DIGITAL COMMONS
@ UNIVERSITY OF SOUTH FLORIDA

University of South Florida
Digital Commons @ University of South Florida

USF Tampa Graduate Theses and Dissertations USF Graduate Theses and Dissertations

March 2024

Utilizing Machine Learning Techniques for Accurate Diagnosis of Breast Cancer and Comprehensive Statistical Analysis of Clinical Data

Myat Ei Ei Phyo
University of South Florida

Follow this and additional works at: https://digitalcommons.usf.edu/etd

Part of the Statistics and Probability Commons

Scholar Commons Citation
Phyo, Myat Ei Ei, "Utilizing Machine Learning Techniques for Accurate Diagnosis of Breast Cancer and Comprehensive Statistical Analysis of Clinical Data" (2024). *USF Tampa Graduate Theses and Dissertations.*
https://digitalcommons.usf.edu/etd/10233

This Thesis is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact digitalcommons@usf.edu.

Utilizing Machine Learning Techniques for Accurate Diagnosis of Breast Cancer and Comprehensive Statistical Analysis of Clinical Data

by

Myat Ei Ei Phyo

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts
Department of Mathematics & Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Jiwoong Kim, Ph.D.
Lu Lu, Ph.D.
Seung-Yeop Lee, Ph.D.

Date of Approval:
March 14, 2024

Keywords: machine learning, breast cancer, decision tree, SVM, naïve Bayes, logistic regression, AdaBoost and Bagging

Copyright © 2024, Myat Ei Ei Phyo

DEDICATION

To my parents for their sacrifices, belief in my abilities, and relentless encouragement to pursue my dreams. To my dear mom, whose strength, wisdom, and love have been my greatest inspiration. To Ko Ko, for always being a pillar of support and a source of wisdom and guidance. And to my extended family, for their continuous encouragement and belief in my potential. This thesis is dedicated to all of you, with heartfelt gratitude and love.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor Dr. Jiwoong Kim, for his unwavering support, guidance, and mentorship through the journey of completing this thesis. As an international student, navigating the complexities of academic research in a foreign environment presented numerous challenges, but Dr. Kim was always there to offer valuable insights and encouragement. He not only provided academic advice but also took the time to understand my perspective and unique challenges. His patience, understanding, and willingness to listen have been in shaping my academic growth and development.

I extend my sincere appreciation to Dr. Lu Lu and Dr. Seung-Yeop Lee, the members of the thesis committee, for their valuable feedback, insightful suggestions, and contributions to this thesis. I am truly grateful for their expertise and dedication, which have greatly enriched the quality and depth of my thesis.

I am profoundly thankful for my family and friends' support, encouragement, and understanding throughout this challenging yet rewarding journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Breast cancer represents a formidable malignancy, presenting a substantial threat to global health and individual well-being. Conventionally, it is widely held that the prognosis for breast cancer patients hinges predominantly upon the timing of diagnosis and the extent of cancer progression, typically delineated by its stage. However, emerging evidence from robust regression and machine learning analyses challenges this prevailing notion. The results indicate that survival months cannot be solely attributed to diagnosis and socio-economic factors. Instead, additional variables such as existing diseases and treatment complexities may contribute to the intricate landscape of breast cancer outcomes.

This research aims to delve into the factors that influence the survival status of breast cancer patients beyond the traditional understanding. By harnessing the power of advanced regression and machine learning techniques, this study explores the complex interplay of various variables that may impact a patient's survival status. The underlying premise of this research is rooted in addressing a critical inquiry that profoundly impacts cancer patients: "What factor influences survival outcome status?" Understanding the factors that shape survival status is of paramount importance to individuals battling breast cancer. The results of this study hold substantial relevance for clinical practice and patient care, particularly within machine learning methodologies. By acknowledging the complex nuances of survival outcomes, healthcare practitioners can embrace a comprehensive approach to treatment and patient oversight. The results highlight that survival status outcomes cannot be solely attributed to diagnosis and socio-economic factors but necessitate a thorough assessment of individualized factors. Coexisting

diseases and treatment intricacies may significantly influence the prognosis, diagnosis, and overall survival of the patients.

Using survival status as a central hypothesis reflects cancer patients' pressing concerns and uncertainties. By comprehensively exploring the factors that underpin survival months, this research aims to provide valuable insights that empower healthcare providers to deliver personalized care and support. Recognizing the complexity of breast cancer outcomes will enable clinicians to tailor treatment plans, consider individualized variables, and address the unique needs and concerns of patients. Ultimately, this study endeavors to enhance the understanding of breast cancer prognosis, optimize patient outcomes, and improve the quality of care for patients navigating the challenging journey of breast cancer.

This study presents an analysis of breast cancer data employing various machine learning algorithms to predict survival status. The dataset encompasses information crucial for diagnosis, including whether cancer is present or absent, classified as malignant or benign. Notably, survival status, typically associated with post-treatment prognosis, may not directly align with the context of the initial diagnosis. Longitudinal studies are essential for understanding survival outcomes dynamically, capturing the effects of treatment changes, disease progression, and time-dependent factors. By analyzing individual patient trajectories, previously undetectable patterns emerge, offering insights into breast cancer evolution and its impact on survival.

This study evaluated various classification algorithms, including logistic regression, decision trees, naive Bayes, support vector machine (SVM), AdaBoost, and bagging, to predict survival status. While logistic regression remains a conventional statistical tool, its limitations in capturing complex relationships in survival prediction were evident. In contrast, machine learning algorithms demonstrated advantages, particularly in handling nonlinear relationships and imbalanced datasets. Advanced algorithms, such as bagging, performed better in accurately predicting breast cancer survival status, surpassing logistic regression and other methods. This reinforces the critical role of advanced machine learning techniques in enhancing the accuracy and reliability of breast cancer survival prediction models.

CHAPTER ONE: INTRODUCTION

1.1 Background

In the entire world, breast cancer is indeed one of the leading causes of death for women. Breast cancer is the most prevalent type of cancer in women, with roughly 2.3 million new diagnoses in 2020 alone, according to the World Health Organization (WHO). Breast cancer has become more common in recent decades, having a substantial impact on the health and well-being of individuals and their families. Given the tremendous impact of breast cancer on individuals and society, there is a rising need for effective methods for the disease's early detection, diagnosis, and treatment. The availability of breast cancer data, including genetic and clinical data, has risen dramatically in recent years, opening new avenues for better understanding the illness and designing more effective interventions.

The goal of this research is to create a model for the prediction of breast cancer outcomes based on data such as age, marital status, cancer stage, survival or death, and months alive. This approach can be used to identify high-risk patients and guide treatment decisions. A thesis can specifically try to investigate the association between factors and breast cancer outcomes. Develop and test multiple machine learning models for overall survival and progression-free survival. In this study, the author rigorously examined the way to predict breast cancer outcomes using the aforementioned machine learning models. These models will be used to identify the variables that have the most significant impacts on breast cancer outcomes and to investigate potential interactions between these variables.

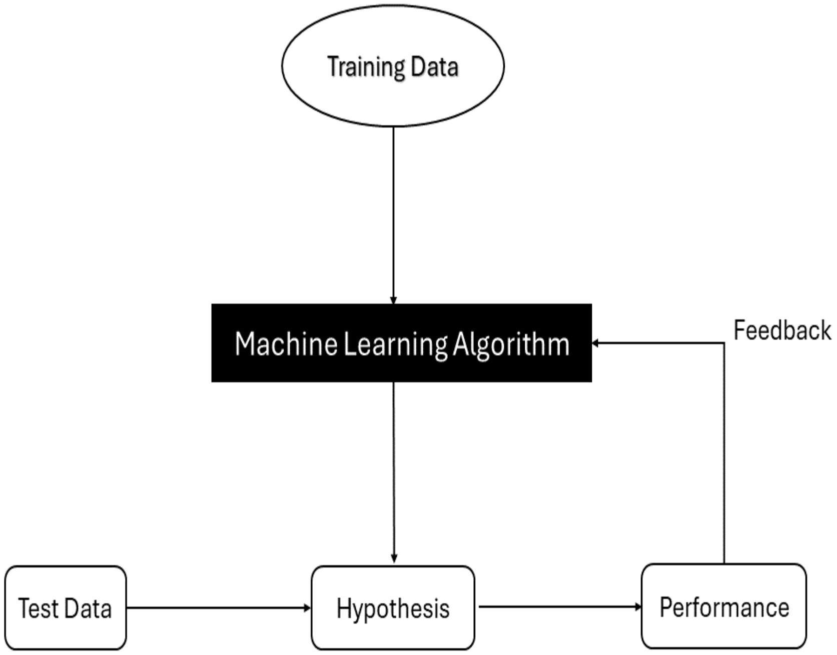


Figure 1.1

Unveiling Machine Learning Mechanisms of Comprehensive Understanding

This study also compares the performance of different models and assesses their generalizability to new patient populations. This evaluation will help determine which model provides the most accurate predictions and is suitable for broader use in clinical practice. Analyzing the results, it is tried to provide insights and recommendations for clinical practice. They guide individualized treatment decisions based on patient characteristics and expected outcomes. This research can provide valuable information about breast cancer treatment and improve patient care by tailoring treatment plans to each person's unique circumstances. This study uses existing breast cancer data to develop predictive models to improve treatment outcomes and inform clinical decision-making.

1.2 Literature Review

Note: In this section, we conduct a thorough review of existing literature on the prediction of breast cancer using machine learning methods. Our goal is to offer a comprehensive analysis that includes insightful perspectives, critical evaluation of methodologies, and synthesis of connections among studies.

Prediction of Breast Cancer Using Machine Learning Approaches: A Review of the Literature

By Reza Rabiei, Seyed Mohammad Ayyoubzadeh, Solmaz Sohrabei, Marzieh Esmaeili, and Alireza Atashi

Breast cancer stands as a pervasive health concern affecting women globally. Its multifaceted etiology, influenced by clinical, lifestyle, social, and economic factors, demands innovative approaches for early detection and effective management. In response to this pressing need, machine learning techniques have emerged as a promising avenue, offering the potential to unveil hidden patterns within complex datasets. The article titled "Prediction of Breast Cancer using Machine Learning Approaches," authored by Reza Rabiei, Seyed Mohammad Ayyoubzadeh, Solmaz Sohrabei, Marzieh Esmaeili, and Alireza Atashi, contributes significantly to this discourse by investigating the predictive capabilities of diverse machine-learning methods, while incorporating demographic, laboratory, and mammographic data.

In the wake of the digital age, where data-driven insights have transformed various industries, healthcare remains ripe for such advancements. Breast cancer, with its wide-ranging risk factors, necessitates a nuanced approach to prediction. The authors' pursuit of harnessing machine learning to predict breast cancer encompasses an ambitious goal - to leverage a trifecta of factors: demographic attributes, laboratory measurements, and mammography features. The synthesis of these elements allows for a more comprehensive analysis, potentially revealing subtle interplays that could otherwise remain concealed. Central to the study's foundation is an extensive database housing over 5,000 independent records. Among these, a quarter corresponds to breast cancer patients. This corpus of data, containing 24 attributes per record, serves as the bedrock for the authors' investigation. The study orchestrates a symphony of machine-learning methodologies, each with its intricacies and strengths. Random forest, a popular ensemble learning technique, makes its debut alongside neural networks, gradient-boosting trees, and genetic algorithms, each poised to decode the data's complexity.

The methodological approach is meticulously devised. The authors opt for an incremental unveiling of insights. Initially, the models are trained solely on demographic and laboratory attributes, thereby extracting the potential influence of these dimensions on predictive accuracy. The subsequent phase widens the scope, encompassing all three dimensions - demographic, laboratory, and mammographic. This intricate choreography of model training serves a dual purpose: first, it showcases the relative impact of mammographic features, a facet critical in modern breast cancer detection; second, it emphasizes the intricate relationships between variables, necessitating a holistic approach to predictive modeling. The results of the study are both revealing and instructive. Among the array of techniques employed, the random forest emerges as the proverbial torchbearer, boasting a remarkable accuracy of 80%. This result is accompanied by a sensitivity of 95% and a specificity of 80%, underscoring the model's robustness in identifying true positives and avoiding false positives. The area under the curve (AUC) 0.56 further reinforces the model's efficacy. Notably, gradient boosting, with an AUC of 0.59, exhibits a compelling performance, overshadowing the neural network, a testament to the variability in algorithm performance across different datasets.

The study's implications are profound. It underscores the pivotal role of a multifaceted approach to predictive modeling. By amalgamating demographic, laboratory, and mammographic attributes, the study embraces a holistic vantage point, one that mirrors the intricate nature of breast cancer causation. Such comprehensive modeling holds the potential for accurate predictions and offers a platform for devising tailored care plans that are pivotal in early diagnosis. Moreover, the article's call for data-driven intelligence resonates with the contemporary healthcare landscape. The digitization of medical records, combined with the exponential growth of data, presents an opportunity to refine predictive models continually. The authors highlight the significance of data collection, storage, and management in the modern healthcare paradigm. These prerequisites pave the way for intelligent systems that amalgamate diverse risk factors, enhancing predictive accuracy and efficacy.

In conclusion, the article "Prediction of Breast Cancer using Machine Learning Approaches" navigates the complex realm of breast cancer prediction with a deft blend of machine-learning methodologies and comprehensive data analysis. The authors' approach, exemplified by the incremental unveiling of insights, offers a roadmap for future endeavors in this domain. By accentuating the significance of multidimensional modeling, the study echoes the complex interplay of factors inherent to breast cancer etiology. The findings emphasize the potential of machine learning to revolutionize early detection and disease management while advocating for a data-driven, multidisciplinary approach that aligns with the evolving healthcare landscape. This article's contribution stands as a beacon in the ongoing quest to enhance breast cancer prediction and underscores the transformative potential of machine learning in healthcare.

1.3 Machine Learning Methods

Machine learning methods such as naive Bayes, decision trees, AdaBoost and Bagging, support vector machines (SVM), and logistic regression are commonly used to analyze breast cancer data because they can handle large datasets and detect complex patterns and relationships. These methods are chosen for their performance, interpretability, and ability to handle different data types, such as categorical and continuous variables. Moreover, these methods have been widely used in previous breast cancer studies, making them a reliable option for data analysis.

Decision tree is a widely used machine learning technique for analyzing breast cancer data due to its ability and interpretability to handle both categorical and continuous variables. Naive Bayes is a pioneering machine learning method that can effectively handle high-dimensional data and is widely used for breast cancer diagnosis and classification. AdaBoost and Bagging are ensemble learning techniques that can improve the accuracy and strength of predictive models. SVM is a powerful machine learning technique that can handle complex and nonlinear relationships in data, making them well-suited for analyzing breast cancer data. Logistic regression is a widely used method for solving binary classification problems and effectively predicts breast cancer outcomes. These methods were selected for this study due to their proven efficacy in breast cancer research and their ability to handle different data types and relationships between them.

The dataset used in this study includes multiple variables such as age, race, marital status, T stage for tumor size, N stage for degree of lymph node involvement, Stage 6 for overall cancer stage according to the 6th edition American Joint Committee on Cancer (AJCC), Differentiation for degree of cancer differentiation, Grade for tumor type, A stage for comprehensive cancer stage according to AJCC 8th edition, tumor size, estrogen status, progesterone status, regional nodes examined, regional nodes positive, survival months and status. This dataset contains information about breast cancer patients and various factors related to their prognosis and diagnosis. These variables are essential in predicting breast cancer outcomes and are commonly used in breast cancer research.

Table 1

Comparison of Clinical Characteristics Between Alive and Dead Groups with corresponding p-Values

Status	Alive (N=3,408)	Dead (N=616)	Overall (N=4,024)	P -Value
Age (years)				
Mean (SD)	53.8 (8.81)	55.2 (9.70)	54.0 (8.95)	0.01
Median (Min-Max)	54.0 (30-69)	56.5 (30-69)	54.0 (30-59)	0.5
T Stage				
T1	1,446	157	1,603	0.43
T2	1,483	303	1,786	0.37
T3	417	116	533	0.33
T4	62	40	102	0.14
N Stage				
N1	2,462	270	2,732	0.43
N2	655	165	820	0.34
N3	291	181	412	0.15
Tumor Size (mm)				
Mean (SD)	29.3 (20.3)	37.1 (24.1)	30.5 (21.1)	0.07
Median (Min-Max)	23 (1–140)	30(1–140)	25 (1-140)	0.5
A Stage				
Distant	57	35	92	0.15
Regional	3,351	581	3,932	0.39
Estrogen Status				
Negative	161	108	269	0.12
Positive	3,247	508	3,755	0.4
Progesterone Status				
Negative	494	204	698	0.25
Positive	2,914	412	3,326	0.41

Note. The analysis of a cohort comprising 3,408 individuals (Alive) and 616 individuals (Dead) revealed significant age differences (mean 53.8 vs. 55.2 years, P = 0.01). Tumor and lymph node stages showed no significant variations. Tumor size exhibited a non-significant trend (P = 0.07). Anatomical stage (A stage), estrogen status, and progesterone status did not differ significantly between the alive and dead groups. These findings underscore the importance of age in prognosis within the cohort.

Table 1 presents the descriptive statistics for the ages of the patients in three different groups: Alive and dead. In terms of the median age, both the "Alive" and "Overall" groups have a median age of 54 years. This indicates that half of the individuals in these groups are 54 years old or younger, while the other half are 54 years old or older. On the other hand, the "Dead" group has a median age of 56.5 years. This implies that half of the individuals who died were 56.5 years old or younger, while the remaining half were 56.5 years old or older.

The T stage indicates the size and extent of the primary tumor and represents mm as the tumor size in the table. The majority of patients had T2-stage tumors (44.4%), followed by T1-stage tumors (39.8%) and T3-stage tumors (13.2%). A smaller proportion of patients had T4-stage tumors (2.5%). The N stage reflects the involvement of lymph nodes and represents the number of patients who have the N stage in Table 2. Most patients had N1 stage (67.9%), followed by N2 (20.4%) and N3 (11.7%). The tumor size ranged from 1 to 140 mm, with a mean length of 30.5 mm and a standard deviation of 21.1 mm. The median tumor size was 25.0 mm. Regarding the A stage, a small proportion of patients had distant metastasis (2.3%), while most had regional metastasis (97.7%).

Finally, it provides information on the patients' hormone receptor status. The majority of patients had estrogen-positive tumors, 93.3%, while a smaller proportion had estrogen-negative tumors, 6.7%. Similarly, most patients had progesterone-positive tumors, 82.7%, while a smaller proportion had progesterone-negative tumors, 17.3%. Overall, this demographic and survival characteristic table provides essential information on the characteristics of patients with breast cancer, which offers invaluable insights in guiding clinical decision-making and further research endeavors.

Table 2

Demographic and Survival Characteristics by Race

Race	Black (N=291)	Other (N=320)	White (N=3413)	Overall (N=4024)	P-Value
Ages					
Mean (SD)	52.6 (9.05)	51.4 (9.54)	54.3 (8.85)	54 (8.95)	<0.01
Median (Min-Max)	53 (31-69)	51 (30-59)	55 (30-59)	54 (30-69)	<0.01
Survival Months					
Mean (SD)	66.6(24.8)	73.2 (23.1)	71.5 (22.7)	71.3 (22.9)	<0.01
Median (Min-Max)	57 (4-107)	77 (1-107)	73 (2-107)	73 (1-107)	<0.01
Status					
Alive	218	287	2,903	3,408	0.33
Dead	73	33	510	615	0.31
Marital Status					
Divorced	40	29	417	486	0.33
Married	113	237	2,293	2,543	0.34
Separated	8	4	33	45	0.33
Single	102	33	480	615	0.28
Windowed	28	17	190	235	0.30

Table 2 presents descriptive statistics for a sample of 4,024 individuals, including 291 Black, 320 Other, and 3,413 White. This table reports the mean and standard deviation (SD) for age and survival months and each variable's median, minimum, and maximum values. The mean age was highest among Whites, 54.3 years, followed by the overall sample, 54 years, and 52.6 years for Black individuals. The mean survival months were highest among other individuals (73.2 months), followed by White individuals (71.5 months) and Black individuals (66.6 months). In terms of status, many individuals in the sample were alive at the time of data collection 84.7%, with the highest percentage of survivors among White individuals at 85.1% and the lowest rate among Black individuals at 74.9%. The overall percentage of deceased individuals was 15.3%, with the highest percentage of deaths among White individuals being 14.9% and the lowest among other individuals (10.3%).

Regarding marital status, most individuals in the sample were married (65.7%), with the highest percentage of married individuals among other individuals being 74.1%, and the lowest among Black individuals was 38.8%. The overall rate of divorced individuals was 12.1%, with the highest percentage of divorced individuals among White individuals at 12.2% and the lowest among other individuals at 9.1%. The rate of single individuals was highest among Black individuals at 35%, followed by White individuals at was14.1%, and Other individuals at 10.3%. The percentage of separated and widowed individuals was lowest among Other individuals at 1.3% and 5.3%, respectively, and highest among White individuals at 1.0% and 5.6% each. This table provides a comprehensive overview of the sample's demographics, including age, survival status, and marital status, disaggregated by race. It can help identify potential disparities in health outcomes and social factors among different racial groups.

