with an accuracy of 88%. The precision for class 1 was 0.60, and the F1-score for class 1

increased to 0.64. This demonstrates that the SVM model handled the class imbalance better than Ridge Logistic Regression.

3.3 K-Nearest Neighbors

The K-Nearest Neighbors algorithm was tuned with 11 neighbors, which produced the best results. On the original dataset, KNN achieved an accuracy of 90%, similar to Ridge Logistic Regression and SVM. The precision for class 0 was 0.92, and the recall was 0.97. However, the model's performance on class 1 was suboptimal, with a recall of only 0.50 and an F1-score of 0.60.

After oversampling, the performance for class 1 improved slightly, with a recall of 0.59 and a precision of 0.44. However, the overall accuracy dropped to 82%. The KNN model demonstrated lower effectiveness compared to other models when dealing with the imbalanced dataset.

3.4 Neural Network

The Neural Network model was tuned with the following best parameters: `alpha` = 0.0001, `hidden_layer_sizes` = (50,), and `max_iter` = 200. On the original dataset, the Neural Network achieved an accuracy of 90%, with a precision of 0.91 and a recall of 0.98 for class 0. The recall for class 1 was 0.49, and the F1-score for class 1 was 0.60.

After applying oversampling, the recall for class 1 increased to 0.69, and the precision was 0.53. The overall accuracy was 86%. Despite the improvement in performance for the minority class, the Neural Network did not outperform the SVM in terms of F1-score or accuracy.

3.5 Random Forest Classifier

The Random Forest Classifier was evaluated using the oversampled dataset with the following best hyperparameters: `n_estimators` = 150, `max_samples` = 0.8, `max_features` = `sqrt`, and `max_depth` = None. It achieved an accuracy of 87% and an F1-score of 0.59 for class 1. The recall for class 1 was 0.62, and the precision was 0.56, indicating that the model performed well but did not surpass the SVM in handling the class imbalance.

On the original dataset, the Random Forest was tuned with the best hyperparameters: `n_estimators` = 100, `max_samples` = 0.7, `max_features` = `None`, and `max_depth` = None. It achieved an accuracy of 91%, with a precision of 0.78 and a recall of 0.58 for class 1.

3.6 Model Comparison

The table below summarizes the performance of each model based on both the original and oversampled datasets. It shows the accuracy, precision, recall, and F1-score for class 1 (minority class - Dead) for each model:

Model	Dataset	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Ridge Logistic Regression	Original	0.90	0.76	0.49	0.59
	Oversampled	0.84	0.49	0.77	0.60
SVM	Original	0.90	0.78	0.50	0.61
	Oversampled	0.88	0.60	0.67	0.64
KNN	Original	0.90	0.75	0.50	0.60
	Oversampled	0.82	0.44	0.59	0.50
Neural Network	Original	0.90	0.78	0.49	0.60
	Oversampled	0.86	0.53	0.69	0.59
Random Forest	Original	**0.91**	**0.78**	**0.58**	**0.66**
	Oversampled	0.87	0.56	0.62	0.59

Table 2: Comparison of Classification Models on Original and Oversampled Datasets

From the table, the Random Forest Classifier achieved the best performance on the original dataset, with the highest F1-score for class 1 (0.66) and accuracy (91%). Although oversampling improved performance for other models like SVM and Neural Networks, the RFC with the original dataset remained superior overall.

This decline in performance for the oversampled dataset is likely attributed to the SMOTE oversampling technique failing to adequately capture the true underlying characteristics of the "Dead" class, thereby impeding the model's ability to generalize effectively. Additionally, the features possibly do not contain all the necessary information to truly discriminate between classes, further limiting the effectiveness of oversampling.

4 Final Model

The final model selected for the classification task is the Random Forest Classifier rained on the original dataset. Moreover, the chosen hyperparameters help in reducing computational cost and improving generalization:

- `n_estimators`: 100
- `max_samples`: 0.7, indicates that each tree in the forest will be trained on a random sample consisting of 70% of the original training dataset. This ensures that each tree is slightly different, contributing to the diversity of the forest and reducing the computational cost of the model.
- `max_features`: None, the algorithm will consider all features for splitting at each tree node.
- `max_depth`: None

The RFC model was trained on the original dataset to leverage the dataset's inherent structure without introducing synthetic data. The key strengths of the model include:

- High accuracy (91%) and F1-score (0.66) for the minority class, demonstrating its effectiveness in handling the original imbalanced dataset.

