A picture containing text, clock, sign

Description automatically generated

**ADTA 5910: Analytics Capstone Experience**

**Dr. Philpot Denise**

**Complete Draft of Project paper**

**Team Members**

**Rahul Nalluri**

**Dileep Kumar Yarra**

**Anvesh Konduru**

**Table of Contents**

**Chapters:**

1. **Introduction** …......................................................................................4

Dataset Description ........................................................................................4

Research Questions …..................................................................................6

1. **Literature Review** …..........................................................................................7

Introduction.......................................................................................................7

Conclusion........................................................................................................11

1. **Methodology** ………………………………………….......................................12

Software and tools.........................................................................................12

Python..........................................................................................................12

Numpy Pandas ….......................................................................................12

Scikit-Learn …...........................................................................................13

Jupyter Notebook …...................................................................................13

Exploratory Data Analysis ….....................................................................14

Cleaning the data.........................................................................................15

Data Visuvalisation ….................................................................................15

Numerical Variable Plots …........................................................................15

Categorical Variables Plots ….....................................................................18

Insights from Correlation Matrix …............................................................22

BoxplotsInsights ….....................................................................................23

1. **Modeling the Data** ….......................................................................................................33

K-means clustering ….....................................................................................33

Regression Model Building….........................................................................36

Decision Tree Classifier…...............................................................................42

Random Forest Classifier …............................................................................43

1. **Summary**…........................................................................................................................44

**References** ….......................................................................................................................48

**Appendix – Code** …............................................................................................................40

**Table of Figures**

Figure1: overview of Data columns............................................................................15

Figure2: Histogram of age.............................................................................................16

Figure3: Histogram of Tumor size...............................................................................17

Figure4: Histogram of Regional node Examined. ...................................................17

Figure5: Histogram of regional node positive Distribution....................................18

Figure6: Histogram of Survival months distribution................................................18

Figure7: categorical Variable plot of race distribution.............................................19

Figure8: categorical Variable plot of martial status distribution............................20

Figure9: categorical Variable plot of T stage distribution........................................21

Figure10: categorical Variable plot of N stage distribution.....................................22

Figure11: categorical Variable plot of Grade distribution........................................22

Figure12: categorical variable plot of Estrogen and progesterone distribution...23

Figure13: categorical variable plot of Estrogen status...............................................23

Figure14: Correlation matrix of Numerical variables................................................24

Figure15: Box plot of Tumor size vs T stage.................................................................25

Figure16: Box plot of Tumor size vs N stage.................................................................26

Figure17: Box plot of Tumor size vs Grade....................................................................27

Figure18: Box plot of Tumor size vs Estrogen status...................................................28

Figure19: Violin plots of numerical variables vs 6th stage.........................................29

Figure20: Violin plots of regional node examined vs 6th stage.................................30

Figure21: Violin plots of survival months vs 6th stage................................................31

Figure22: Violin plots of regional node positive vs 6th stage.....................................32

Figure23: Violin plots of regional node examined vs 6th stage..................................32

Figure24: Violin plots of survival months vs 6th stage.................................................33

Figure25: cumulative distribution function of Age.........................................................34

Figure26: Density curve of Tumor size..............................................................................35

Figure27: Random feature Importance graph....................................................................43

Figure28: significance of variables in Decision tree model...........................................45

Figure29: Significance of variables in Random Forest classifier.................................46

**Chapter 1**

**Introduction:**

Breast cancer is a disorder in which abnormal breast cells multiply uncontrolled, resulting in tumors. If the tumors are not treated, they might spread throughout the body and result in death. Breast cancer cells first manifest in the milk ducts and/or milk-producing lobules of the breast. There is no danger to life during the early period. Adjacent breast tissue might contract malignant cells. As a result, tumors are created, resulting in lumps or thickening. Invasive cancers have the potential to spread to nearby lymph nodes or other organs. Death might arise from metastasis. Treatment is based on the patient, the type of cancer, and how far advanced it is. The treatment includes surgery, radiation therapy, and medication. The design and analysis of epidemiologic research, investigations using animal testing, and mechanistic studies of breast cancer biology are beginning to be influenced by innovative hypotheses that attempt to explain how environmental influences may alter the causes of breast cancer. The limits of the present understanding of breast cancer are highlighted by the fresh insights on the condition. The data, collected by the SEER Program of NCI in 2017, comprises information from the years 2006 to 2010.