Table 3 presents statistical summaries for various clinical and demographic characteristics of a group of 4,024 patients with different stages of cancer. The patients were categorized based on tumor stage (T-stage), lymph node stage (N-stage), and anaplastic grade IV status. The mean age of the patients was 54 years, with a standard deviation of 8.96 years. The age distribution was relatively consistent across all T-stages, N-stages, and anaplastic grade IV status groups. However, the anaplastic grade IV group had a slightly higher mean age of 52.3 years and a more significant standard deviation of 10.8 years, suggesting more variability in age within this group.

Table 3

Correlation of Tumor Grade, Cancer Stage, and Clinical Parameters with Survival Outcomes

Grade	1 (N=543)	2(N=2,351)	3(N=1,111)	Grade IV (N=19)	Overall (N=4,024)	P-Value
Mean (SD)	55.3 (6.37)	54.43 (8.83)	2.6 (9.34)	52.3 (10.8)	54 (8.96)	<0.01
Median (Min-Max)	55 (32-69)	55 (30-69)	53 (30-69)	52 (37-69)	54 (30-69)	<0.01
T Stage						
T1	219	975	344	5	1,603	0.15
T2	200	1,025	556	5	1,786	0.14
T3	55	303	168	7	533	0.14
T4	9	48	43	2	102	0.12
N Stage						
N1	436	1,635	651	10	732	0.14
N2	74	489	253	4	820	0.15
N3	33	227	207	5	472	0.13
Tumor Size						
Mean (SD)	26.4 (20.8)	29.7 (20.4)	33.8(22.1)	44.2 (25.6)	30.5 (21.1)	0.03
Median (Min-Max)	20 (2-100)	24 (1-140)	27(1-140)	40 (13-100)	25 (1-140)	<0.01
Survival Month						
Mean (SD)	72.9 (21.0)	72.2 (22.2)	68.7(24.9)	64.4 (32.7)	71.3 (22.9)	<0.01
Median (Min, Max)	74 (5-107)	73 (1-107)	70 (2-107)	75 (9-102)	75(1-103)	<0.01

Tumor size was highest in the anaplastic grade IV group, with a mean of 44.2 mm and a median of 40 mm. In contrast, T1 tumors had the smallest mean size of 26.4 mm and a median of 20 mm. The tumor size distribution was positively skewed for all T-stages, N-stages, and anaplastic grade IV status groups, with a few extreme values that skewed the mean. The mean survival time was 71.3 months, with a standard deviation of 22.9 months. The anaplastic grade IV group had the lowest mean survival time of 64.4 months, while the other groups had mean survival times ranging from 68.7 to 72.9 months. The median survival time for all groups was around 73 months, except for the T1 group, which had a slightly higher median survival time of 74 months.

Overall, the N-stage had the most significant effect on patient outcomes, with patients in the N3 stage having the lowest mean and median survival times and the highest tumor size. T-stage and anaplastic grade IV status had a more minor but still significant effect on tumor size and survival time.

1.3.1 Decision Tree

Decision tree is a trendy nonparametric machine learning algorithm that can be used for both classification and regression tasks. A supervised learning algorithm builds a tree-like model of decisions and their consequences. Decision tree algorithms are popular in academia and industry because they are easy to understand and interpret. In the decision tree, each inner node represents a test for an attribute, each branch represents the test result, and each leaf node represents a class label or number.

A decision tree is created by partitioning the data according to given attribute values, which are then iteratively partitioned into subsets until the stopping condition is met. The stopping criteria can be based on factors such as tree depth, the number of instances in leaf nodes, or subset impurities. Various metrics, such as precision and misclassification error, can measure subset contamination, but the analyzing its impact on Gini index can provide insights into its effects, particularly in the decision tree where the Gini index is used as a measure the impurity to guide the splitting of nodes.

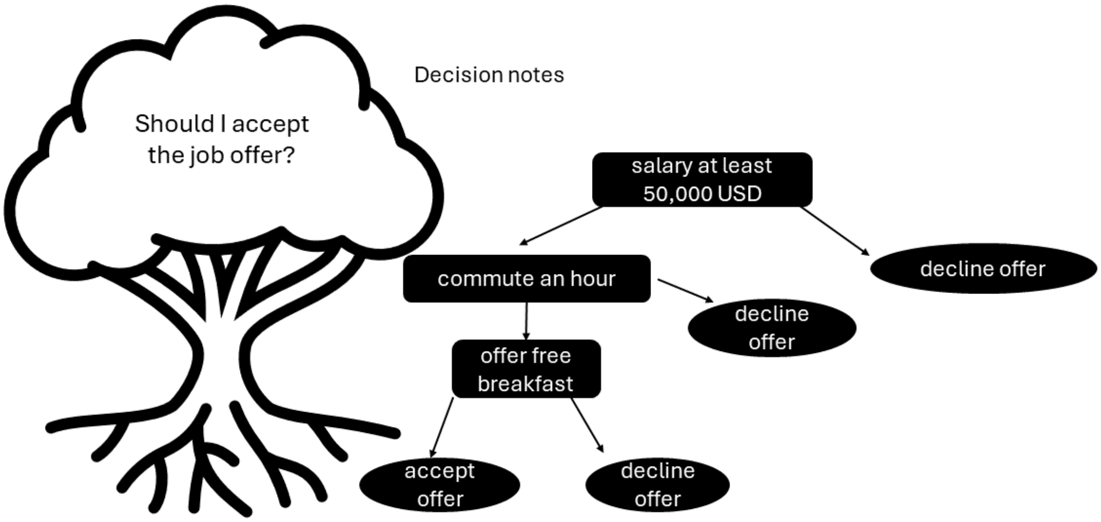


Figure 1.2

Concept of Decision Tree

Mathematically, Gini Index is expressed as:

$$Gini = 1 - \sum_{i=1}^{j} p_i{}^2$$

where p_i is the probability of an object being classified to a particular class.

Entropy measures the level of randomness or disorder within a subset, and the Gini index measures the probability of misclassifying a random instance within a subgroup. The decision tree algorithm aims to minimize subset impurities by choosing attributes that maximize the separation of the cases into their respective classes. A comprehensive breakdown of the Gini Index and its role in decision trees, including its computational aspects follows.

The Gini index $Gini(B)$, for a data set B with J classes is calculated as:

$$Gini(B) = 1 - \sum_{j=1}^{J}(p_j)^2,$$

where p_j is the probability of an instance being classified as class j within dataset B.

At the each node of decision tree, the algorithm evaluates the potential splits based on the different attributes. For example, a given attribute A with n possible values $A_1, A_2, \dots, A_n$, the algorithm calculates the Gini index for each split, denotes as $Gini_{split}(B, A)$. The Gini index for split A is computed as a weighted average of the Gini index for each resulting sub note B_i:

$$Gini_{split}(B, A) = \sum_{i=1}^{n} \frac{|B_i|}{|B|} \cdot Gini(B_i),$$

where$|B_i|$ is the number of instances in sub note B_i, and $|B|$ is the total number of instances in the root node before the split.

The decision tree algorithm selects the attribute A and corresponding split point that minimizes the Gini Index for the resulting sub notes. Mathematically, this can be presented as:

$$A^* = \arg min_A((Gini_{split}(B, A))$$

where A^* is the optimal attribute selected for the split.

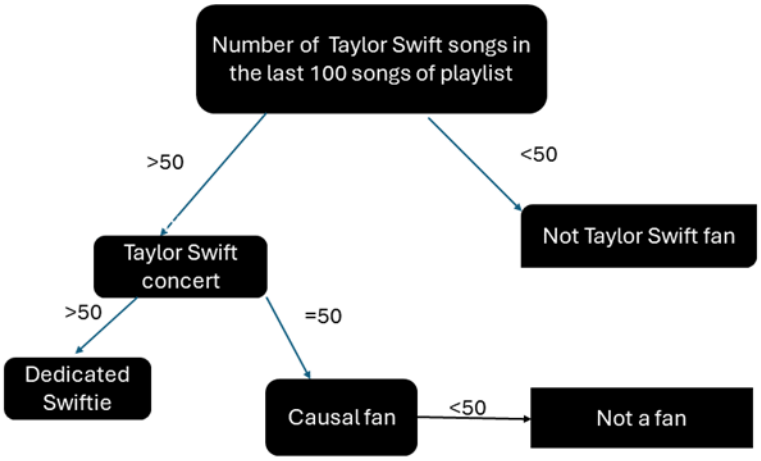


Figure 1.3

Decision Tree to Classify Taylor Swift Fans

The benefit of a decision tree is that it can handle both categorical and numerical attributes. Categorical attributes are easily handled by dividing the data into subsets based on the attribute's possible values. Numeric attributes can be discretized into categories or divided into intervals based on specific criteria, such as information retrieval or variance reduction. Another benefit of a decision tree is that it can handle missing values by assigning the most common or mean value of the attributes in the subset. However, this approach may introduce bias if the values are not missed by chance. A weakness of a decision tree is that it tends to overfit, especially if the tree is too deep or the data contains irrelevant attributes. Overfitting occurs when the tree captures noise in the data instead of the underlying pattern. However, various techniques can be used to avoid overfitting. Pre-pruning and post-pruning are the main techniques for overfitting. Pre-pruning in construction of a decision tree involves stopping the tree's growth before it reaches its maximum depth or the minimum number of samples required for a split, preventing overfitting. In contrast, post-pruning, also known as pruning or cutting back, entails trimming branches of an already-grown decision tree based on various criteria to improve its generalization performance. The post-pruning generates a tree with fewer branches than would otherwise generate a complete tree and then removes parts. Pruning is the data compression technique in data mining that reduces the size of the decision tree by removing the sections of the tree that are non-critical to classify the instances.

Tree pruning involves removing some branch or leaf nodes that do not improve the tree's accuracy concerning validation data. Limiting the tree depth reduces the model complexity and prevents overfitting. Ensemble techniques such as bagging, boosting, and random forest can combine multiple decision trees to improve model accuracy and robustness. There are two differences in the performance of random forest and gradient boosting. That is, the random forest can build each tree individually. In contrast, gradient boosting can build one tree one by one, so the performance of the random forest is not very comparable to slope elevation. Regularization techniques such as L1 (Lasso) or L2 (Ridge) regularization can add a penalty term to the decision tree algorithm to prevent overfitting. They aim to counter the overfitting model by lowering the variance while increasing some of the bias. The L1 and L2 regularizations work by adding a penalty term to the model's cost function. This penalty term is based on the magnitude of the model's coefficients. Regularization aims to balance reducing variance (which leads to overfitting) and increasing bias (which makes the model less flexible). In the L1 regularization, the penalty term is the absolute value of the model's coefficients multiplied by a regularization parameter (lambda). It encourages the model to set some coefficients to precisely zero, effectively performing feature selection. The L1 regularization prevents overfitting and helps identify and exclude less essential features from the model. In the L2 regularization, the penalty term is the square of the model's coefficients multiplied by the regularization parameter. Unlike L1, the L2 regularization does not force coefficients to become exactly zero but penalizes significant coefficients. It helps reduce the impact of individual features without excluding them entirely, leading to a more stable model.

In construction of a decision tree, the tree pruning involves selectively eliminating non-contributory branches or leaf nodes, aiming to enhance the accuracy of the tree concerning validation data. By limiting the depth of the tree, its overall complexity is reduced to prevent overfitting. Ensemble techniques like random forest and gradient boosting play pivotal roles in aggregating multiple decision trees, each exhibiting distinct construction approaches. Random forest constructs individual trees independently while gradient boosting incrementally builds one tree at a time, resulting in differing performance characteristics. Concurrently, regularization techniques such as L1 and L2 mitigate overfitting in decision tree algorithms. The L1 regularization facilitates feature selection by introducing a penalty term based on the absolute values of the model's coefficients, effectively identifying and excluding less pertinent features. On the other hand, the L2 regularization, employing a penalty term based on the square of coefficients, focuses on reducing the impact of individual features without entirely excluding them. These methodologies collectively strive to balance minimizing variance (associated with overfitting) and increasing bias, ultimately enhancing decision tree models' overall robustness and performance.

A decision tree has been used a lot in health care: it has been popular to diagnose diseases, predict treatment outcomes, or recommend treatment. It can also have useful real-world applications in many other subfields of science. For example, a decision tree can detect fraud, assess credit risk, or predict market trends in finance. Engineering can use it to optimize processes, detect anomalies, and predict failures. In summary, a decision tree is a powerful and versatile machine-learning algorithm that can be used for various tasks and domains. Its simplicity, interpretability, and flexibility are popular with machine learning beginners and experts.

1.3.2 Naive Bayes

The naive Bayes method is a classification algorithm commonly used in learning applications. It is based on Bayes' theorem, mathematically expressed as

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where

- A and B are events with $P(B) \neq 0$,
- $P(A|B)$ is the posterior probability of A given B,
- $P(B|A)$ is the likelihood of A given a fixed B,
- $P(A)$ and $P(B)$ are prior probability and marginal probability, respectively.

In classification, a hypothesis is the class designation of an instance, and proof is a set of attributes or features that describe the instance. Given the class labels, the naive Bayes method assumes that the attributes are conditionally independent, which simplifies computing the likelihood of a hypothesis given to evidence.

In the classification, a hypothesis constitutes the designated class of an instance, while proof encapsulates the set of attributes or features delineating said instance. Given the assigned class labels, the naive Bayes method predicates its efficacy on the conditional independence assumption among these attributes. This foundational assumption streamlines the computation of the likelihood of a hypothesis predicated on the provided evidence. To enhance precision, "proof" in this context can be defined as the distinct set of features characterizing an instance, and "class labels" denote the assigned categories or classes within which instances are categorized.

Illustrating the computational aspect, consider a pragmatic example involving email classification as "spam" or "not spam" where features encompass words or phrases in the emails. The hypothesis pertains to classifying a new email as either spam or not spam with the proof comprising the specific words or phrases in the email. The likelihood computation involves determining the probability of observing the given set of features (words or phrases) about the assigned class label (spam or not spam). Leveraging the naive Bayes assumption of conditional independence, given the class label, this probability is delineated as the product of individual probabilities for each feature. This methodological simplification optimizes computational efficiency.

The naive Bayes method is called "naive" because it strongly assumes conditional independence between attributes called the probabilistic model: more detailed explanation for being naive is given in next two pages. Still, in practice, this is not always the case. However, this simplistic assumption makes the method work well even with limited training data and applies to various classification problems. The naive Bayesian methods can be used for binary or multiclass classification problems. In the case of binary classification, there are two possible classifications for the class, and the probability of each classification is calculated from the evidence using Bayes' theorem. For instance, the label with the highest posterior probability is chosen as the expected class label. In the case of multiclass classification, where there are more than two possible class classifications, Bayes' theorem calculates the probabilities of all labels given the evidence.

The naive Bayes method optimizes the probability $p(C_k|x_1, \dots, x_n)$ for each of the possible outcomes, or classes C_k, given a problem instance represented by a value of the future $x = (x_1, \dots, x_n)$. The joint probability model can be expressed using the chain rule for conditional probability, which leads to a product of conditional probabilities:

$$p(C_k, x_1, \dots, x_n) = p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

Under the naive conditional independence assumption, where features are assumed to be mutually independent given the class C_k, the joint model can be simplified to

$$p(C_k|x_1, \dots, x_n) \propto p(C_k) \prod_{i=1}^{n} p(x_i|C_k),$$

where the symbol $\propto$ denotes proportionality. Thus, the independence assumptions implies conditional distribution will be

$$p(C_k|x_1, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k),$$

where $Z = p(x) = \sum_k p(C_k) \cdot p(x|C_k)$ is a scaling factor dependent only on $x_1, \dots, x_n$.

The ongoing discourse has solidified the framework of the independent feature model, recognized as the naive Bayes probability model. Incorporated within the naive Bayes classifier, this model converges with a decision rule. A prevalent decision strategy entails choosing the hypothesis that maximizes the probability, thus mitigating the likelihood of misclassification. The Bayes classifier assigns a class label $\hat{y} = C_k$ for some k can be written as follows:

$$\hat{y} = \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

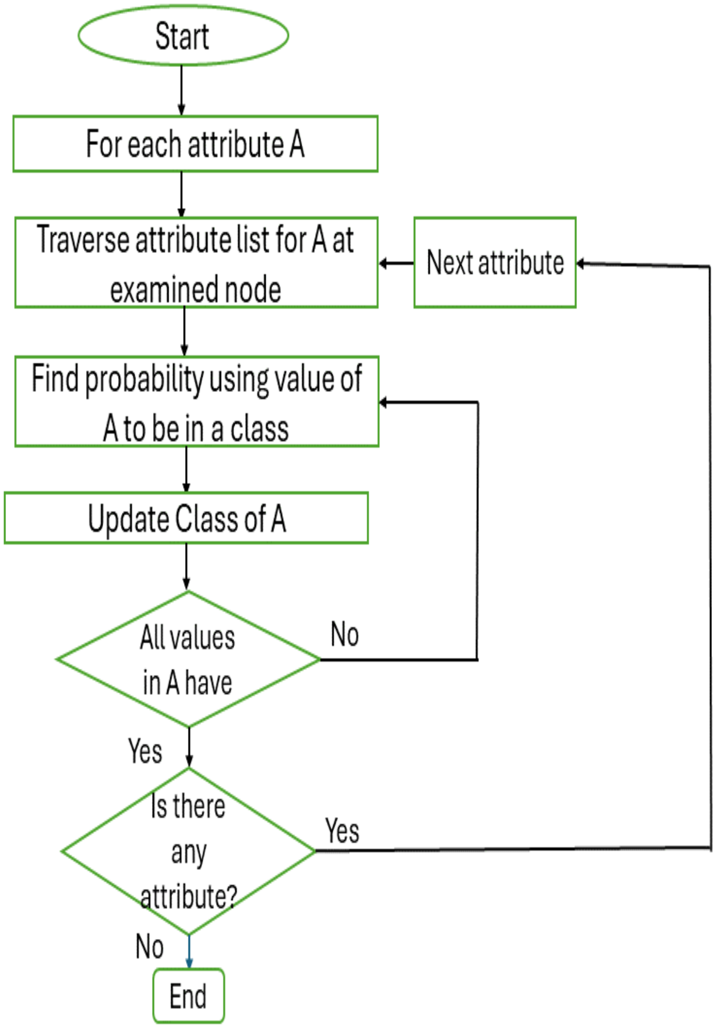


Figure 1.4

Flow Chart of naive Bayes Classification

The naive Bayes method has several variations, depending on the type of probability distribution expected for the attributes. The three most common variants are the Bernoulli method, the polynomial method, and the Gaussian-naive Bayes method. The Bernoulli-naive Bayes method is used when the attribute is binary and indicates the presence or absence of the characteristic. Multinomial Naive Bayes is used when the attributes are discrete and represent the number or frequency of traits. The Gaussian naive Bayes method is used when the attribute is continuous, and a Gaussian or normal distribution can approximate the distribution of the attribute values. These three variants of naïve naïve Bayes classifier are mathematically expressed as follows.

1. Bernoulli Naïve Bayes Classifier:

$$p(x|C_k) = \prod_{i=1}^{n} p_{ki}^{xi}(1 - p_{ki})^{(1-x_i)},$$

where x_i is Boolean expressing the occurrence or absence of the i^{th} term, and p_{ki} is the probability of class C_k generating the term x_i,

2. Multinomial Naïve Bayes Classifier:

$$p(x|C_k) = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} {p_{ki}}^{xi},$$

where $p_{ki} = p(x_i|C_k)$,

3. Gaussian Naïve Bayes Classifier:

$$p(x = v|C_k) = \frac{1}{\sqrt{2\pi{\sigma_k}^2}} e^{-\frac{(v-\mu_k)^2}{2{\sigma_k}^2}},$$

where

v = some observation value,

μ_k= the mean of the value in x associated with class C_k,

σ_k^2= the Bessel corrected variance of the variance in x associated with class C_k.

These variants cater to different types of data and assumptions about feature distributions, making naive Bayes a choice for classification tasks across various domains.

The naive Bayes method is easy to implement and computationally efficient, making it suitable for large data sets and real-time applications. It also works well with high-dimensional data where the number of attributes or features is much larger than the number of instances. However, it can be sensitive to irrelevant or redundant attributes affecting classification accuracy. Also, the training data is assumed to represent the population, and data bias and sampling error can affect classification accuracy.

Despite its limitations, the naive Bayes method is widely used in many applications, such as spam filtering, sentiment analysis, text classification, and medical diagnosis. Spam filtering uses the naive Bayes method to classify the email as spam or non-spam based on email content. Sentiment analysis uses the naive Bayes method to classify text documents as positive or negative based on the sentiment expressed in the text. In medical diagnosis, the naive Bayes method is used to classify patients as healthy or sick based on their symptoms and medical history. In brief, the naive Bayes approach is a versatile and efficient classification algorithm, leveraging Bayes' theorem and assuming conditional independence among attributes. Its simplicity and computational efficiency make it applicable to diverse classification challenges.