- Consistent performance across cross-validation, with low variance and robust generalization error estimates.
- Ability to handle non-linear relationships and interactions between features effectively, making it particularly well-suited for complex medical datasets.

This model will be used to predict outcomes in future data, and its performance will be monitored to ensure sustained effectiveness. Additional optimization or retraining may be considered as new data becomes available.

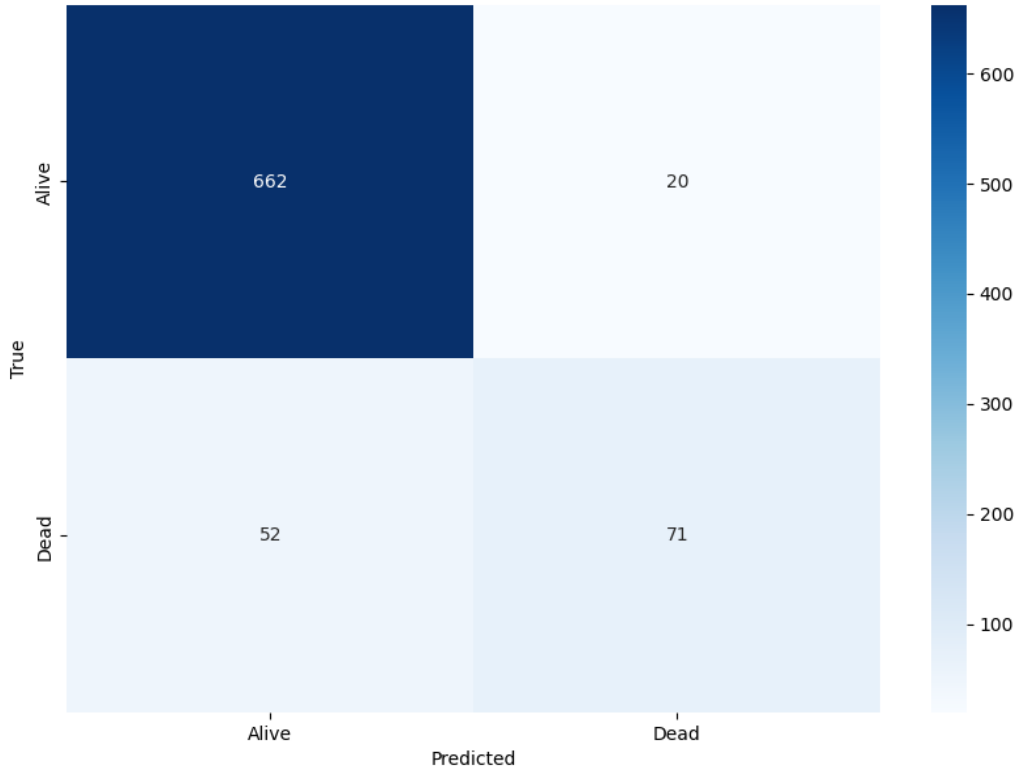

Figure 15: Confusion Matrix of the final model

5 Conclusions

In this study, we developed and evaluated several machine learning models to predict breast cancer survival outcomes using a dataset of 4024 patients. After preprocessing, feature selection, and rigorous evaluation, the Random Forest Classifier trained on the original dataset emerged as the best-performing model.

The RFC achieved an accuracy of 91% and an F1-score of 0.66 for the minority class (deceased patients) on the original dataset, outperforming other models, including those trained on the oversampled dataset. While oversampling with SMOTENC improved minority class recall for some models, such as the Support Vector Machine and Neural Network, the RFC trained on the original dataset provided a superior balance between precision and recall for both classes.

Key takeaways from the study include:

- The RFC demonstrated consistent performance across cross-validation, with low variance and robust generalization, making it the most reliable model for deployment.
- The oversampled dataset struggled to adequately capture the underlying characteristics of the minority class, limiting its effectiveness and resulting in reduced generalizability.
- Feature selection, including techniques like Recursive Feature Elimination and Principal Component Analysis, enhanced model interpretability and reduced overfitting.