**Dataset Description:**

This Dataset contains 4024 observations and 15 variables. There are 5 nominal variables, 5 ordinal variables, 5 numerical variables in the dataset. Dataset variable descriptions and datatypes are shown below. In the dataset we can see patients age range from 30 to 69.

AGE(Numerical): This variable represents the patient's age at the time of diagnosis, given in years.

RACE(Nominal): This variable classifies patients into many racial categories, such as White, Black, Other (which includes American Indian/AK Native and Asian/Pacific Islander).

MARITAL STATUS(Nominal): The patients' marital status at the time of diagnosis is determined by this variable, which includes characteristics such as Single, Married (including common law), Separated, Divorced, and Widowed.

T STAGE(Ordinal): This factor is related to the size of the tumor and the degree of primary tumor involvement. It uses categories T1, T2, T3, and T4.

T1: Tumor ≤ 20 mm.

T2: Tumor > 20 mm and ≤ 50 mm.

T3: Tumor > 50 mm.

T4: Any size tumor that has a direct extension to the skin or the chest wall.

N STAGE(Ordinal): This variable represents the extent of regional lymph node involvement and uses categories N1, N2, and N3.

N1: The cancer has reached 1 to 3 axillary lymph nodes.

N2: The cancer has reached 4 to 9 axillary lymph nodes under the arm.

N3: The cancer has reached 10 or more axillary lymph nodes.

6TH STAGE(Ordinal): This variable uses categories to characterize the cancer stage according to the AJCC 6th Edition staging method(IIA, IIB, IIIA, IIIB, and IIIC).

GRADE(Ordinal): It gives details on the grade and differentiation of cancer cells, with classifications ranging from Grade I to Grade IV.

A STAGE(Ordinal): This variable shows the cancer's historical stage, which is divided into two categories: regional and distant.

TUMOR SIZE(Numerical): information on the tumor's exact size in millimeters.

ESTROGEN STATUS(Nominal): Detects if the breast cancer has estrogen receptors that are positive or negative.

PROGESTERONE STATUS(Nominal): Determines if the breast cancer has progesterone receptors that are positive or negative.

REGIONAL NODES EXAMINED(Numerical): The total number of regional lymph nodes that the pathologist examined.

REGIONAL NODES POSITIVE(Numerical): The exact amount of local lymph nodes that the pathologist examined and discovered to have tumors.

SURVIVAL MONTHS(Numerical): Provides data on patient survival time, estimated using entire dates, including days.

STATUS(Nominal): Indicates the patients' vital state as of the research cutoff date, as either Alive or Dead.

**Research Questions:**

1. **Does size of the tumor depends on the Age, Race, A stage, Regional nodes examined, Regional nodes positive?**

Linear Regression can be used to model the relationship between Age, Race, A stage, Regional Nodes examined, Regional nodes positive (Independent variables) and Tumor size (Dependent Variable). Most important factor for determining tumor size are Regional nodes examined and Regional nodes positive.

1. **Can we cluster breast cancer patients based on their characteristics such as Age, Race, A size, Tumor size, Regional nodes examined, and Regional nodes positive?**

K-means clustering can be employed to group breast cancer patients based on their tumor characteristics, including N stage, A stage, tumor size, estrogen status, progesterone status, regional nodes examined, and regional nodes positive. The optimal number of clusters (K) can be determined using methods like the elbow method or silhouette analysis.