1.3.3 SVM

SVM is one of the popular machine learning methods for tackling classification and regression problems. SVM is considered highly accurate and efficient compared to other machine learning methods such as decision trees, random forests, and artificial neural networks. It can be applied in finance, bioinformatics, image classification, and text classification. An example of how

SVM can be used in finance is credit scoring, which classifies and analyzes individuals based on their credit-related features and access to their creditworthiness. SVM can predict whether a borrower is likely to repay a loan. To do this, historical data about borrowers' credit scores, income levels, and loan history can be fed into the SVM algorithm. It learns to classify the data into two categories: those with bad loans and those without loans. It can use this learned model to predict whether a new borrower is likely to default based on credit scores and other relevant information.

SVM can also be used in other disciplines, such as image recognition, natural language processing, and bioinformatics. For example, in image recognition, a support vector machine can classify an image into different categories, such as "cat" or "dog," based on its features. In natural language processing, SVM can classify text documents into various categories, such as "positive" or "negative" emotions. Bioinformatics can be used to classify genes based on their expression patterns.

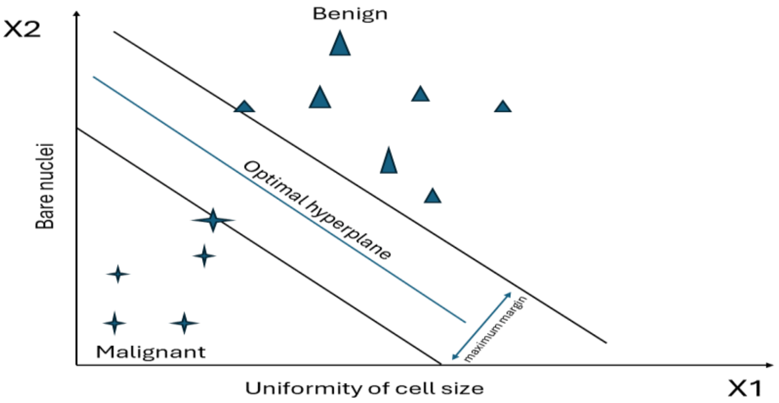


Figure 1.5

Support Vector Classification

Finding the hyperplane that divides the data points into two classes in high-dimensional space is the basis of SVM. An ideal hyperplane maximizes the difference between the two classes. Margin is defined as the distance between the hyperplane and the closest data point for each class. SVM aims to maximize the margin as it is expected to perform better generalization for unseen data. The SVM training process is based on the mathematical optimization problem of finding the hyperplane with the most significant margin. This algorithm takes training data points and assigns each point a class label. These data points are then mapped into a high-dimensional space where a hyperplane is chosen to separate them. The optimal hyperplane is determined by solving an optimization problem involving minimizing the classification error and maximizing the margin. Optimization problems can be formulated as quadratic programming problems - minimization or maximization of objective functions subject to limits, linear equality, and inequality constraints that can be solved using numerical optimization techniques.

An important aspect of SVM is the choice of kernel function used to map the data to the high-dimensional space without explicitly computing the coordinates of the data points. SVM uses kernel functions such as linear, polynomial, and radial basis functions (RBFs). Which kernel functions are chosen depends on the type of data and the specific task.

$$K(x, y) = f(x) \cdot f(y),$$

where $K(x, y)$ is a kernel function, x and y are n-dimensional inputs, and $f(\cdot)$ is a mapping from an n-dimensional space to an m-dimensional space. Examples of the kernel function include linear kernels, polynomial kernels, and Gaussian kernels. The merits of linear kernels are computationally efficient, especially for high-dimensional data, often used as a baseline or starting point for other kernel functions, and sensitive to linear relationships in the data. The demerit is limited expressive power compared to other kernel functions. The polynomial kernel can capture nonlinear relationships between features, model higher-order interactions between features, and be more expressive than linear kernels. However, it requires careful selection of the polynomial degree, which can be time-consuming and computationally expensive and can be sensitive to overfitting if the degree is too high — widely used in many applications due to its flexibility and effectiveness. Gaussian kernels can capture complex nonlinear relationships between features and provide a measure of similarity between data points. The demerits of Gaussian kernels are computationally expensive, especially for large datasets, and can be sensitive to the choice of the kernel bandwidth parameter.

To compute the SVM classifier is equivalent to solving the problem of minimizing the expression of

$$[\frac{1}{n}\sum_{i=1}^{n} \max\,(0,1 - y_i(w^T x_i - b))] + \lambda||w||^2.$$

There are two optimization problems: primal and dual. The primal problem can be written as

$$minimize\ \frac{1}{n}\sum_{i=1}^{n} \varsigma_i + \lambda||w||^2,$$

subject to $y_i(w^T x_i - b) \geq 1 - \varsigma_i,\ \varsigma_i \geq 0$ for all i while the dual problem can be written as

$$maximize\ f(c_1 \dots c_n) = \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i C_i (x_i^T x_j) y_j c_{j,}$$

subject to $\sum_{i=1}^{n} c_i y_i = 0, and\ 0 \leq c_i \leq \frac{1}{2n\lambda}$ for all i.

In addition to binary classification, SVM can be used for multiclass classification and regression tasks. In the pursuit of multiclass classification, the utilization of methodologies such as one-versus-all, in which a single classifier is trained to distinguish one class from the rest of the classes combined, or one-versus-one, in which a separate classifier is trained for every pair of classes that represents strategic approaches for systematically managing the categorization of diverse courses within the framework of the research. In this case, multiple binary classifiers are trained to distinguish each pair of classes. For the regression, SVM aims to find the optimal hyperplane for the data in the continuous output space. One of the main advantages of SVM is its ability to process high-dimensional data with relatively small sample sizes. It is also known to be robust against outliers and performs well on both linearly and nonlinearly separable data. However, it is sensitive to the kernel function and regularization parameter choice, which can affect the model’s generalization performance.

In summary, SVM can handle non-linear decision boundaries using kernel functions. However, it can be sensitive to the choice of kernel function and hyperparameters, and its training time can increase significantly for large datasets. SVM is memory-efficient and computationally fast, making it suitable for large datasets. It is widely used in many fields and has been proven to be a powerful tool for solving various machine-learning problems.

1.3.4 AdaBoost and Bagging

Machine learning models are trained on large datasets to identify patterns and relationships that can help make accurate predictions. However, some datasets may contain complex relationships that are difficult to capture with a single model. In such cases, ensemble learning techniques like AdaBoost and Bagging can be used to improve the accuracy and robustness of the models. Ensemble learning techniques involve combining multiple models to achieve better performance than what can be achieved with a single model. A weak model is a model that performs minimally better than random guessing. The idea is to combine a series of weak models and combine them to create a strong one. The final prediction is obtained by combining the outputs of the weak models using a weighted sum. The weights of the weak models are determined based on their performance on the training data. The better the performance of the weak model, the higher the weight assigned to it.

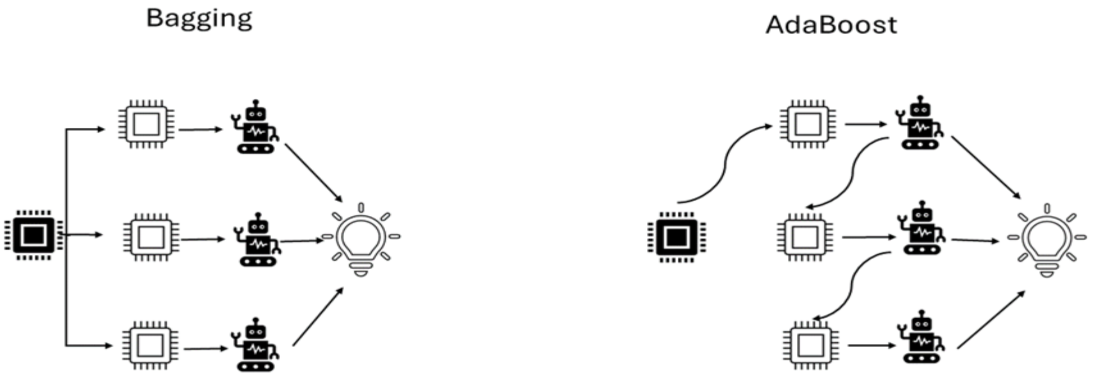


Figure 1.6

Bagging parallel power meets AdaBoost Sequential Elegance

AdaBoost, short for Adaptive Boosting, is a popular machine-learning ensemble method that combines multiple weak models to form a robust model. The main idea behind AdaBoost is to assign weights to the training examples so that misclassified examples are given higher weights. AdaBoost offers several advantages over traditional machine learning methods. First, it enhances accuracy by combining multiple weak models. These weak models capture different aspects of the data, and when combined, they create a robust model that outperforms individual models. Second, AdaBoost demonstrates validity by effectively handling outliers and noise in the data. It accomplishes this by assigning higher weights to misclassified examples, ensuring subsequent models focus on improving their classification. Lastly, AdaBoost is known for its speed, making it suitable for training on large datasets.

AdaBoost classifier is

$$F_T(x) = \sum_{t=1}^{T} f_t(x),$$

where f_t is a weak learner of an object x.

Each weak learner produces the hypothesis h for each sample in the training set.

$$E_t = \sum_i E[F_{t-1}(x_i) + \alpha_t h(x_i)],$$

where $F_{t-1}(x)$ is the boosted classifier, and $h(x)$ is the weak learner that is being considered for addition to the final classifier. The discrete AdaBoost algorithm example can be expressed as follows.

- Samples:$x_{1,} \dots , x_n$,
- outputs: $y_{1,} \dots , y_n, \; y \in \{-1,1\}$,
- initial weights: $w_{n,1}$, set to $\frac{1}{n}$
- error function: $E(f_x), y, i) = e^{-y_i f(x_i)}$
- weak learners $h: x \rightarrow \{-1,1\}$.

For t in 1…T, choose $h_t(x)$ as a weak learner and find it that minimizes.

$$\varepsilon_t = \sum_{i=1}^{n} w_{i,t}.$$

Next, choose

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1-\varepsilon_t}{\varepsilon_t}\right),$$

and add to ensemble as

$$F_t(x) = F_{t-1}(x) + \alpha_t h_t(x).$$

Finally, update the weights as

$$w_{i,t+1} = w_{i,t} e^{-y_i \alpha_t h_t(x_i)} \text{ for } i \text{ in } 1, 2, \ldots, n,$$

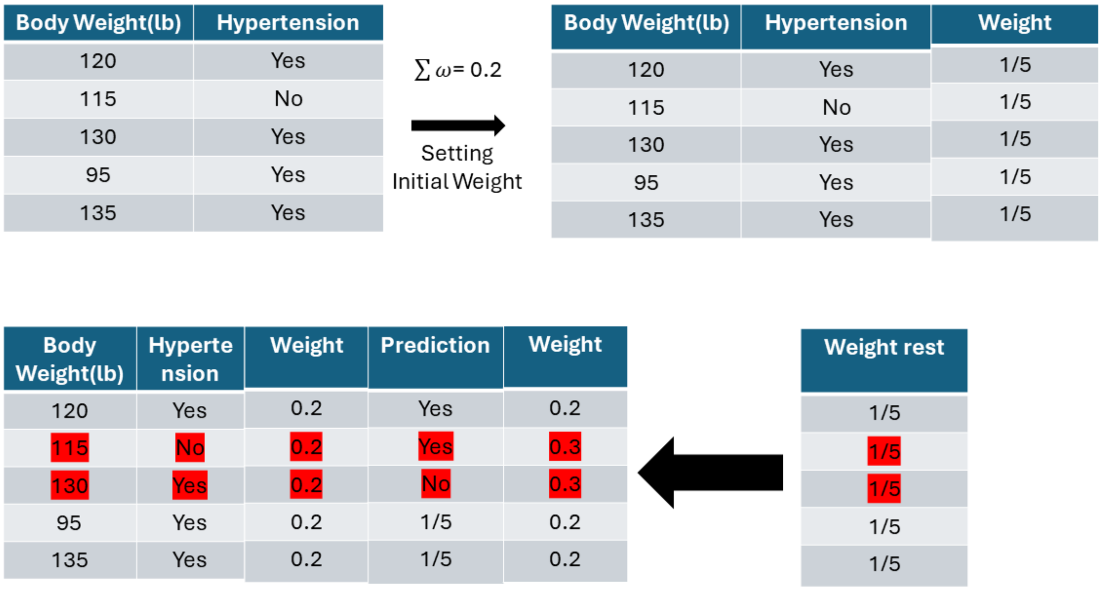
and renormalize the weight such as

$$\textstyle\sum_i w_{i,t+1} = 1.$$

However, AdaBoost does have its limitations. It can be sensitive to outliers in the data, as the higher weights assigned to misclassified examples may lead to overfitting on the training data. Additionally, AdaBoost may struggle with noisy datasets, as the weak models might misclassify a significant number of examples. Furthermore, the use of multiple weak models can make interpretation and explanation challenging. To implement AdaBoost, a series of steps should be followed. Initially, equal weights are assigned to all training examples. Then, a weak model is trained using the training data. The model's performance on the training data is evaluated. Afterward, the weights of the misclassified examples are increased. Finally, another weak model is trained using the updated training data, and the process can be repeated iteratively to enhance the model's performance further.

The AdaBoost algorithm operates in a series of steps to create a strong predictive model. Initially, equal weights are assigned to all training examples, treating them as equally important. Then, a weak model, often referred to as a "base learner," is trained using the training data. This weak model captures certain patterns or characteristics of the data but is not highly accurate on its own. Next, the performance of the weak model is evaluated by assessing its predictions on the training data. The algorithm identifies the examples that were misclassified and assigns them higher weights. By increasing the weights of the misclassified examples, AdaBoost emphasizes the importance of these challenging instances, forcing subsequent weak models to focus on them.

The process is repeated, with another weak model trained on the updated training data, incorporating the adjusted weights. This iterative training continues for a predefined number of models, each building upon the previous ones, refining the overall predictive capability. Finally, to obtain the final prediction, the outputs of the weak models are combined. Typically, a weighted combination or voting scheme is employed, where the models with better performance contribute more to the final prediction. This aggregation of the weak models' predictions results in a strong and accurate model that can provide robust predictions on new, unseen data.



Body Weight(lb)	Hypertension
120	Yes
115	No
130	Yes
95	Yes
135	Yes

$\sum \omega = 0.2$

Setting Initial Weight

Body Weight(lb)	Hypertension	Weight
120	Yes	1/5
115	No	1/5
130	Yes	1/5
95	Yes	1/5
135	Yes	1/5

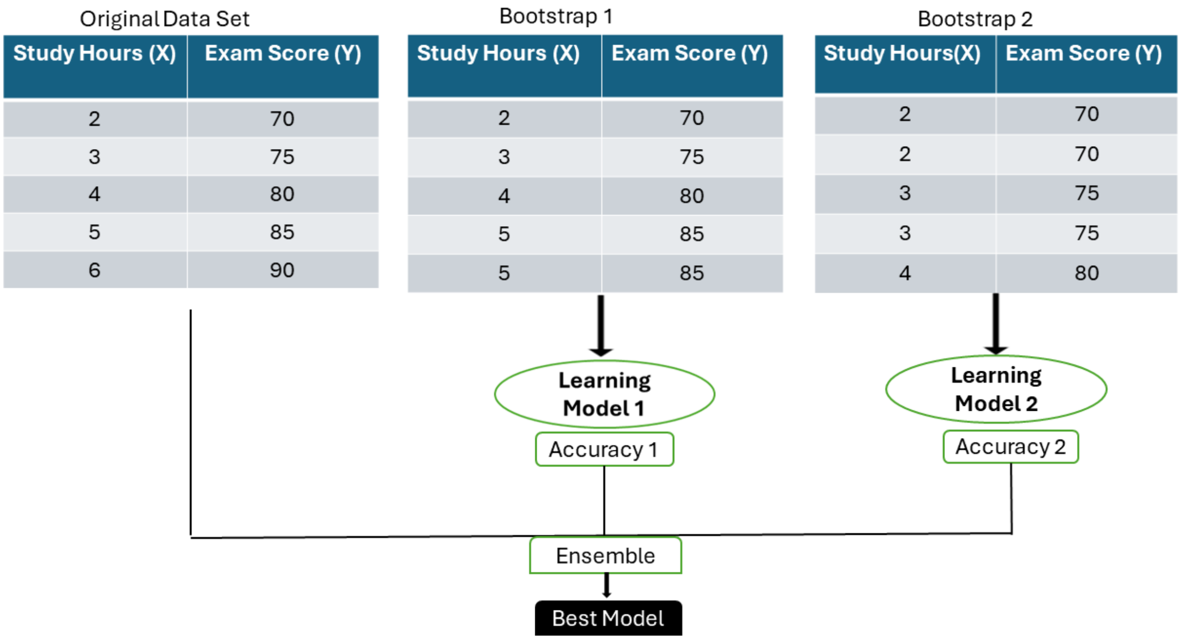
Weight rest
1/5
1/5
1/5
1/5
1/5

Body Weight(lb)	Hypertension	Weight	Prediction	Weight
120	Yes	0.2	Yes	0.2
115	No	0.2	Yes	0.3
130	Yes	0.2	No	0.3
95	Yes	0.2	1/5	0.2
135	Yes	0.2	1/5	0.2

Figure 1.7

Weighting Scheme Diagram

Bagging, or Bootstrap Aggregating, is a popular ensemble method in machine learning that combines multiple models to improve overall performance. The idea behind bagging is to create multiple subsets of the original training dataset by resampling with replacement and training a model on each of these subsets. The final prediction is obtained by combining predictions from multiple models. Bagging's fundamental aim is to enhance model stability by averaging predictions, contributing to improved accuracy and reliability in predictive modeling. This can be done by creating multiple subsets of the original training data by resampling with replacement. Each of these subsets is then used to train a separate model, which can be a decision tree, random forest, or any other model that can handle the resampled data.



Original Data Set

Study Hours (X)	Exam Score (Y)
2	70
3	75
4	80
5	85
6	90

Bootstrap 1

Study Hours (X)	Exam Score (Y)
2	70
3	75
4	80
5	85
5	85

Bootstrap 2

Study Hours(X)	Exam Score (Y)
2	70
2	70
3	75
3	75
4	80

Figure 1.8

Bootstrapping

Once all the models are trained, they are used to make predictions on new data. The final prediction is obtained by aggregating the outputs of these models. This aggregation can be done in various ways, such as taking the majority vote (for classification problems) or averaging the predictions (for regression problems). The reason why bagging works is that by resampling the data and creating multiple models, we introduce diversity into the models. Each model is trained on a slightly different subset of the data and thus captures different aspects of the data. By combining these models, we can reduce the variance of the final prediction. Bagging offers several advantages over traditional machine learning methods. One of its key advantages is the reduction of variance in the model. By creating multiple models and aggregating their outputs, bagging diminishes the impact of outliers and noise in the data, leading to more reliable predictions. Improved accuracy is another benefit of bagging. By mitigating overfitting, where the model becomes too complex and captures noise in the training data, bagging enhances the model's performance. The introduction of diversity through resampling helps to alleviate overfitting and improves the accuracy of the model on new data.

Bagging is also known for its robustness in handling outliers and noise. Since each model is trained on a slightly different subset of the data, it captures different aspects of the data, making the ensemble more resilient to irregularities in the dataset. Despite its advantages, bagging does have some limitations. One limitation is increased computational complexity. Training multiple models on different subsets of the data can be computationally expensive, requiring more resources and time. Another limitation is the lack of interpretability. Bagging generates multiple models, making it challenging to interpret and explain their contributions to the final prediction.

Furthermore, bagging may not perform well on imbalanced datasets where one class significantly outweighs the others. The resampling process can amplify the dominant class, leading to biased predictions. Several steps are involved in implementing bagging. First, multiple subsets of the original training data are created through resampling with replacement. Each subgroup is used to train a separate model. Next, these models are used to make predictions of new data. Finally, the outputs of these models are aggregated to obtain the final prediction, typically achieved through majority voting or averaging.

1.3.5 Logistic Regression

Logistic regression is a popular statistical modeling and machine learning algorithm mainly used for binary classification tasks. Its importance comes from estimating the likelihood that an instance belongs to a particular class based on its characteristics, and unlike linear regression, which predicts continuous values, logistic regression uses a logistic function, often called the sigmoid function $p(x)$, to model the relationship between the probabilities of the independent and dependent variables, which is defined as

$$p(x) = \frac{\boldsymbol{e^{(\beta_0+\beta_1 x)}}}{\boldsymbol{1+e^{(\beta_0+\beta_1 x)}}},$$

where x = input value,

β_0 = intercept term,

β_1= coefficient for input.