Overall, the results highlight the effectiveness of the RFC in handling imbalanced datasets without introducing synthetic data and its ability to generalize well to unseen data.

6 Limitations and Future Work

6.1 Limitations

Despite the success of the Random Forest Classifier, this study has several limitations:

- **Class Imbalance:** The dataset's inherent class imbalance posed challenges for accurately predicting the minority class, even with techniques like oversampling and class weighting.
- **Feature Representation:** The dataset's features may not fully capture all factors influencing survival outcomes. Incorporating additional clinical, genetic, or molecular data could enhance model performance.
- **Oversampling Limitations:** The SMOTENC technique used for oversampling introduced synthetic data that failed to adequately represent the minority class, which negatively impacted model generalization.
- **Model Complexity:** While the RFC performed well, its computational cost can be significant, particularly when fine-tuning hyperparameters or handling larger datasets.

6.2 Future Work

Building on the findings of this study, several directions can be explored to address the limitations and further improve predictive performance:

- **Expanding the Dataset:** Gather more recent and diverse datasets that include a broader range of patient demographics, treatments, and outcomes.
- **Improving Feature Representation:** Include richer clinical features, such as imaging data, genetic markers, or patient lifestyle information, to improve model accuracy and interpretability.
- **Explainability and Clinical Relevance:** Implement explainable AI (XAI) techniques, such as SHAP or LIME, to provide insights into how the model makes predictions, aiding clinical decision-making.

Appendices

Feature	**Type**	**Levels**
Age	Numerical	30-69
Race	Categorical	White, Black, Other
Marital Status	Categorical	Married, Single, Divorced, Widowed, Separated
T Stage	Categorical	T1 (<2cm), T2 [2cm,5cm], T3 (>5cm), T4 (spread into the chest wall) [3]
N Stage	Categorical	N1 (<3 lymph nodes), N2 [4,9] lymph nodes, N3 (>10 lymph nodes) [3]
6th Stage	Categorical	IIA, IIIA, IIB, IIIB, IIIC [3]
Differentiate	Categorical	Well, Moderately, Poorly, Undifferentiated
Grade	Categorical	1 (Well), 2 (Moderately), 3 (Poorly), 4 (Undifferentiated)
A Stage	Categorical	Regional, Distant
Tumor Size	Numerical	1-140 mm
Estrogen Status	Categorical	Positive, Negative
Progesterone Status	Categorical	Positive, Negative
Regional Node Examined	Numerical	1-61
Regional Node Positive	Numerical	1-46
Survival Months	Numerical	1-107 months
Status	Categorical	Alive, Dead

Table 3: Feature Types and Levels

Feature	**Description**
Age	Age of the patient at diagnosis
Race	Race or ethnicity of the patient
Marital Status	Marital status of the patient
T Stage	Size and extent of the primary tumor
N Stage	Spread of cancer to lymph nodes
6th Stage	Overall stage of cancer-based on T, N, and M stages
Differentiate	How abnormal the cancer cells look under a microscope
Grade	Grade associated with differentiation
A Stage	Extent of cancer spread beyond the primary tumor site
Tumor Size	Size of the primary tumor
Estrogen Status	Status of estrogen receptor in tumor tissue
Progesterone Status	Status of progesterone receptor in tumor tissue
Regional Node Examined	Number of regional lymph nodes that were removed and examined by the pathologist
Regional Node Positive	Number of regional lymph nodes examined by the pathologist that were found to contain metastases
Survival Months	Months survived after diagnosis
Status	Patient status at the last follow-up

Table 4: Feature Descriptions

References

[1] Marc Fortó Arnau Torruella. Github Repository. https://github.com/marcforto14/breast_cancer_survival, 2024.

[2] NCI. Breast Cancer Dataset. https://www.kaggle.com/datasets/reihanenamdari/breast-cancer, 2010.

[3] Penn Medicine. Stages grades. https://www.pennmedicine.org/cancer/types-of-cancer/breast-cancer/breast-cancer-staging.