In the other form, the logistic function of logistic regression can be expressed as

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

where μ and s are location and scale parameters, respectively.

A core assumption of logistic regression is that the log-odds transformation (logit) of the dependent variable has a linear relationship with the independent variable. This assumption allows the estimation of coefficients that quantify the influence of each independent variable on the probability of predicting an event. When applying logistic regression, it is important to consider the assumptions of linearity, independence of errors, and absence of multicollinearity. One of the most crucial test in logistic regression is the deviance and likelihood ratio test. According to the value of the test, the evaluation of the test would be performed as a smaller value of deviance is better. The deviance and likelihood ratio test is expressed as

$$D = -2ln\frac{likehood\ of\ the\ fitted\ model}{likehood\ of\ the\ saturated\ model}.$$

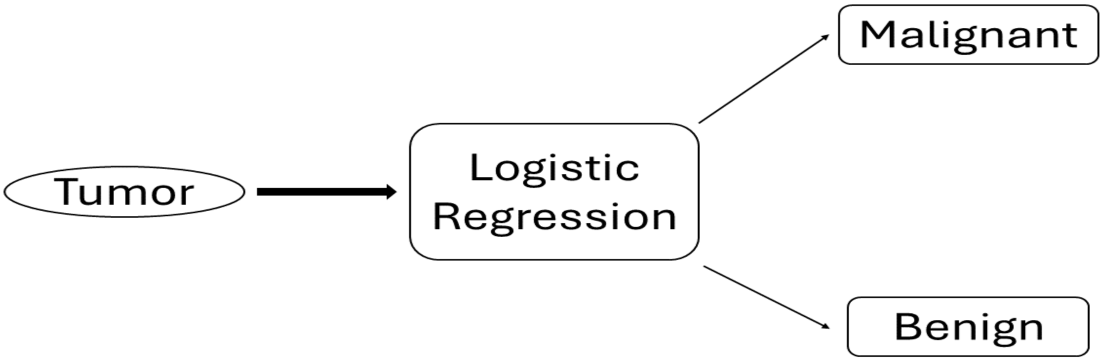


Figure 1.9

Breast Cancer Prediction Using Logistic Regression

Training a logistic regression model involves estimating the coefficients that best fit the data. This is usually achieved by maximum likelihood estimation, which aims to find a set of coefficients that maximizes the likelihood of observed data in a given model. Alternatively, an optimization algorithm such as gradient descent can iteratively update the coefficients until convergence. The performance of logistic regression models is evaluated using various metrics such as accuracy, precision, recall, and F1 score to assess the model's ability to classify instances correctly. Interpreting the coefficients of a logistic regression model is essential for understanding the relationship between independent variables and outcomes. A positive coefficient indicates a positive relationship, suggesting that an increase in the corresponding independent variable leads to an increase in the log probability of the predicted event. Conversely, a negative coefficient indicates a negative relationship. The size of the coefficient represents the strength of the association.

Logistic regression is commonly used for binary classification tasks but can also be extended to handle multiclass problems. One approach is the one-vs-all method. This involves training multiple logistic regression models, each representing one class relative to all others. Another approach is multinomial logistic regression, which simultaneously models the probabilities of various courses. These enhancements enable logistic regression to handle more complex classification scenarios. Logistic regression is used in many fields. Medical diagnoses can predict the probability of disease based on various symptoms and risk factors. In credit risk assessment, logistic regression helps determine the likelihood of a loan applicant's default. Logistic regression is also used in marketing analytics to predict customer behavior by analyzing various demographic and behavioral characteristics.

Advantages of logistic regression include its simplicity, interpretability, and efficiency. This transparent model provides insight into the relationships between variables and outcomes. You can interpret the coefficients to understand the impact of each variable on your predictions.

In addition, logistic regression works well even with limited data and can handle high-dimensional datasets efficiently. However, logistic regression also has limitations. It assumes a linear relationship between the independent variables and the logarithmic rate, which may not be accurate in complex scenarios. Non-linear relationships may require different modeling techniques. In addition, logistic regression is susceptible to outliers, and multicollinearity between independent variables can affect the stability of coefficient estimates.

Logistic regression is a versatile and widely used algorithm for binary classification tasks. Its ability to estimate probabilities and provide interpretable results has value in several areas. By understanding the assumptions, training process, metrics, and interpretation of coefficients, practitioners can effectively use logistic regression for predictive modeling and extract valuable insights from their data. Despite its limitations, logistic regression remains a powerful tool for understanding and predicting binary outcomes.

CHAPTER TWO: STATISTICAL ANALYSIS

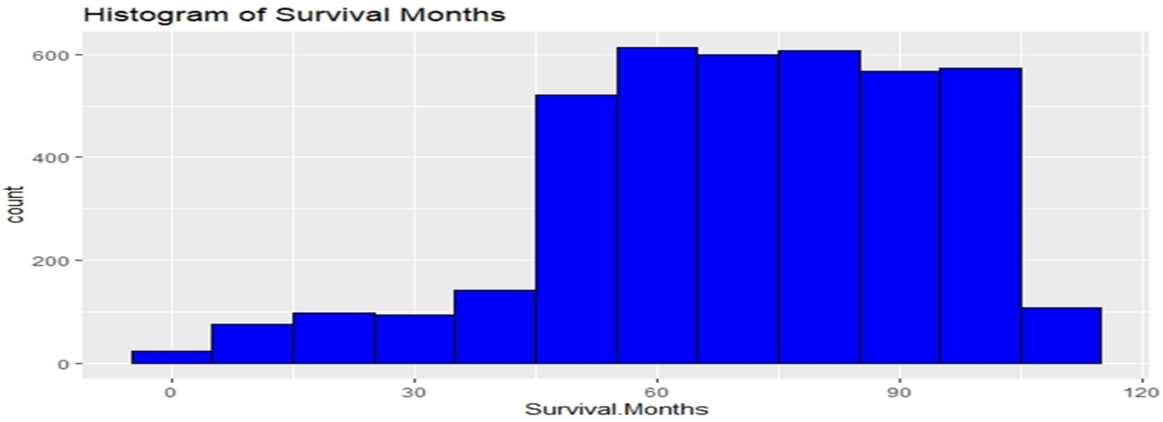
2.1 Multiple Linear Regression Analysis

All analyses of this study were meticulously conducted with the R Studio environment, utilizing the latest version of R Studio (2023.12.1) to ensure access to the most advanced tools, techniques and methodologies tailored to the specific needs of the research project.

Descriptive statistics were employed to summarize the dataset. For the variable "Age," the minimum and maximum ages were 30 and 69 years, respectively, while the average age was approximately 54 years, with a median of 54 years. Most patients were 47 to 61 years (1st quartile to 3rd quartile). Race distribution indicated 291 Black, 320 from other races, and 3413 White patients. Marital status was categorized into four different classes: divorced (486), married (2643), separated (45), single (615), and widowed (235).

In breast cancer, the T stage, N stage, and A stage collectively serve as pivotal indicators guiding diagnosis, treatment planning, and prognosis. The T stage provides crucial information on the size and extent of the primary tumor within the breast tissue, aiding clinicians in determining the appropriate therapeutic interventions. T1 represents a small primary tumor, T2 indicates a giant tumor, T3 signifies further extension into surrounding tissues, and T4 denotes an advanced stage with infiltration into adjacent structures or organs. The N stage focuses on regional lymph nodes, offering insights into the spread of cancer beyond the primary site. Understanding lymph node involvement is vital for assessing disease progression and tailoring treatment strategies. N1 suggests the presence of cancer in nearby lymph nodes, and increasing numerical values (N2, N3) signify more extensive lymph node involvement, reflecting the potential spread of cancer beyond the primary site. Last, assessing distant metastasis (A stage) is indispensable in identifying whether cancer has spread to distant organs, influencing the aggressiveness of treatment and prognosis. A comprehensive evaluation of these stages is imperative in breast cancer management, enabling healthcare professionals to devise personalized and effective strategies for each patient based on the specific characteristics and extent of the disease. Cancer grading classifies tumors based on the degree of abnormality and differentiation of cells, where Grade 1 indicates well-differentiated and less aggressive cells, Grade 2 signifies moderately differentiated cells with intermediate aggressiveness, Grade 3 denotes poorly differentiated cells with high aggressiveness, and Anaplastic (Grade IV) refers to undifferentiated cells displaying extreme aggressiveness and rapid growth.

Analysis of tumor stage (T Stage) revealed 1603 T1, 1786 T2, 533 T3, and 102 T4 cases. The lymph node involvement (N Stage) degree included 2732 N1, 820 N2, and 472 N3 instances. Additionally, information on cancer stage (6th Stage), differentiation grade (differentiate), tumor grade (Grade), cancer spread (A Stage), tumor size, estrogen status, progesterone status, number of examined regional nodes which is the lymph nodes near a primary tumor site that are examined for signs of metastasis in cancer staging, number of positive regional nodes (Regional Node Positive), and survival months were analyzed.



o

Figure 2.1

Survival Spectrum Histogram

The histogram of survival months indicates that most survival times fall within the 60 to 90-month range. This suggests a concentrated distribution of survival durations within this timeframe.

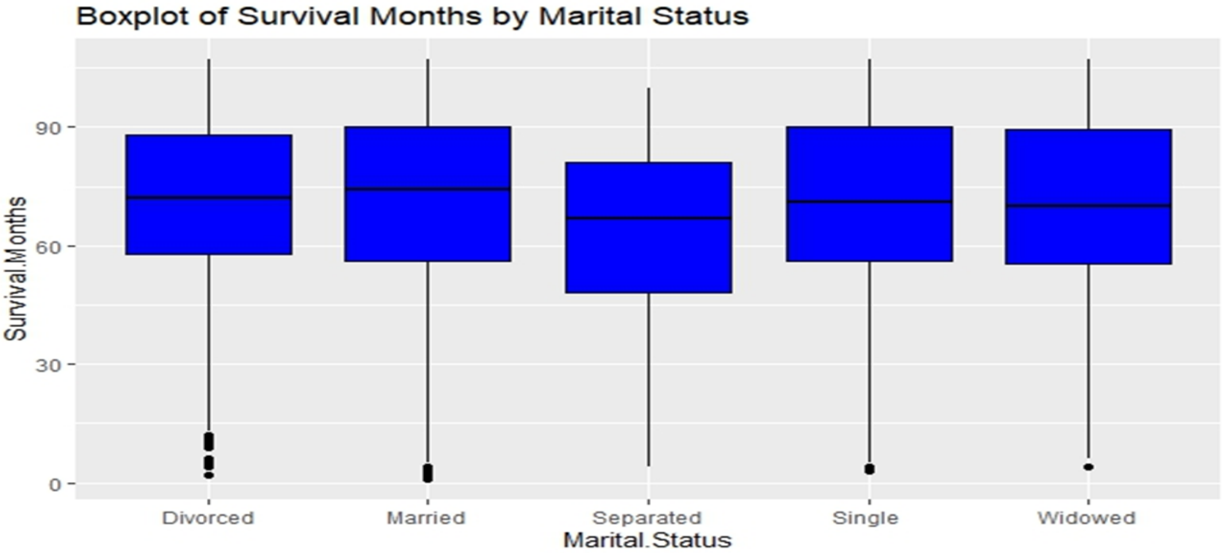


Figure 2.2

Marital Milestone Survival Box Plot

The graphical representation demonstrates a consistent relationship between the categories and survival duration, with minimal variation across most groups. However, individuals with a separated marital status exhibit a slightly lower median survival duration. These findings suggest that marital status, particularly separation, may significantly influence the observed survival days.

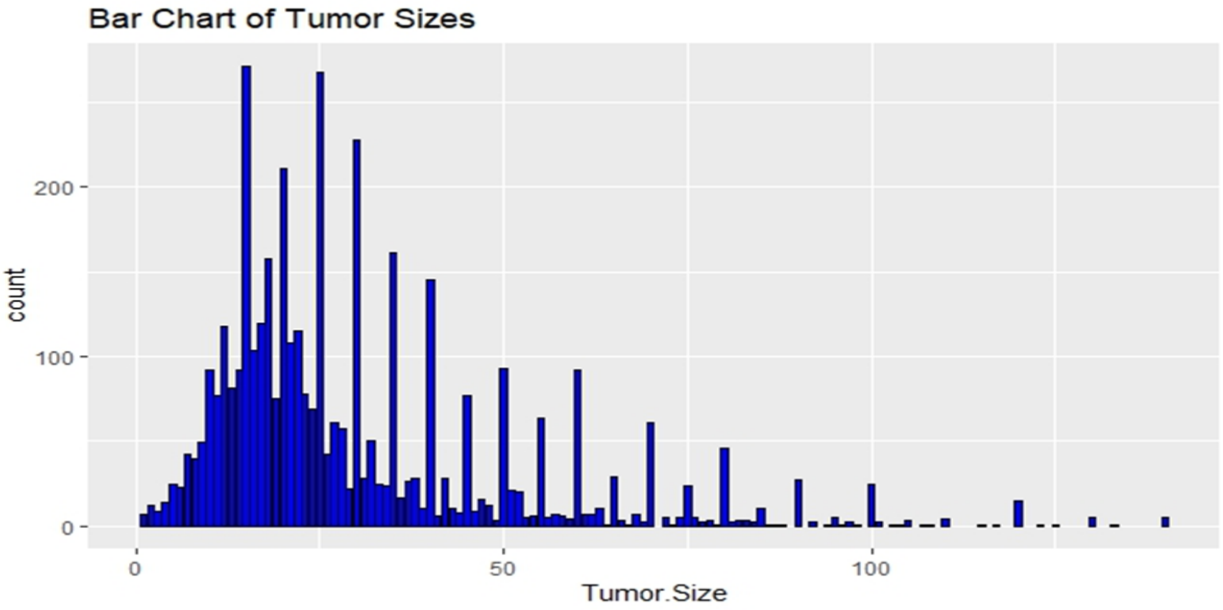


Figure 2.3

Tumor Size Distribution Bar Chart

The bar graph illustrating tumor size distribution reveals a predominant pattern: the majority of tumor sizes are smaller than 100 units. This observation underscores the prevalence of relatively smaller tumor sizes within the studied population.

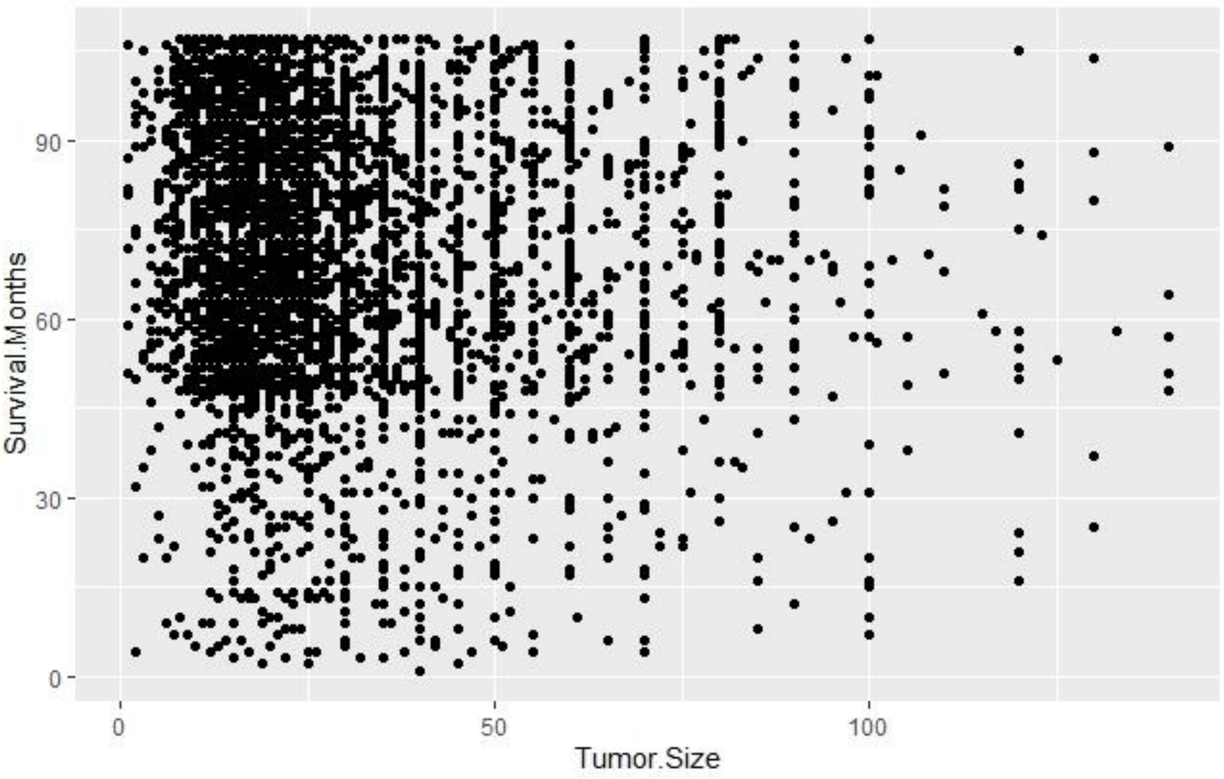


Figure 2.4

Tumor Size Vs. Survival Months Scatterplot

The relationship between tumor size and survival months exhibits the association whereby larger tumor sizes are associated with shorter survival durations. This relationship underscores the potential impact of tumor size on patient outcomes, suggesting that greater tumor size may indicate a more aggressive disease progression or reduced treatment efficacy. Further investigation is warranted to explore the underlying mechanisms driving this observed relationship and its implications for clinical management.

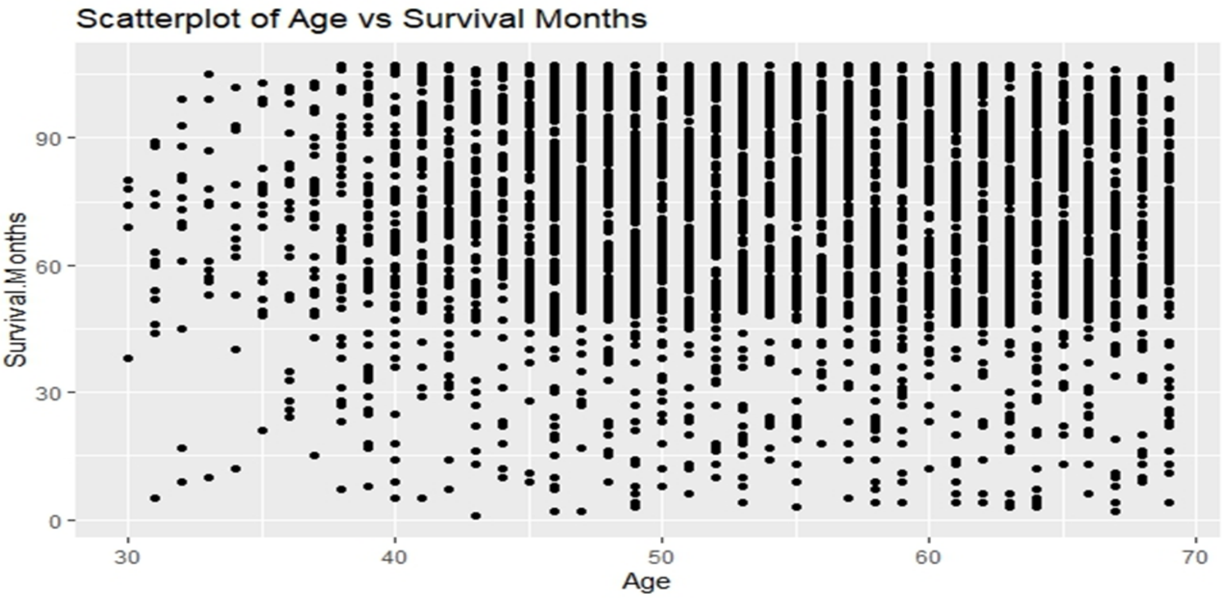


Figure 2.5

Age – Survival Correlation Chart

The analysis reveals a lack of significant association between age and survival months, indicating no discernible relationship between these variables within the examined dataset. This finding suggests that age alone may not serve as a prognostic factor for survival duration in the context of the studied population. Additional research is warranted to explore potential confounding factors or underlying mechanisms contributing to the absence of an age-survival relationship.

The results of figure 2.6 indicate a correlation between tumor grade and survival duration; higher tumor grades are associated with shorter survival times. This finding suggests that tumor grade may serve as an important prognostic factor, with higher-grade tumors potentially indicating a more aggressive disease course and reduced survival. The survival months gradually decrease from Grade 1 at 75 to Grade 3 at 69 months.

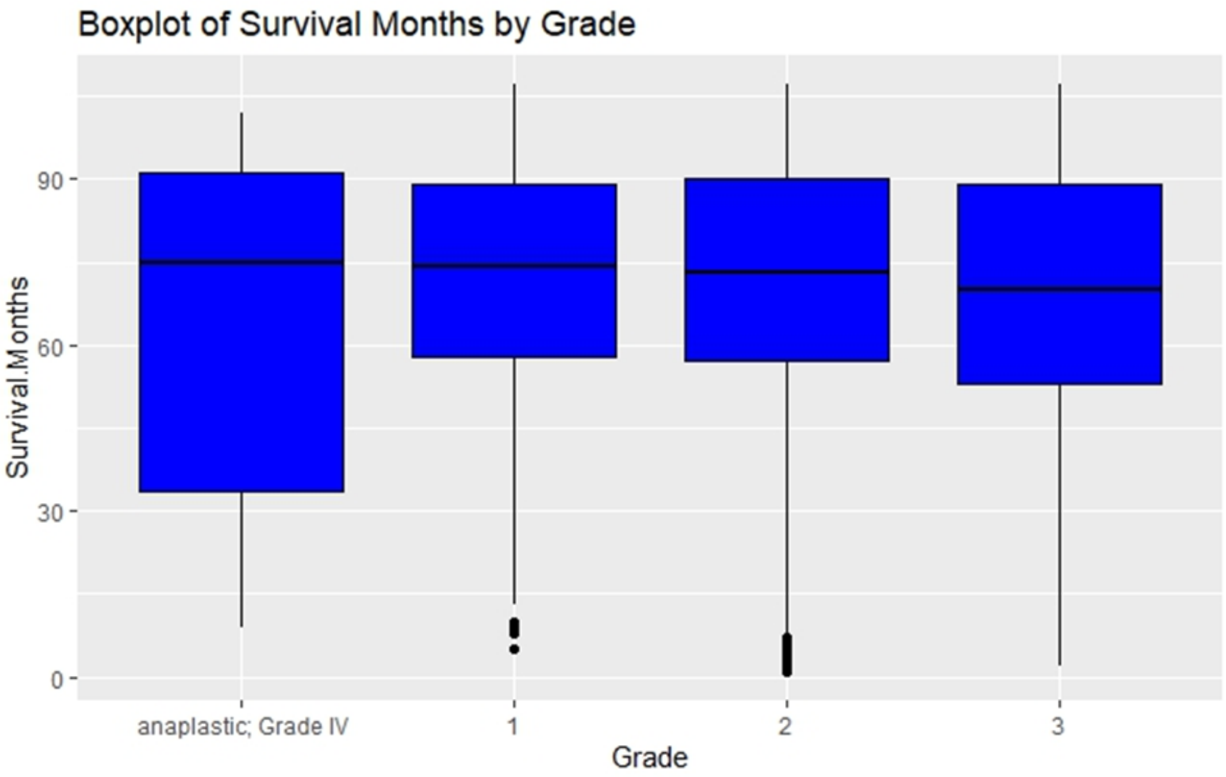


Figure 2.6

Survival Month: Insights Box Plot

The correlation analysis reveals exciting insights into the relationships among the variables. Age shows a strong positive correlation with itself, which is expected. It also exhibits small negative correlations with tumor size, regional node examined, regional node-positive, and survival months. This suggests that older patients tend to have slightly smaller tumor sizes, fewer regional nodes examined, fewer positive regional nodes, and somewhat shorter survival months, although these correlations are weak. Tumor size displays a slight negative correlation with age, indicating that younger patients tend to have slightly larger tumors. It shows small positive correlations with regional node examined and regional node-positive, implying that patients with larger tumors tend to have more regional and positive regional nodes examined. Furthermore, tumor size exhibits a slight negative correlation with survival months, indicating that patients with larger tumors tend to have slightly shorter survival months.

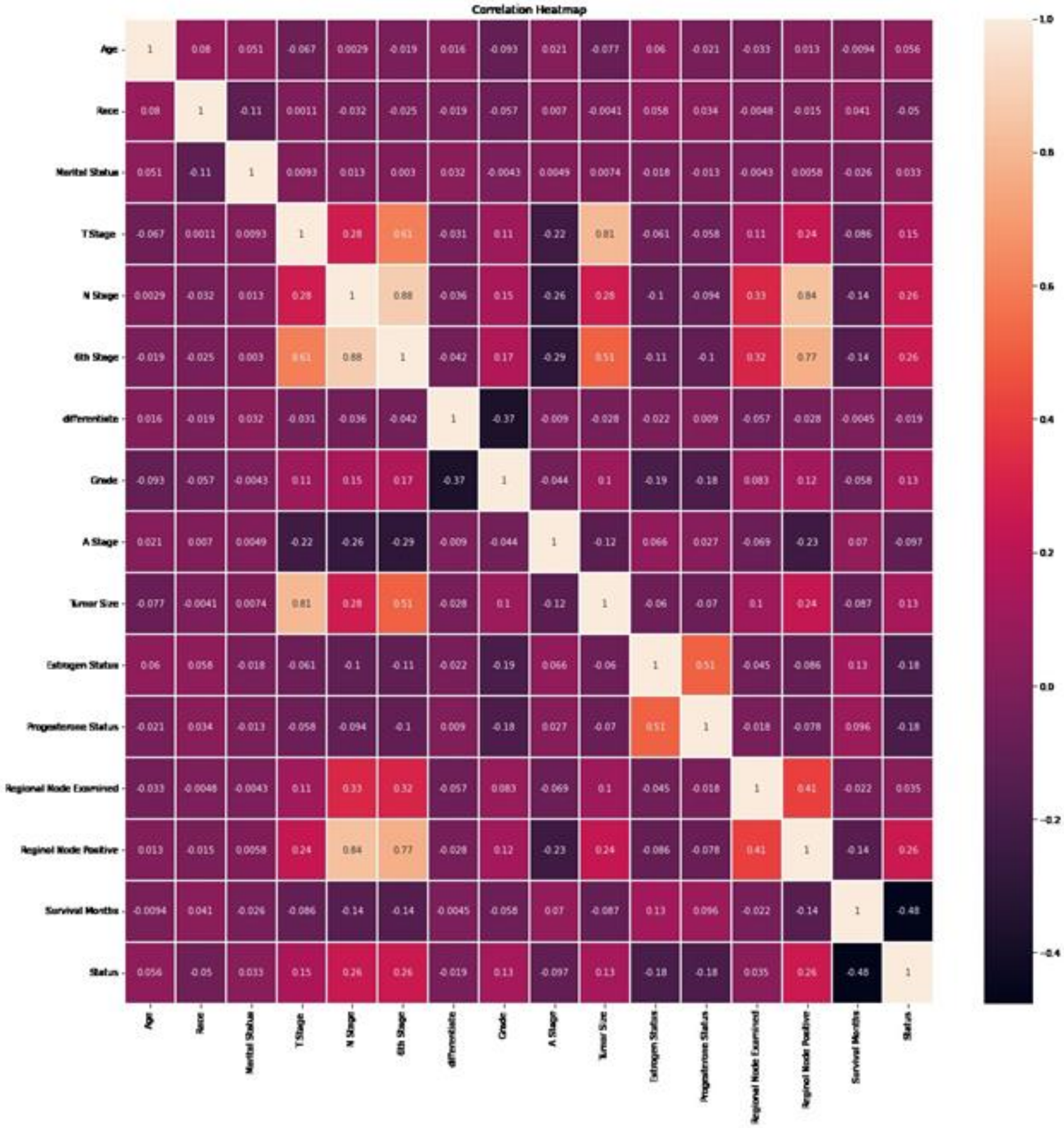


Figure 2.7

Correlation Heatmap

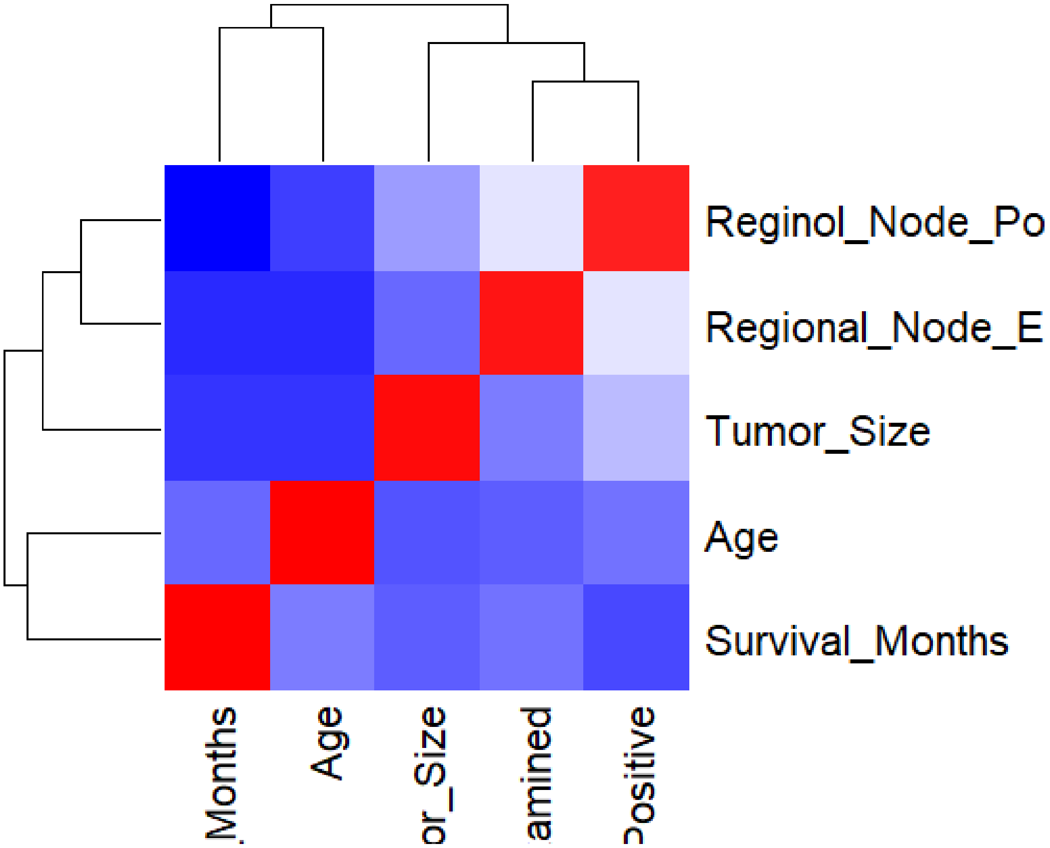


Figure 2.8

Correlation Plot

The number of regional nodes examined has a slight negative correlation with age and a small positive correlation with tumor size and regional node positivity. This suggests that patients with more regional nodes examined tend to be younger, have larger tumors, and have more positive regional nodes. Additionally, there is a minimal negative correlation between the regional node and survival months, indicating that patients with more regional nodes examined tend to have slightly shorter survival months. Regional node-positive, representing the number of positive regional nodes, exhibits a minimal positive correlation with age, a moderate positive correlation with tumor size and regional node examined, and a slight negative correlation with survival months. This implies that patients with more positive regional nodes tend to be slightly older, have larger tumors, have more regional nodes examined, and have shorter survival months. Survival months display small negative correlations with age, tumor size, and regional node examined, and the regional node positive. This indicates that patients who survive longer tend to be younger, have smaller tumors, have fewer regional nodes examined, and have fewer positive regional nodes.

It is important to note that correlation coefficients range from -1 to 1, where coefficients close to 1 or -1 indicate a strong positive or negative relationship. Coefficients close to 0 suggest no significant relationship between the variables. The dataset has 16 variables (including the response) and 4,024 data points. From the summary, it is seen that the variables have no missing values. Now, we fit the new dataset into the multiple linear regression model below:

Survival Months = Age + Race + Marital Status + T Stage + N Stage + 6th Stage+ differentiate + Grade + A Stage + Tumor Size + Estrogen Status + Progesterone Status + Regional Node Examined + Regional Node Positive

The regression model has 14 predictor variables, including Age, Race, Marital Status, T Stage, N Stage, 6th stage, differentiate, Grade, A Stage, Tumor Size, Estrogen Status, Progesterone Status, Regional Node Examined, Regional Node Positive, and Status, whereas the response variable is survival months. The p-values of < 2.2e-16 in R output indicate the statistical significance of each predictor variable, whereas predictors with smaller p-values are considered more statistically significant. The Adjusted R-squared value of 0.22 means that the model explains 22% of the variability in the response variable. The F-statistic of 48.52 and the p-value of less than 2.2e-16 suggest that the model is statistically significant and provides a better fit than a model with no predictors.

2.2 Stepwise Regression

We perform a stepwise multiple regression analysis using Akaike information criterion (AIC) as a selection criterion for further analysis. This analysis aims to identify the subset of independent variables that yields the best-fitting model. According to the Stepwise Regression results, the model with the lowest AIC value is selected as the best model. We chose the lowest AIC value of the model, which is AIC=24166.07. The final model includes the independent variables, "Estrogen Status," "Progesterone Status," and "Regional Node Examined." The model with the lowest AIC value of 24166.07 incorporates the predictor variables mentioned above and can be expressed as below:

Survival Months = 57.43+9.55×Estrogen Status+1.84×Progestrone Status+0.13×Regional Note Examined

2.3 Box-Cox Transformation

The lambda value found from the Box-Cox transform is 1. In other words, if the Box-Cox transformation suggests a lambda value of 1, the original data already meets the assumptions of the statistical model. The data is already suitable for analysis, and there is no need to perform any transformation.

From the Normal Q-Q plot, it seems that though most of the residuals fall along the line, at the upper tail, they deviate a lot, indicating a possibility of violating the normality assumption. From the residuals vs fitted plot, it seems that the residuals are sorted because there is a categorical or continuous variable that has been binned or discretized.

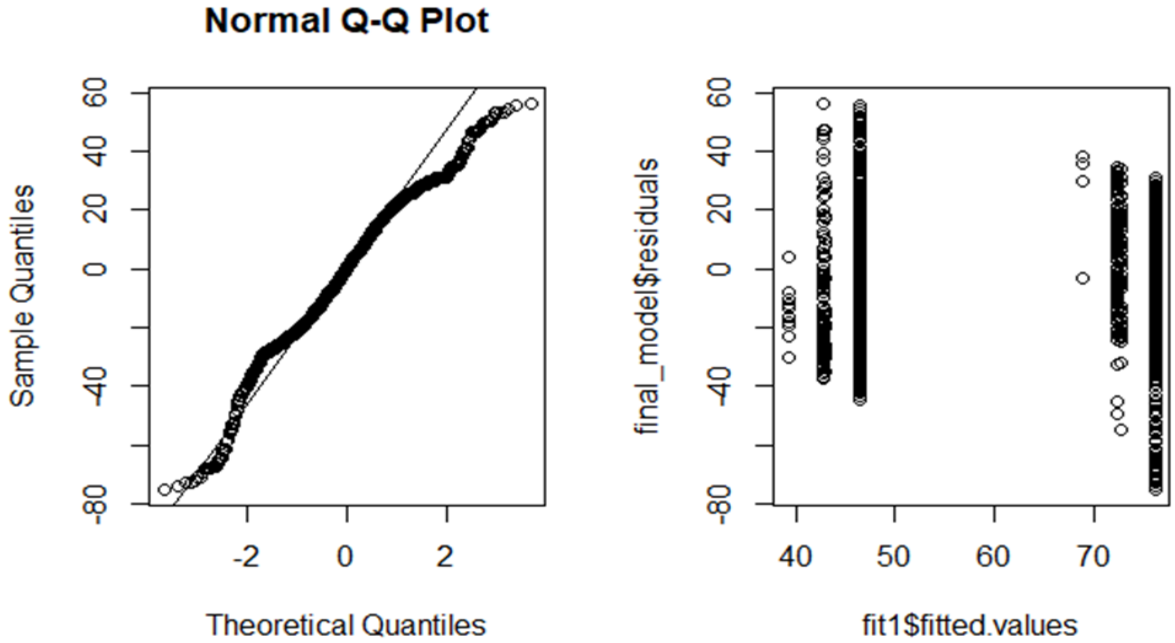


Figure 2.9

Normal Standard QQ Plot and Residual Vs Fitted Value Plot

Table 4

Diagnostic Test Result

Name of the Diagnostic Tests	Test Result	P- value
Shapiro-Wilk Normality Test	W = 0.98	< 0.01
Studentized Breusch-Pagan test	BP = 49.71	<0.01
Durbin-Watson test	DW = 1.97	0.24

The Shapiro-Wilk test is the most robust for checking the normality of the residuals. As shown in Table 2.1, the p-value from the test suggests strong evidence that the data significantly departs from a normal distribution. We conclude that the data do not follow a normal distribution. Breusch-Pegan is a statistical test that assesses the presence of heteroscedasticity or unequal variance of errors in a regression model. The test involves regressing the squared residuals on the independent variables, and the significance of this auxiliary regression is used to determine if heteroscedasticity is present. As reported in Table 2.1, the Breusch-Pagan test also gives p-value less than 0.01, indicating that the assumption of homoscedasticity is violated, thereby indicating the presence of heteroscedasticity.

Though autocorrelation is mainly related to time-series data, it is also often viable in a regression framework. The Durbin-Watson test statistic value of close to 2 indicates no autocorrelation. The statistic value ranges from 0 to 4, where a value close to 0 indicates a positive correlation and a value close to 4 indicates a negative correlation. From the DW test in Table 2.1, we found that the statistic value was close to 1.9778, and the p-value showed that the null was retained. Both ensure that there is no autocorrelation among the residuals.

2.3 Conclusion

Recall the final model:

Survival Months = 57.43+ 9.55×Estrogen Status Positive+1.84×Progestrone Status +

0.13×Regional Note Examined.

According to the final model, the Status of Estrogen, Progesterone Status, and Regional Node Examined positively relate to the Survival Months. This means that an increase in these positively related variables will increase the Survival Months in terms of their hormonal status and lymph node.

CHAPTER THREE: MACHINE LEARNING ALGORITHMS

When employing machine learning methodologies of research such as decision tree, naïve Bayes, AdaBoost, bagging, SVM and logistic regression within R Studio environment, the “caret” package serves as a comprehensive engine. For decision tree, the “rpart” package, available on CRAN (comprehensive R archive network), facilitates tree construction and visualization through recursive partitioning. Naïve from CRAN, employing Bayes’ theorem with strong independence assumptions. To implement AdaBoost bagging techniques, the “ipred”, and “adabag” package extend functionalities beyond base algorithms, offering options such as cross-validation. For SVM, the “e1071” package provides tools for classification and regression tasks, equipped with diverse kernel functions and hyperparameter tuning, At last, logistic regression modeling can be achieved using “glm” package, an integral component of the base R installation, allowing for the fitting of generalized linear models for binary or multinomial outcomes.

3.1 Decision Tree

Several vital components provide valuable insights into the model's performance and feature importance in the decision tree model analysis. First, the total number of observations (n) utilized in the model stands at 2818, reflecting the size of the dataset employed for training and evaluation. Additionally, crucial metrics such as the complexity parameter, number of splits, relative error, cross-validated error, and standard error of the cross-validated error contribute to assessing the model's complexity and potential overfitting. Determining the complexity parameter that yields the minor cross-validated error while still considering the standard error is a common practice to determine an optimal level of pruning for the decision tree, mitigating overfitting concerns.

Decision tree analysis is a cornerstone methodology within medical research, providing a robust framework for exploring the intricate relationships between various prognostic factors and patient outcomes. In healthcare, decision trees are valuable for uncovering predictive patterns and identifying critical disease progression, treatment response, and survival determinants. By systematically partitioning patient data based on a series of decision rules, decision tree analysis facilitates the identification of critical predictors and their relative importance in predicting clinical endpoints. This introductory paragraph sets the stage for a detailed examination of decision tree analysis within medical research, emphasizing its significance in advancing our understanding of disease processes and informing evidence-based healthcare practices.

The concept of variable importance (VI) offers insights into the relative significance of predictor variables within the model. Moreover, the node information provides detailed statistics for each node in the tree. This includes the number of observations and complexity parameters. Additionally, it highlights the primary and surrogate splits employed to partition the data. The "improve" value of the variables indicates the degree of improvement achieved in the model fit because of each split. The decision tree model under examination is constructed based on a classification framework, employing a range of predictor variables encompassing demographic characteristics, tumor staging parameters, and biomarker statuses. Specifically, the model is formulated using the categorical variables race, marital status, N Stage, T stage, A stage, estrogen status, and differentiate to predict the binary outcome variable denoting patient status as "Alive" or "Dead."

In analyzing decision tree models, pruning constitutes a pivotal technique to refine the model's structure to mitigate overfitting and enhance its generalization capabilities. The model output provides insights into this process by presenting complex parameters and the resulting tree structure. The CP values indicate the model's complexity at different stages of tree growth. Initially, with a CP value of 0.0173, the tree remains in its maximal complexity state, devoid of any splits. As the tree grows and additional splits are incorporated, the CP value influences the decision to further partition nodes based on their predictive utility. Subsequently, with a second CP value of 0.01 and an incremented number of split values, the pruning process commences, wherein nodes with marginal improvements in predictive accuracy are systematically pruned from the tree. This iterative process balances model complexity and predictive performance, ensuring the final decision tree is interpretable and robust. By selectively removing unnecessary nodes, pruning facilitates the construction of a parsimonious model capable of capturing the underlying patterns in the data while minimizing the risk of overfitting. Thus, within the framework of the decision tree model, the CP values offer valuable guidance for optimizing model performance through effective pruning strategies.

An integral aspect of the decision tree analysis lies in assessing VI and elucidating the relative contributions of individual predictors toward predicting survival outcomes. In particular, the study reveals N stage, estrogen status, and A stage as the most influential variables, with respective importance scores of 76, 17, and 7. These scores are critical in tumor staging parameters and hormonal statuses in prognostication, informing clinical decision-making processes and treatment strategies. The structural configuration of the decision tree unveils insightful patterns and hierarchical relationships between predictor variables and patient outcomes. Beginning with the root node, which classifies most observations as "Alive," the subsequent nodes delineate specific characteristics associated with differing survival probabilities. Its splits, such as N stage, estrogen status, and differentiation, provide nuanced insights into the heterogeneity of patient populations and the multifactorial nature of survival outcomes.

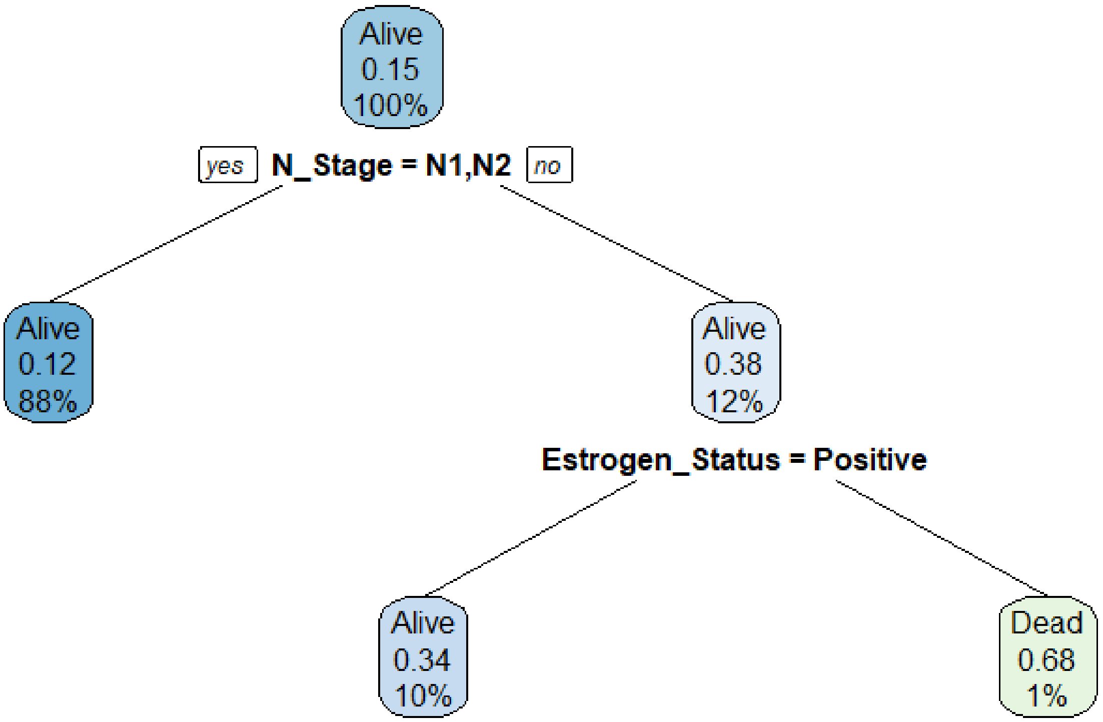


Figure 3.1

Decision Tree Model

The decision tree model evaluation outcome reveals vital performance metrics concerning the test dataset. With an accuracy of 85% in predicting survival status, the decision tree model demonstrates commendable performance. This metric indicates that the model correctly classified the survival outcomes for 85% of the instances within the testing dataset. Such accuracy suggests that the model has effectively learned relevant patterns or relationships from the available features to differentiate between survival and non-survival status. However, while this accuracy rate is relatively high, it's crucial to contextualize it within the specific problem domain and dataset characteristics. The interpretation of the decision tree nodes offers a nuanced understanding of how predictor variables influence survival outcomes. Each node represents a distinct combination of variables, providing valuable insights into the classification of survival predictions. Evaluation metrics such as precision, recall, and F1 serve as critical benchmarks for assessing the model's predictive accuracy and error distribution.

Precision measures the proportion of correctly predicted "live" cases out of all instances predicted as "live" is 0.85. It indicates that when the model predicts that a person is "alive", it is correct about 85% of the time. Conversely, recall is reported as 0.99, indicating that the model captures almost all cases of "live" individuals in the dataset. This high recall value suggests that the model has a meager false negative rate, meaning that it rarely fails to detect "live" individuals. Finally, the F1 score, which combines precision and recall into a single measure, is 0.91. It suggests that the model achieves a good balance between precision and recall, indicating strong performance. In conclusion, the decision tree model shows strong predictive ability with high accuracy, precision, recall, and F1 scores, making it a promising tool for predicting survival status in breast cancer.

3.2 Naive Bayes Method

Upon analyzing the summary of the naive Bayes model, it becomes evident that the model object encompasses several essential attributes, each providing unique insights into its structure and characteristics. For instance, the ‘apriori’ attribute encapsulates the prior probabilities of the classes within the dataset. This information is stored as a numeric table, offering a foundational understanding of the class distribution and its potential impact on classification outcomes. Delving deeper, the ‘tables’ attribute emerges as a crucial component, furnishing comprehensive details regarding conditional probability tables for each predictor variable relative to the class variable. Comprising a list of 15 elements, these tables serve as a cornerstone for the model's decision-making process, facilitating the computation of probabilities and aiding in the class assignments.

Further exploration reveals the 'levels' attribute, shedding light on the levels present within the class variable. In the context of binary classification, as indicated by the two levels, this attribute underscores the dichotomous nature of the classification problem, offering clarity on the potential outcomes the model considers. Moreover, the 'numeric' attribute offers valuable insights into the nature of predictor variables. With 15 logical values, this attribute delineates whether each predictor variable is numeric, providing essential context for understanding the data's structure and informing subsequent analysis. Finally, the 'call' attribute encapsulates the function call utilized to fit the naive Bayes model, encapsulating the formula and data employed in the modeling process. Serving as a reference point for reproducibility and transparency, this attribute underscores the meticulousness of the modeling approach and facilitates a deeper understanding of the modeling methodology employed.

The summary of the naive Bayes model transcends mere numerical outputs, offering an advanced glimpse into the model's architecture, assumptions, and underlying mechanisms. By dissecting each attribute, researchers can glean valuable insights into the model's construction and performance, paving the way for informed decision-making and further exploration of classification tasks. The accuracy and kappa values serve as critical metrics for assessing the performance of the Naive Bayes classification model. Accuracy, computed as the ratio of correctly classified instances to the total number of cases, reveals the model's overall correctness in predictions. With an accuracy of 0.80, 80% of cases are correctly classified by the Naive Bayes model, indicating a relatively strong performance in accurately predicting class labels. However, accuracy alone may not comprehensively assess the model's effectiveness, especially in scenarios with imbalanced datasets or varied consequences of misclassification.

On the other hand, the kappa statistic, also known as Cohen's kappa, offers insight into the agreement between the model's predictions and the actual classes, accounting for agreement occurring by chance. Kappa value of 0.25 signifies the extent of agreement beyond what would be expected by random chance alone. While this value suggests some degree of agreement between the model's predictions and the actual classes, its moderate magnitude implies room for improvement in capturing the nuances of the classification task. It is essential to contextualize these metrics within the specific requirements of the problem domain and consider additional evaluation measures to gain a comprehensive understanding of the naive Bayes model’s performance.

In the precision-recall curve, the vertical axis denotes precision, varying from 0 to 1, while the horizontal axis signifies recall, or sensitivity, also ranging from 0 to 1. This graphical representation illustrates the delicate balance between these two fundamental metrics across diverse classification thresholds. For a binary classification task with two distinct categories, such as "alive" and "dead" in survival status prediction, each point on the curve corresponds to a specific threshold governing the classification of instances as positive (e.g., "alive") or harmful (e.g., "dead"). In the context of survival status prediction, the curve initiates from the origin (0.0, 0.0), indicating a threshold where no instances are designated as positive, thus resulting in both precision and recall being zero. As the threshold escalates, more instances are classified as positive, augmenting the recall.

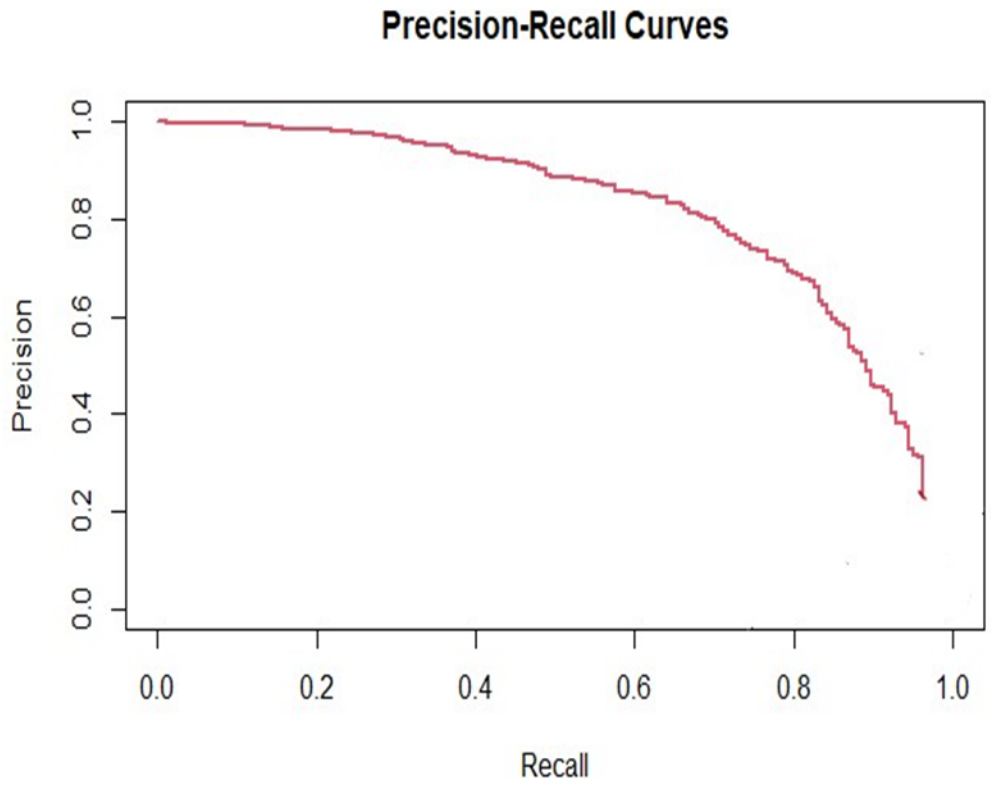


Figure 3.2

Naive Bayes Model Precision Recall Plot

The trajectory of the curve typically ascends towards the upper-right corner of the plot, indicative of elevated precision and recall values, signifying superior model performance. In our scenario, the curve's culmination at the point (1.0, 1.0) denotes that all instances classified as positive indeed represent "alive" individuals (100% recall), and all optimistic predictions are accurate (100% precision). The precision-recall curve is a valuable tool for gauging the model's capability to identify instances of interest while concurrently minimizing false positives accurately. A higher area under the precision-recall curve (AUC-PR) denotes enhanced model performance in distinguishing between positive and negative instances across a spectrum of classification thresholds. The precision, recall, and F1 score of the naive Bayes model are 0.88, 0.87, and 0.88, respectively. This assessment metric holds particular significance in scenarios characterized by class imbalance or when prioritizing the identification of positive instances over the correct classification of negative cases.

3.3 SVM

The SVM model employed in this analysis is designed explicitly for regression tasks and falls under epsilon-insensitive regression, where a margin of error is allowed to predict the continuous values. This model's chosen SVM kernel function is the radial basis function. It is known for its ability to map data into infinite dimensions, creating complex decision boundaries that can capture intricate relationships within the data. The radial basis function (RBF) is a commonly used kernel function in SVM for classification and regression tasks. The RBF kernel function is defined as:

$$K\left(x_i, x_j\right) = (\exp(-\gamma \cdot \left\|x_i - \ x_j\right\|^{\ 2}),$$

here, x_i, x_j represent data points in the feature space, and γ, gamma parameter, serves as the positive constant that determines the shape of the decision boundary.

The model's cost parameter is set to 1, indicating a balance between allowing some training errors and maintaining a reasonable margin. This parameter governs the trade-off between the flexibility of the decision boundary and the tolerance for misclassifications. A soft margin is created by setting the cost parameter to 1, allowing for a certain degree of misclassification. Larger values of the cost parameter increase the cost associated with misclassifying data points, leading to a more sensitive decision boundary that can be influenced by individual data points, resulting in higher variance and lower bias. The gamma parameter, associated explicitly with the RBF kernel, is set to 1 divided by the data dimension, resulting in a value of 0.033 in this context.

Gamma plays a crucial role in determining the influence of data points on the decision boundary. A lower gamma value emphasizes the contribution of distant points to the decision boundary. A higher gamma value gives more weight to points closer to the decision boundary, leading to a more intricate decision boundary with greater flexibility.

The epsilon parameter in the loss function of epsilon-SVR (Support Vector Regression) is set to 1. It defines the epsilon tube within which no penalty is incurred for errors. Errors more minor than the specified epsilon value are considered negligible and are not penalized, allowing some tolerance in the model's performance.

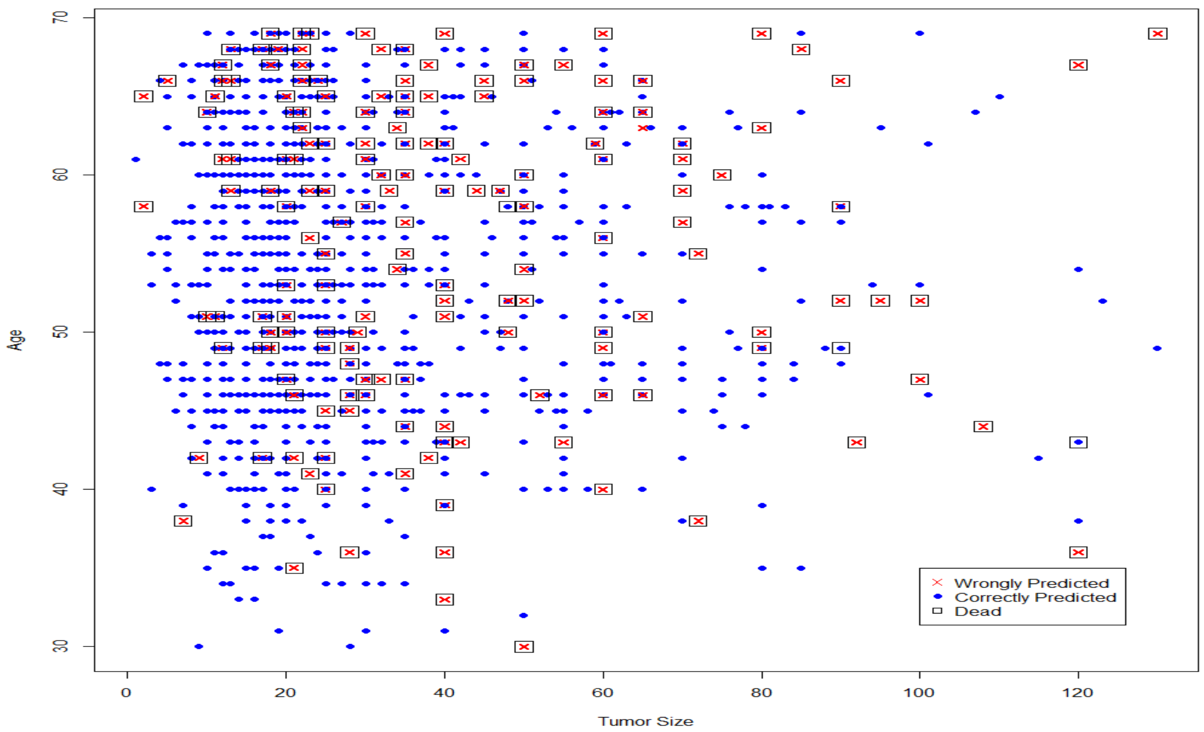


Figure 3.3

SVM Prediction Plot

The prediction scatterplot in Figure 3.3 generated from the SVM model offers valuable insights into the predictive performance regarding cancer survival based on tumor size and age. Blue dots denote the instance that individuals are correctly predicted to be alive or dead. “x” represents a wrongly predicted individual while the square box shows a deceased patient. For example, x with square box represents a deceased patient who was predicted to survive. This plot depicts a concentration of data points primarily between 40 to 70 on the Y-axis, representing age, and 0 to 40 on the X-axis, representing tumor size. This clustering suggests that a significant portion of the data falls within this age and tumor size range, indicating a solid pattern within this demographic. Most of the patients who are between 40 and 70 years old have tumor sizes up to 40mm. Using SVM, most of the predictions are corrected. The rest of the wrongly predicted are mostly deceased patients. The prediction plot shows that the SVM method can predict the survival outcomes of breast cancer patients, and it is beneficial for clinical usage.

The SVM model was constructed utilizing a radial kernel function, chosen for its efficacy in handling non-linear classification tasks by transforming data points into higher-dimensional space. This kernel choice allows the model to capture complex patterns and relationships that may not be discernible in the original feature space. Additionally, the SVM was configured with C-classification, a parameter determining the penalty associated with the misclassification of training examples. In this instance, the cost parameter was set to 1, reflecting a balanced approach where the model aims to minimize misclassifications without overly biasing towards specific data points.

Central to SVM modeling is identifying support vectors, data points crucial for defining the decision boundary. These vectors lie closest to the boundary and are pivotal in shaping its position. In the present model, 996 support vectors were identified, further highlighting the model's reliance on critical data points for accurate classification. Of these, 567 support vectors corresponded to the "Alive" class and 432 to the "Dead" class, suggesting that both classes contribute significantly to delineating the decision boundary. This observation shows the importance of considering the distribution of support vectors in understanding the model's decision-making process.

Moreover, the SVM model was tailored for binary classification, where it distinguishes between two distinct classes: "Alive" and "Dead." This binary classification framework aligns with the specific task of predicting survival outcomes based on the provided dataset. The model aims to categorize individuals into discrete survival categories by focusing on these two classes, providing valuable insights for clinical decision-making and prognosis assessment. Overall, the SVM model's configuration, including the choice of kernel function, cost parameter, and identification of support vectors, reflects a comprehensive approach toward achieving accurate and robust classification performance in predicting patient survival status.

The performance evaluation of the SVM model yielded promising results, with an accuracy of 84 %. These metrics provide insights into the model's ability to classify survival outcomes based on the breast cancer dataset correctly. The high accuracy score indicates that the model accurately predicts the survival status of patients in the test dataset, with 84% of predictions matching the actual labels. Overall, the evaluation of the SVM model underscores its efficacy in predicting survival outcomes based on the provided dataset. The high accuracy and moderate kappa coefficient indicate that the model distinguishes between different survival categories, offering valuable insights for clinical decision-making and patient prognosis assessment.

However, accuracy cannot guarantee the model-validated performance, precision, recall, and F1 score are calculated. According to the SVM model that predicts the "status" of individuals, the precision is 0.84. This shows that when a model predicts a certain "status", it is correct about

84% of the time. On the other hand, the recall is reported as 0.99, suggesting that the model captures almost all occurrences of the specified "status" in the data set. A model with a high recall value has very few false negatives, meaning it rarely fails to detect cases that match the specified "status." In addition, the F1 score is 0.91. This score indicates that the model achieves a balanced performance between precision and recall, indicating the strength of its predictive capabilities. The SVM model exhibited considerable accuracy and coefficient in its predictions of survival outcomes, suggesting its proficiency in classifying patient survival statuses. In summary, the SVM model has an encouraging high potential for predicting survival outcomes, offering valuable insights for clinical decision-making.

3.5 AdaBoost and Bagging

Upon examining the summary output provided for the AdaBoost model, it becomes evident that the model encompasses a comprehensive array of elements crucial for predictive analysis. Each component within the summary holds significance in elucidating the model's architecture and operational dynamics. First, the AdaBoost model delineates the formulation employed for model construction, encapsulating the relationship between predictor variables and the target variable. This foundational aspect guides the model's predictive inference process, facilitating a structured data interpretation and analysis approach. Moreover, the summary reveals the ensemble nature of the AdaBoost model, manifesting in the presence of multiple decision trees denoted as 'trees.' With 150 trees comprising the ensemble, the model harnesses the collective intelligence of diverse decision-making units to enhance predictive accuracy and robustness. Each tree contributes unique insights from the training data, collectively enriching the model's predictive capabilities.

The allocation of weights to individual trees, as depicted in the 'weights' component, underscores the strategic orchestration of each tree's influence on the final prediction. These weights govern the contribution of each tree to the ensemble's collective decision-making process, facilitating optimal amalgamation of diverse predictions and enhancing model performance. Furthermore, the 'votes' and 'prob' components signify the probabilistic outputs generated by the AdaBoost model. These outputs offer insights into the model's confidence levels regarding individual predictions, empowering stakeholders with a nuanced understanding of predictive uncertainty and facilitating informed decision-making. Additionally, the model's capacity to handle categorical outcomes is evident from the presence of the 'class' component. This feature underscores the model's versatility in accommodating diverse data types and predictive scenarios, ensuring its applicability across various domains and use cases. Last, the 'importance' section elucidates the relative importance of predictor variables in influencing the model's predictive outcomes. By quantifying the significance of each variable, this component guides feature selection and model refinement endeavors, enabling stakeholders to streamline model development processes and enhance predictive performance.

The AdaBoost model, which was 85% accurate, with a precision value of 0.84, suggests that when a model predicts a particular outcome, it is correct about 84% of the time. On the other hand, the recall value is 0.97, indicating that the model effectively captures almost all occurrences of the specified outcome in the dataset. With such a high recall value, the model exhibits minimal false negatives, meaning that cases associated with a given outcome are rarely missed. This F1 score is 0.9, which reflects a balanced performance between precision and recall, indicating the model's strength in correctly identifying cases and minimizing false positives and negatives. In conclusion, the AdaBoost model has vital precision, recall, and F1 scores, suggesting its effectiveness in predicting outcomes in predicting breast cancer survival status. The AdaBoost model summary offers a comprehensive overview of its architecture, functionality, and predictive capabilities. Through elucidating the model's internal components and operational dynamics, the summary empowers valuable insights, fostering a deeper understanding of the model's underlying mechanisms and facilitating effective utilization in real-world scenarios.

The boosting model employed in this investigation provides the framework for analyzing and understanding the intricate relationships between various predictors and outcomes within a specific domain. Among the predictors examined, variables such as 'Age,' 'Marital Status,' and 'Regional Node Examined' emerge as pivotal determinants, underscoring their substantial impact on the model's predictive accuracy. By systematically integrating these predictors into the model architecture, researchers gain valuable insights into the multifaceted nature of the phenomenon under investigation, facilitating a more nuanced understanding of the factors influencing the outcomes of interest.

The VI scores derived from the boosting model offer valuable insights into the relative significance of predictors in predicting outcomes. Moreover, the VI scores shed light on the hierarchical structure of predictors, guiding researchers in prioritizing variables for further investigation and model refinement. Furthermore, the interaction between specific predictors elucidates complex relationships and potential confounding factors within the dataset. Such interactions enrich our understanding of the underlying mechanisms driving the observed outcomes and provide valuable insights into the dynamic nature of the phenomenon under study. By accounting for these interactions, the boosting model enhances the interpretability and predictive accuracy of the analysis, enabling researchers to uncover hidden patterns and associations within the data.

The findings from this analysis have important implications for clinical practice and decision-making processes in healthcare settings. By identifying key predictors and their relative importance in predicting outcomes, the model offers valuable guidance for healthcare practitioners in devising personalized treatment strategies and interventions. Moreover, the insights from the analysis can inform the development of risk stratification models and decision support systems, facilitating more targeted and effective patient care. Additionally, the systematic evaluation of predictors and their interactions enhances our understanding of disease progression and prognosis, paving the way for advancements in precision medicine and personalized healthcare delivery.

The boosting model employed in this study provides a comprehensive framework for analyzing predictors and predicting outcomes within a specific context. The model offers valuable insights into the complex relationships underlying the observed outcomes by leveraging variable importance scores and examining predictor interactions. These findings advance our understanding of the phenomenon under investigation and have important implications for clinical practice and healthcare delivery. Continuing research in predictive modeling and data analysis can further enhance our ability to predict and intervene in complex healthcare scenarios, ultimately improving patient outcomes and quality of care.

In evaluating the Bagging model's performance, it is evident that it demonstrates commendable accuracy, achieving an accuracy of approximately 84% and a Kappa statistic of roughly 0.54. These metrics signify the model's ability to classify instances and perform beyond random chance correctly. A post-resampling analysis is conducted to assess the model's effectiveness further, yielding comprehensive insights into its predictive capabilities. The achieved accuracy and Kappa statistics attest to the model's robustness and efficacy in handling classification tasks, thereby instilling confidence in its applicability and reliability for real-world applications. Moreover, the post-resampling analysis comprehensively assesses the model's predictive performance, facilitating informed decision-making and model refinement efforts.

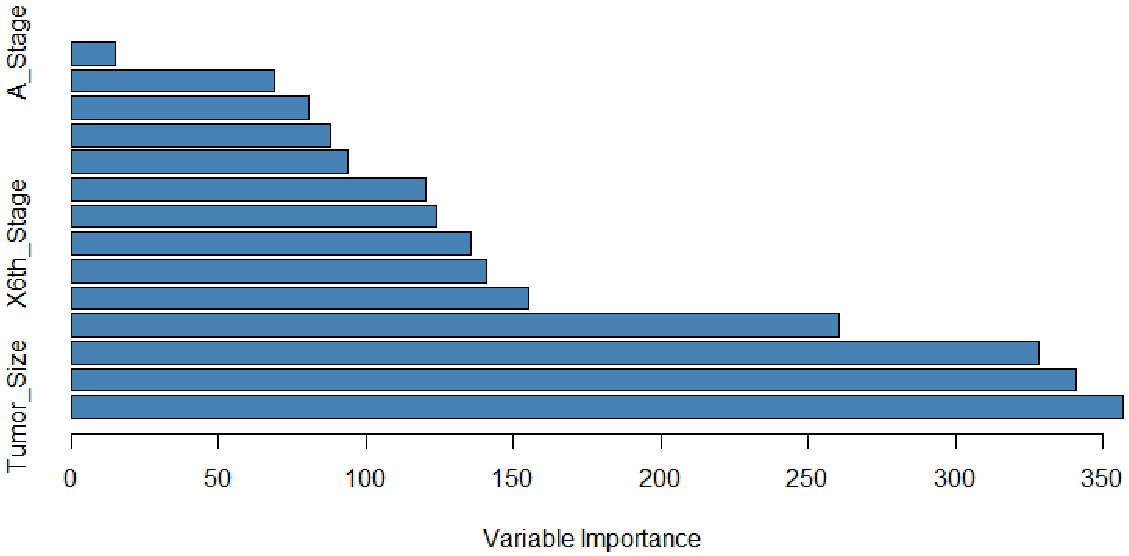


Figure 3.4

Bar Plot of the Importance Variable in the Bagging Model

Moving forward, it is imperative to address the model’s precision, recall, and F1, which are 0.86, 0.97, and 0.91, respectively. This indicates the model's ability to strike a favorable trade-off between correctly identifying instances and minimizing false positives and negatives. Bagging, or Bootstrap Aggregating, is a powerful ensemble learning technique that combines multiple base models to improve overall predictive performance. By training various models on different subsets of the data and then aggregating their predictions, Bagging reduces overfitting and variance while enhancing accuracy and robustness.

3.5 Logistic Regression

Based on the provided data, the logistic regression model was fitted using Status as a dependent variable since it is for a binary classification problem. The model output includes several components and metrics for evaluating the model's performance. The Deviance Residuals section displays the minimum, first quartile, median, third quartile, and maximum values of the deviance residuals. These residuals measure the discrepancy between the observed and predicted values, indicating how well the model fits the data. In this case, the deviance residuals range from -2.0358 to 3.3404. This wide range suggests that the logistic regression model's predictions vary significantly across different observations in the dataset. They may indicate areas where the model is not accurately capturing the underlying patterns in the data. Overall, examining the range of deviance residuals helps assess the model's fit and identify potential areas for improvement.

The "Coefficients" section of table 4 presents the estimated coefficients for each predictor variable in the model. However, due to singularity issues, four coefficients are not defined. Each coefficient is accompanied by its estimate, standard error, t-value, and corresponding p-value. The intercept term has an estimated coefficient of 1.693, and several predictor variables, such as age, race, T stage 3, poorly differentiated, undifferentiated, and well differentiated show statistically significant associations with the outcome variable, as indicated by the p-values.

Table 5

Coefficients of Logistic Regression Model

	β	Standard Error	t – Statistics	P-Value
Intercept	1.693	0.728	2.326	0.02
Age	0.019	0.007	2.512	0.01
Race				
Other	-0.631	0.341	-1.851	0.06
White	-0.324	0.224	-1.446	0.14
Marital Status				
Married	0.017	0.203	0.088	0.92
Separated	1.148	0.577	1.987	0.04
Single	0.070	0.253	0.277	0.78
Widowed	0.420	0.307	1.36	0.17
T Stage				
T-2	0.622	0.290	2.145	0.031
T-3	1	0.460	2.179	0.029
T-4	1.337	0.744	1.796	0.072
N Stage				
N-2	0.844	0.334	2.523	0.011
N-3	0.157	0.441	0.357	0.721
6th Stage				
IIB	-0.154	0.330	-0.468	0.639
IIIA	-0.670	0.419	-1.598	0.110
A Stage Regional	0.074	0.422	0.177	0.859
Tumor Size	-0.01	0.005	-0.270	0.786
Differentiate				
Poorly	0.490	0.146	3.340	0.0008
Undifferentiated	1.311	0.908	1.444	0.148
Well	-0.856	0.254	-3.362	0.0007
Estrogen Status Positive	-0.315	0.273	-1.153	0.0175
Progesterone Status Positive	-0.435	0.183	-2.376	0.017
Regional Node Examined	-0.027	0.009	-2.894	0.003
Regional Node Positive	0.062	0.022	2.744	0.006

The confusion matrix presented depicts the performance evaluation of a logistic regression model. The model's predictions are compared against the actual outcomes for the Alive and Dead classes. The matrix reveals that out of all the instances classified as Alive, only 22 were accurately predicted, while a staggering 1000 instances were falsely classified as Alive when they were Dead. Similarly, the model correctly predicted 26 instances for the Dead class but misclassified 158 instances as Alive.

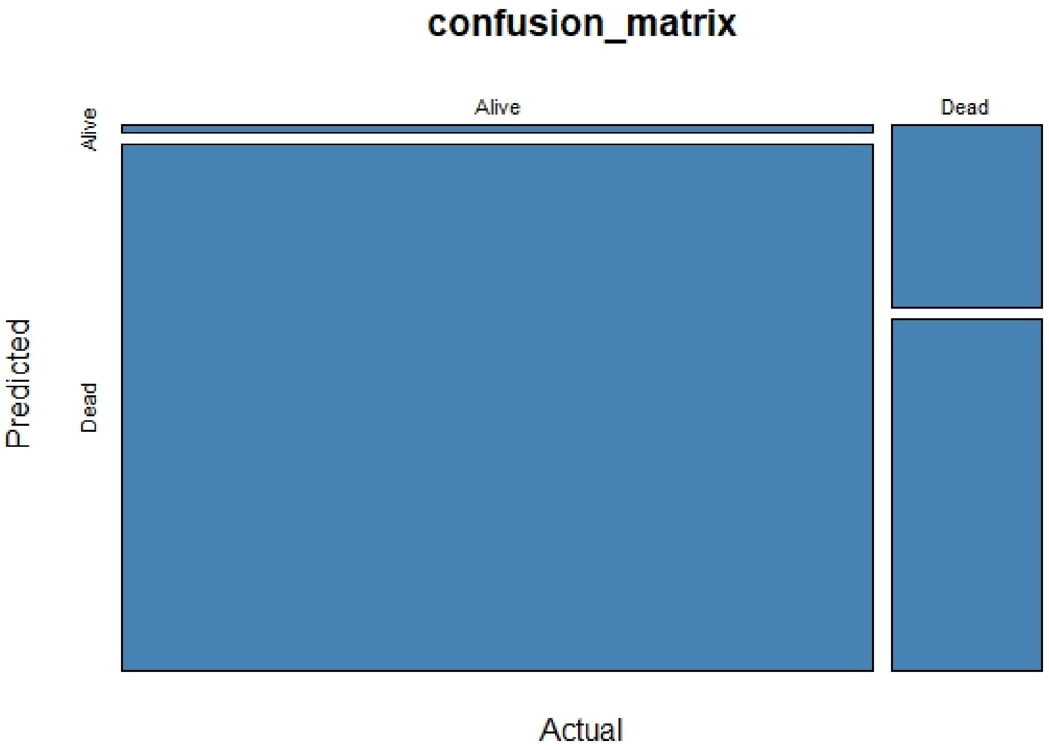


Figure 3.5

Confusion Matrix of Predicted and Actual Survival Status

The model's accuracy, which measures the overall correctness of predictions, is calculated to be 14%. This indicates that the logistic regression model has performed poorly in predicting the outcomes for this dataset. The low accuracy suggests that the model's predictions align with the actual outcomes for only a tiny fraction of cases as the further validations, precision, recall, and F1 are performed as 0.09, 0.85, and 0.23, respectively. These metrics collectively suggest that the logistic regression model performs poorly in correctly identifying positive instances while minimizing false positives. Several key factors contribute to the poor performance of logistic regression models. First, the most essential patterns in the data may not have been correctly captured. In addition, the uneven distribution of classes in a dataset can make it difficult to accurately predict minority classes, leading to biased predictions and poor performance metrics such as sparseness and recall. Logistic regression also assumes a linear relationship between the independent variables and the log odds of the outcome, which may not apply to situations where the relationship is null. As a result, models can struggle to capture these complex relationships, leading to inaccurate predictions. Finally, because of its simplicity, logistic regression may not be suitable for data sets with complex relationships between variables, which makes practical issues difficult.

After we thoroughly examined the limitations, it became evident that the primary factor contributing to the model's poor performance was the presence of non-linear relationships between variables. Consequently, the model's predictive power diminishes, and its performance metrics, such as precision, recall, and accuracy, suffer due to the above.

CHAPTER FOUR: CONCLUSION

After conducting an extensive analysis to evaluate the performance of five distinct classification algorithms applied to a breast cancer dataset, the primary objective was to assess the accuracy of each algorithm in predicting the presence or absence of cancer, which is typically characterized as malignant or benign. This evaluation is crucial in informing clinical decision-making and improving patient outcomes in oncology. The classification algorithms employed in this analysis encompassed various methodologies, including decision trees, logistic regression, Naive Bayes, SVM, AdaBoost, and Bagging. Each algorithm was rigorously trained and validated using established machine-learning techniques to ensure robustness and generalizability.

After the results were carefully examined, it became evident that the accuracy of the classification algorithms varied significantly. The decision tree model exhibited an accuracy of 85%, a substantial improvement over the initially reported value. Conversely, the logistic regression model demonstrated a relatively low accuracy of 14%. These contrasting outcomes highlight the impact of algorithm choice and model complexity on predictive performance in breast cancer classification. Further analysis revealed that the Naive Bayes and SVM models achieved higher accuracy, reaching 83% and 89%, respectively. This notable performance enhancement underscores the efficacy of probabilistic methods like Naive Bayes and the capability of SVM to delineate nonlinear decision boundaries, contributing to their superior performance in distinguishing between malignant and benign tumors.

Of particular interest is the bagging algorithm, which emerged as the top performer among all models, boasting an impressive accuracy of 84%. Bagging, a powerful ensemble learning technique, leverages the collective wisdom of multiple base learners to enhance predictive performance. The exceptional accuracy attained by the bagging algorithm underscores the efficacy of ensemble methods in mitigating overfitting and capturing intricate patterns within the dataset. These findings offer valuable insights into the comparative effectiveness of various classification algorithms in breast cancer diagnosis. However, it is imperative to interpret these results within the context of the dataset's specific characteristics, including its size, class distribution, and potential biases. Future research endeavors may explore additional performance metrics such as precision, recall, F1-score, and ROC-AUC to comprehensively understand the models' performance and their clinical relevance in oncology practice.

The classification analysis was conducted using machine learning algorithms on a breast cancer dataset. This study's target variable of interest is the presence or absence of cancer, typically categorized as malignant or benign. It is important to note that survival status, as mentioned, may not directly apply to this classification problem, as it typically pertains to predicting patient survival after cancer treatment rather than the initial diagnosis.

Longitudinal studies that follow patients over an extended period are also essential for comprehending the dynamic nature of survival outcomes. Such studies can capture the impact of treatment changes, disease progression, and other time-dependent factors on a patient's survival months. By analyzing the trajectories of individual patients, researchers can discern patterns that were previously undetectable, providing valuable insights into the evolution of breast cancer and its impact on survival. Collaborative efforts among researchers, clinicians, and data scientists are vital for advancing our understanding of breast cancer prognosis. By pooling resources, expertise, and datasets, large-scale studies can be conducted to validate the findings of this research and explore additional variables of interest. Additionally, international collaborations can help overcome limitations in sample sizes and facilitate the generalizability of findings across diverse populations.

In examining breast cancer data focused on predicting survival status, we evaluated several classification algorithms, including logistic regression, decision trees, Naive Bayes, support vector machine (SVM), and bagging. While logistic regression is a conventional statistical tool for binary classification tasks, our analysis revealed its limitations in accurately capturing the intricate relationships inherent in breast cancer survival prediction. Logistic regression relies on the assumption of linear relationships between predictor variables and the log odds of the outcome, which may not adequately encapsulate the multifaceted dynamics involved in survival prediction.

Table 6

Performance Metric Summary of Method Learning Algorithms

Method	**Precision**	**Recall**	**F1**	**Accuracy**
Decision Tree	0.84	0.99	0.91	85%
Naïve Bayes	0.88	0.87	0.88	80%
SVM	0.85	0.99	0.91	84%
Bagging	0.86	0.97	0.91	84%
AdaBoost	0.84	0.97	0.90	85%
Logistic Regression	0.09	0.85	0.23	14%

In contrast, machine learning algorithms such as Decision Trees, SVM, AdaBoost, and Bagging demonstrated advantages in our analysis. These methods excel in handling nonlinear relationships, identifying intricate interactions between features, and providing robust predictions even in the presence of outliers and imbalanced datasets. Furthermore, advanced algorithms are adept at accommodating the multifaceted nature of survival prediction tasks by employing ensemble learning techniques, capturing subtle patterns within the data, and effectively addressing class imbalances. The superior performance exhibited by these machine learning methods compared to logistic regression underscores their potential to significantly enhance the accuracy and reliability of breast cancer survival prediction models.

In conclusion, our study on breast cancer survival prediction emphasized the importance of employing accurate classification algorithms. Through rigorous analysis, it became evident that machine learning methods, particularly bagging, outperformed logistic regression and other algorithms in accuracy. Bagging demonstrated the slightly highest perfection among all methods evaluated in our study, as its efficacy in accurately predicting breast cancer survival status. This finding highlights the critical role of advanced machine learning techniques in enhancing the accuracy and reliability of breast cancer survival prediction models. By highlighting trends and correlations that follow accepted statistical principles, statistical tests offer essential insights into the significance of interactions between variables. However, machine learning techniques do well when processing high-dimensional data or capturing intricate nonlinear relationships, where classic statistical approaches would struggle. By utilizing sophisticated algorithms, machine learning can reveal complex patterns and correlations in the data that may not be visible through traditional statistical analysis alone. Additionally, machine learning makes predictive modeling easier, making it possible to create reliable algorithms that can accurately classify data and estimate outcomes and trends.

REFERENCES

[1] Burkov, A. (2019). The Hundred-page Machine Learning Book.

[2] Adamson, A. S., &amp; Welch, H. G. (2019). Machine learning and the cancer-diagnosis problem — no gold standard. New England Journal of Medicine, 381(24), 2285–2287. https://doi.org/10.1056/nejmp1907407.

[3] Haug, C. J., &amp; Drazen, J. M. (2023). Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. New England Journal of Medicine, 388(13), 1201–1208. https://doi.org/10.1056/nejmra2302038.

[4] Wikimedia Foundation. (2023, May 5). Machine learning. Wikipedia. Retrieved May 5, 2023, from https://en.wikipedia.org/wiki/Machine_learning.

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785–794).

[7] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32.

[8] DeSantis, C. E., Ma, J., Goding Sauer, A., Newman, L. A., & Jemal, A. (2017). Breast cancer statistics, 2017, racial disparity in mortality by state. CA: A Cancer Journal for Clinicians, 67(6), 439-448.

[9] Harbeck, N., & Gnant, M. (2017). Breast cancer. The Lancet, 389(10074), 1134-1150.

[10] Li, J., & Xu, W. (2017). A review on mathematical modeling of breast cancer growth. Journal of Cancer Research and Therapeutics, 13(3), 421.

[11] DeSantis, C., Ma, J., Gaudet, M. M., Newman, L. A., Miller, K. A., Sauer, A. G., Jemal, A., & Siegel, R. L. (2019). Breast cancer statistics, 2019. CA: A Cancer Journal for Clinicians, 69(6), 438–451. https://doi.org/10.3322/caac.21583.

[12] Tang, Y., Yang, C., Su, S., Wang, W., Fan, L., & Shu, J. (2021). Machine learning-based Radiomics analysis for differentiation degree and lymphatic node metastasis of extrahepatic cholangiocarcinoma. BMC Cancer, 21(1). https://doi.org/10.1186/s12885-021-08947-6.

[13] Dietterich, T. G. (1999). An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization (D. Fisher, Ed.; pp. 1–22) [Journal-article]. https://csd.uwo.ca/~xling/cs860/papers/mlj-randomized-c4.pdf.

[14] Tripepi, G., Jager, K. J., Dekker, F. W., & Zoccali, C. (2008). Linear and logistic regression analysis. Kidney International, 73(7), 806–810. https://doi.org/10.1038/sj.ki.5002787. .

[15] Peterson, J. R., Cole, J. A., Pfeiffer, J. R., Norris, G. H., Zhang, Y., Lopez-Ramos, D., Pandey, T., Biancalana, M., Esslinger, H. R., Antony, A. K., & Takiar, V. (2023). Novel computational biology modeling system can accurately forecast response to neoadjuvant therapy in early breast cancer. Breast Cancer Research, 25(1). https://doi.org/10.1186/s13058-023-01654-z.

[16] Ramin, C., Veiga, L. H. S., Vo, J. B., Curtis, R. E., Bodelon, C., Bowles, E. J. A., Buist, D. S. M., Weinmann, S., Feigelson, H. S., Gierach, G. L., & De Gonzalez, A. B. (2023). Risk of second primary cancer among women in the Kaiser Permanente Breast Cancer Survivors Cohort. Breast Cancer Research, 25(1). https://doi.org/10.1186/s13058-023-01647-y.

[17] Klimov, S., Miligy, I. M., Gertych, A., Jiang, Y., Toss, M. S., Rida, P. C., Ellis, I. O., Green, A. R., Krishnamurti, U., Rakha, E. A., & Aneja, R. (2019b). A whole slide image-based machine learning approach to predict ductal carcinoma in situ (DCIS) recurrence risk. Breast Cancer Research, 21(1). https://doi.org/10.1186/s13058-019-1165-5.

[18] Hortobagyi, G. N., Stemmer, S. M., Burris, H. A., Yap, Y. S., Sonke, G. S., Hart, L. L., Campone, M., Petráková, K., Winer, E. P., Janni, W., Conte, P., Cameron, D., Pusztai, L., Arteaga, C. L., Zarate, J. E., Chakravartty, A., Taran, T., Gac, F. L., Serra, P., & O'Shaughnessy, J. (2022). Overall Survival with Ribociclib plus Letrozole in Advanced Breast Cancer. The New England Journal of Medicine, 386(10), 942–950. https://doi.org/10.1056/nejmoa2114663.

[19] Sutton, E. J., Onishi, N., Fehr, D., Dashevsky, B. Z., Sadinski, M., Pinker, K., Martinez, D. F., Brogi, E., Braunstein, L. Z., Razavi, P., El-Tamer, M., Sacchini, V., Deasy, J. O., Morris, E. A., & Veeraraghavan, H. (2020). A machine learning model that classifies breast cancer pathologic complete response on MRI post-neoadjuvant chemotherapy. Breast Cancer Research, 22(1). https://doi.org/10.1186/s13058-020-01291-w.

[20] Ming, C., Viassolo, V., Probst-Hensch, N., Chappuis, P. O., Dinov, I. D., & Katapodi, M. C. (2019). Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Research, 21(1). https://doi.org/10.1186/s13058-019-1158-4.

[21] Khan, M. M., Islam, S., Sarkar, S., Ayaz, F. I., Ananda, M. K., Tazin, T., Albraikan, A., & Almalki, F. A. (2022). Machine Learning Based Comparative Analysis for Breast Cancer Prediction. Journal of Healthcare Engineering, 2022, 1–15. https://doi.org/10.1155/2022/4365855.

[22] Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019). Machine Learning Classification Techniques for Breast Cancer Diagnosis. IOP Conference Series, 495, 012033. https://doi.org/10.1088/1757-899x/495/1/012033.

[23] Machine learning techniques to diagnose breast cancer. (2010, April 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/5478895.

[24] Smigal, C., Jemal, A., Ward, E. C., Cokkinides, V., Smith, R., Howe, H. L., & Thun, M. J. (2006). Trends in Breast Cancer by Race and Ethnicity: Update 2006. CA: A Cancer Journal for Clinicians, 56(3), 168–183. https://doi.org/10.3322/canjclin.56.3.168.

[25] Wikipedia contributors. (2024c, February 21). Support vector machine. Wikipedia. https://en.wikipedia.org/wiki/Support_vector_machine.

[26] Wang, Y., Yang, F., Zhang, J., Wang, H., Yue, X., & Liu, S. (2021). Application of artificial intelligence based on deep learning in breast cancer screening and imaging diagnosis. Neural Computing and Applications, 33(15), 9637–9647. https://doi.org/10.1007/s00521-021-05728-x.

[27] Zeebaree, D. Q. (2020). Machine learning and Region Growing for Breast Cancer Segmentation.www.academia.edu. https://www.academia.edu/43377759/Machine_learning_and_Region_Growing_for_Breast_Cancer_Segmentation.

[28] Yassin, N. I., Omran, S., Houby, E. M. F. E., & Allam, H. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. Computer Methods and Programs in Biomedicine, 156, 25–45. https://doi.org/10.1016/j.cmpb.2017.12.012.

[29] Wikipedia contributors. (2002, August 11). Breast cancer. Wikipedia. https://en.wikipedia.org/wiki/Breast_cancer.

[30] Warner, E. (2011). Breast-Cancer Screening. The New England Journal of Medicine, 365(11), 1025–1032. https://doi.org/10.1056/nejmcp1101540.

[31] Mellemkjær, L., Friis, S., Olsen, J. H., Scelo, G., Hemminki, K., Tracey, E., Andersen, A., Brewster, D. C., Pukkala, E., McBride, M. L., Kliewer, E. V., Tonita, J., Kee-Seng, C., Pompe-Kirn, V., Martos, C., Jonasson, J. G., Boffetta, P., & Brennan, P. (2006). Risk of second cancer among women with breast cancer. International Journal of Cancer, 118(9), 2285–2292. https://doi.org/10.1002/ijc.21651.

[32] Wang, Q. Q., Yu, S. C., Qi, X., Hu, Y. H., Zheng, W. J., Shi, J. X., & Yao, H. Y. (2019). Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine], 53(9), 955–960. https://doi.org/10.3760/cma.j.issn.0253-9624.2019.09.018.

[33] Smith, J. G., Whatley, P., & Redburn, J. (1998). Improving survival of melanoma patients in Europe since 1978. European Journal of Cancer, 34(14), 2197–2203. https://doi.org/10.1016/s0959-8049(98)00321-9.

[34] Han, J., & Kamber, M. (2012). Data mining: Concepts and Techniques.

[35] Sieuwerts, A. M., Willis, S., Burns, M. B., Look, M. P., Gelder, M. E. M., Schlicker, A., Heideman, M. R., Jacobs, H., Wessels, L. F. A., Leyland-Jones, B., Gray, K. P., Foekens, J. A., Harris, R. S., & Martens, J. W. (2014). Elevated APOBEC3B Correlates with Poor Outcomes for Estrogen-Receptor-Positive Breast Cancers. Hormones and Cancer, 5(6), 405–413. https://doi.org/10.1007/s12672-014-0196-8.

[36] Wikipedia contributors. (2024b, February 6). Naive Bayes classifier. Wikipedia. https://en.wikipedia.org/wiki/Naive_Bayes_classifier.

[37] Breast cancer - Symptoms and causes - Mayo Clinic. (2022, December 14). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470.

[38] Chua, M. H., Kim, D., Choi, J., Lee, N. G., Deshpande, V., Schwab, J., Lev, M. H., Gonzalez, R., Gee, M. S., & Do, S. (2022b). Tackling prediction uncertainty in machine learning for healthcare. Nature Biomedical Engineering, 7(6), 711–718. https://doi.org/10.1038/s41551-022-00988-x.

[39] Bromham, N., Schmidt-Hansen, M., Astin, M., Hasler, E., & Reed, M. W. (2017). Axillary treatment for operable primary breast cancer. The Cochrane Library, 2019(5). https://doi.org/10.1002/14651858.cd004561.pub3.

[40] Wikipedia contributors. (2024d, March 2). Logistic regression. Wikipedia. https://en.wikipedia.org/wiki/Logistic_regression.

[41] Gorringe, K. L., & Fox, S. B. (2017). Ductal Carcinoma In Situ Biology, Biomarkers, and Diagnosis. Frontiers in Oncology, 7. https://doi.org/10.3389/fonc.2017.00248.

[42] Bayman, E. O., & Dexter, F. (2021). Multicollinearity in Logistic Regression Models. Anesthesia and analgesia, 133(2), 362–365. https://doi.org/10.1213/ANE.0000000000005593.

[43] SEER*Stat Databases: November 2020 submission. (n.d.). SEER. https://seer.cancer.gov/data-software/documentation/seerstat/nov2020/.

[44] Dhahri, H., Rahmany, I., Mahmood, A., Maghayreh, E. A., & Elkilani, W. S. (2020). Tabu Search and Machine-Learning Classification of Benign and Malignant Proliferative Breast Lesions. BioMed Research International, 2020, 1–10. https://doi.org/10.1155/2020/4671349.

[45] Xu, P., Xu, R., Yang, Q., & Zhu, H. (2022). Effect of Intensive Psychological Care on Patients with Benign Breast Lumps after Mammotome-Assisted Tumor Resection. Evidence-based Complementary and Alternative Medicine, 2022, 1–6. https://doi.org/10.1155/2022/9054266.

[46] Turner, N. J., Slamon, D. J., Ro, J., Bondarenko, I., Im, S., Masuda, N., Colleoni, M., DeMichele, A., Loi, S., Verma, S., Iwata, H., Harbeck, N., Loibl, S., Pusztai, L., Theall, K. P., Huang, X., Giorgetti, C., Bartlett, C. H., & Cristofanilli, M. (2018). Overall Survival with Palbociclib and Fulvestrant in Advanced Breast Cancer. The New England Journal of Medicine, 379(20), 1926–1936. https://doi.org/10.1056/nejmoa1810527.

[47] Lobbezoo, D. J., Van Kampen, R. J., Voogd, A. C., Dercksen, M., Van Den Berkmortel, F. W. P. J., Smilde, T. J., Van De Wouw, A. J., Peters, F. H., Van Riel, J. M. G. H., Peters, N. a. J. B., De Boer, M., Borm, G. F., & Tjan-Heijnen, V. C. G. (2013). Prognosis of metastatic breast cancer subtypes: the hormone receptor/HER2-positive subtype is associated with the most favorable outcome. Breast Cancer Research and Treatment, 141(3), 507–514. https://doi.org/10.1007/s10549-013-2711-y.

[48] Jung, S. Y., Rosenzweig, M., Sereika, S. M., Linkov, F., Brufsky, A., & Weissfeld, J. L. (2011). Factors associated with mortality after breast cancer metastasis. Cancer Causes & Control, 23(1), 103–112. https://doi.org/10.1007/s10552-011-9859-8.

[49] Wikipedia contributors. (2023). Artificial intelligence in healthcare. Wikipedia. https://en.wikipedia.org/wiki/Artificial_intelligence_in_healthcare.

[50] Feature selection and classification of breast cancer diagnosis based on support vector machines. (2008, August 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/4631603.

[51] Parkin, D. M. (1998). Epidemiology of cancer: global patterns and trends. Toxicology Letters, 102–103, 227–234. https://doi.org/10.1016/s0378-4274(98)00311-7.

[52] Declining Breast Cancer Mortality Among Young American Women. (1987). Journal of the National Cancer Institute. https://doi.org/10.1093/jnci/78.3.451.

[53] Kandati, D. R., & Gadekallu, T. R. (2023). A Comprehensive Survey on Federated Learning Techniques for Healthcare Informatics. Computational Intelligence and Neuroscience, 2023, 1–19. https://doi.org/10.1155/2023/8393990.

[54] SSVM: a simple SVM algorithm. (2002). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/1007516-

[55] Winters-Hilt, S., & Merat, S. (2007). SVM clustering. BMC bioinformatics, 8 Suppl 7(Suppl 7), S18. https://doi.org/10.1186/1471-2105-8-S7-S18.

[56] Hu, C., Hart, S. N., Gnanaolivu, R., Huang, H., Lee, K. C., Na, J., Gao, C., Lilyquist, J., Yadav, S., Boddicker, N. J., Samara, R., Klebba, J., Ambrosone, C. B., Anton-Culver, H., Auer, P. L., Bandera, E. V., Bernstein, L., Bertrand, K. A., Burnside, E. S., . . . Couch, F. J. (2021). A Population-Based Study of Genes Previously Implicated in Breast Cancer. The New England Journal of Medicine, 384(5), 440–451. https://doi.org/10.1056/nejmoa2005936.

[57] Can Artificial Intelligence Help See Cancer in New Ways? (2022, March 22). National Cancer Institute. https://www.cancer.gov/news-events/cancer-currents-blog/2022/artificial-intelligence-cancer-imaging.

[58] Ytterberg, S. R., Bhatt, D. L., Mikuls, T. R., Koch, G. G., Fleischmann, R., Rivas, J. L., Germino, R., Menon, S., Sun, Y., Wang, C., Shapiro, A. B., Kanik, K. S., & Connell, C. A. (2022). Cardiovascular and Cancer Risk with Tofacitinib in Rheumatoid Arthritis. The New England Journal of Medicine, 386(4), 316–326. https://doi.org/10.1056/nejmoa2109927.

[59] Schoenthaler, R. (2023). The Breast Biopsy and the Buddhist Half-Smile. The New England Journal of Medicine, 388(18), 1642–1643. https://doi.org/10.1056/nejmp2203336.

[60] Adamson, A. S., & Welch, H. G. (2019). Machine Learning and the Cancer-Diagnosis Problem — No Gold Standard. The New England Journal of Medicine, 381(24), 2285–2287. https://doi.org/10.1056/nejmp1907407.

[61] Rajkomar, A., Dean, J., & Kohane, I. S. (2019). Machine Learning in Medicine. The New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/nejmra1814259.

[62] Chen, J. M., & Asch, S. M. (2017b). Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. The New England Journal of Medicine, 376(26), 2507–2509. https://doi.org/10.1056/nejmp1702071.

[63] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. The New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/nejmp1606181.

[64] SVM kernel functions for classification. (2013, January 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/6524743.

[65] Wikipedia contributors. (2024e, March 6). AdaBoost. Wikipedia. https://en.wikipedia.org/wiki/AdaBoost.

[66] Chalup, S. (2005). Predicting Foreign Exchange Rate Return Directions with Support Vector Machines. In Simeon J. Simoff, Graham J. Williams, John Galloway, & Inna Kolyshkina (Eds.), Proceedings of the 4th Australasian Data Mining Conference (Vols. 5–6) [Conference-proceeding]. University of Technology Sydney. https://www.researchgate.net/profile/Stephan-Chalup/publication/235962509_Predicting_Foreign_Exchange_Rate_Return_Directions_with_Support_Vector_Machines/links/0f317538398a47612b000000/Predicting-Foreign-Exchange-Rate-Return-Directions-with-Support-Vector-Machines.pdf#page=

[67] Flach, P. A., & Lachiche, N. (2004). Naive Bayesian classification of structured data. Machine Learning, 57(3), 233–269. https://doi.org/10.1023/b:mach.0000039778.69032.ab

[68] Machine learning for improved clinical management of cancers of unknown primary. (2023). Nature medicine, 29(8), 1920–1921. https://doi.org/10.1038/s41591-023-02501-6.

[69] Akbulut, S., Küçükakçalı, Z., & Çolak, C. (2023). Predicting the Risk of Duodenal Cancer in Patients with Familial Adenomatous Polyposis Using a Machine. The Turkish journal of gastroenterology : the official journal of Turkish Society of Gastroenterology, 10.5152/tjg.2023.22346. Advance online publication. https://doi.org/10.5152/tjg.2023.22346

[70] And Alternative Medicine E. C. (2023). Retracted: An Improved Machine Learning Model for Diagnostic Cancer Recognition Using Artificial Intelligence. Evidence-based complementary and alternative medicine : eCAM, 2023, 9873948. https://doi.org/10.1155/2023/9873948.

[71] Wikipedia contributors. (2024a, February 3). Decision tree. Wikipedia. https://en.wikipedia.org/wiki/Decision_tree.

[72] A review of supervised machine learning algorithms. (2016, March 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/7724478.

[73] Davies, B. K., Hibbert, A. P., Roberts, S. J., Roberts, H. C., Tickner, J. C., Holdsworth, G., Arnett, T. R., & Orriss, I. R. (2023). A Machine Learning-Based Image Segmentation Method to Quantify In Vitro Osteoclast Culture Endpoints. Calcified tissue international, 10.1007/s00223-023-01121-z. Advance online publication. https://doi.org/10.1007/s00223-023-01121-z.

[74] Gomez-Zaragoza, L., Marin-Morales, J., Vargas, E. P., Giglioli, I. A. C., & Raya, M. A. (2023). An Online Attachment Style Recognition System Based on Voice and Machine Learning. IEEE journal of biomedical and health informatics, PP, 10.1109/JBHI.2023.3304369. Advance online publication. https://doi.org/10.1109/JBHI.2023.3304369.

[75] Weissman, G. E., & Joynt Maddox, K. E. (2023). Guiding Risk Adjustment Models Toward Machine Learning Methods. JAMA, 10.1001/jama.2023.12920. Advance online publication. https://doi.org/10.1001/jama.2023.12920

[76] Zehra, S. S., Turabee, Z., & Rawalia, M. A. (2023). Early but Quality Diagnosis: On Breast Cancer and Its Risk Factors [Letter]. Breast cancer (Dove Medical Press), 15, 549–550. https://doi.org/10.2147/BCTT.S431476.

[77] Healthcare Engineering J. O. (2023). Retracted: Machine Learning Based Comparative Analysis for Breast Cancer Prediction. Journal of Healthcare Engineering, 2023, 9870523. https://doi.org/10.1155/2023/9870523.

[78] Riedl, R., Brandstätter, E. & Roithmayr, F. Identifying decision strategies: A process- and outcome-based classification method. Behavior Research Methods 40, 795–807 (2008). https://doi.org/10.3758/BRM.40.3.795

[79] And Alternative Medicine E. C. (2023). Retracted: An Improved Machine Learning Model for Diagnostic Cancer Recognition Using Artificial Intelligence. Evidence-based complementary and alternative medicine: eCAM, 2023, 9873948. https://doi.org/10.1155/2023/9873948.

[80] Boulesteix, A., & Schmid, M. (2014). Machine learning versus statistical modeling. Biometrical Journal, 56(4), 588–593. https://doi.org/10.1002/bimj.201300226.

[81] Alessy, S. A., Alhajji, M., Rawlinson, J., Baker, M., & Davies, E. A. (2022). Factors influencing cancer patients' experiences of care in the USA, United Kingdom, and Canada: A systematic review. EClinicalMedicine, 47, 101405. https://doi.org/10.1016/j.eclinm.2022.101405

[82] Shigei, N., Miyajima, H., Maeda, M., & Ma, L. (2009). Bagging and AdaBoost algorithms for vector quantization. Neurocomputing, 73(1–3), 106–114. https://doi.org/10.1016/j.neucom.2009.02.020

[83] Tufail, M., Hu, J. J., Liang, J., He, C. Y., Wan, W. D., Huang, Y. Q., Jiang, C. H., Wu, H., & Li, N. (2024). Predictive, preventive, and personalized medicine in breast cancer: targeting the PI3K pathway. Journal of translational medicine, 22(1), 15. https://doi.org/10.1186/s12967-023-04841-w.

[84] Ahuja A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. Peer J, 7, e7702. https://doi.org/10.7717/peerj.7702.

[85] Kalinli, A., Sarikoc, F., Akgun, H., & Ozturk, F. (2013). Performance comparison of machine learning methods for prognosis of hormone receptor status in breast cancer tissue samples. Computer methods and programs in biomedicine, 110(3), 298–307. https://doi.org/10.1016/j.cmpb.2012.12.005

[86] 꽁냥이부. (2022b, November 5). 15. AdaBoost(Adaptive Boost) 알고리즘에 대해서 알아보자 with Python. 부자 되고픈 꽁냥이. https://zephyrus1111.tistory.com/195

[87] Sungkee. (2021, June 28). [머신러닝] 앙상블 학습 - 2) Bagging. 책 읽는 성키. https://sungkee-book.tistory.com/9