1. **Can we predict the “A Stage” of breast cancer based on patient demographics and tumor characteristics?**

This research question focuses on predicting the “A stage” of breast cancer using features such as patient age, race, marital status, tumor size, T stage, N stage, grade, estrogen status, and progesterone status. This classification problem is specific, feasible, and complex enough to merit a detailed answer, as it requires analyzing multiple variables and their relationships to predict the breast cancer stage.

**Chapter 2**

**Literature Review:**

**Introduction:**

The most common cancer in women is breast cancer. It is the second most frequent cancer-related mortality among women worldwide. Breast cancer develops slowly, and the majority of cases are found through routine screenings. In addition to reviewing the causes, presentation, and diagnosis of breast cancer, this exercise also emphasizes the management of the disease's multidisciplinary team. Artificial Intelligence(AI) and Machine Learning(ML) have played a role, in advancing breast cancer research, diagnosis, prognosis and treatment. These innovative technologies are enhancing the precision, effectiveness and individualized care provided for breast cancer patients.

The Surveillance, Epidemiology and End Results (SEER) Program provides information on cancer statistics in an effort to decrease the burden of cancer among Americans. SEER is funded by the Division of Cancer Control and Population Sciences (DCCPS) of the NCI's Surveillance Research Program (SRP). SEER breast cancer datasets are mostly commonly used in breast cancer studies which uses ML techniques to detection and diagnosis.

This review's main focus is on factors that causes the breast cancer. The review emphasizes the AI and ML techniques for preventing, finding, and diagnosing breast cancer patients. This review examines the previous works done on breast cancer diagnosis. This review can help in what extra work need to be done using machine learning with new factors that causes breast cancer.

In previous studies (Momenimovahed & Salehiniya, 2019) genetics, lifestyle and environment are the various factors that mostly causes breast cancer. Moreover, literature concludes lactation, parities and exercise are crucial in lowering risk of the breast cancer. This research focuses on geographical and socioeconomic factors. It also offers a viewpoint on the increasing incidence of breast cancer while observing a decline in mortality rates.

There are limitations in ML algorithms. Recent research is directed towards the Deep Learning from ML (Shah et al., 2022). Deep Learning algorithms have the ability to extract all features from input data. In the past, pathologists and radiologists would manually review breast images and draw findings after consulting with other medical specialists. Observing several images is a time-consuming process which can give false outcomes. To observe large number of images that are generated there is a need for AI based automated breast cancer systems.

In an article (Watkins, 2019) it is discussed that breast cancer is a diverse disease that necessitates the expertise of healthcare professionals, from various disciplines to effectively diagnose and treat. Healthcare providers working outside oncology settings would be able to provide care for patients, in both the stages of diagnosis and later stages of treatment if they possess a thorough understanding of the latest research findings. Breast cancer is, on the rise among all groups in the United States and this includes an increase, in cases of estrogen positive breast cancer.

Triple negative breast cancer (TNBC) constitutes 15-20% of all diagnosed breast cancer cases and is the sole subtype that currently lacks specific treatment options. While TNBC may affect anybody, it more frequently affects young and Black women is mentioned in the article (Zagami & Carey, 2022) and is associated with a number of negative consequences, therefore we must consider all costs associated with the illness in addition to the requirement for better and more accurate treatment. New technologies may enable the targeting of tumors, blocking of immune responses, and perhaps even the eradication of TNBC. The priority must continue to be on enhancing toxicity management, providing social and financial support to persons with TNBC, and enhancing therapy.

The researchers carried out a thorough investigation titled "Comparing Endocrine Therapy Adherence in Male and Female Breast Cancer: An Investigation using SEER-Medicare Data." With the use of a sizable dataset, this study aimed to examine the endocrine therapy (ET) adherence rates and discontinuation patterns of male and female breast cancer patients. It's interesting to note that a sizable portion of male patients did not receive adjuvant endocrine therapy even though male breast cancer is typically hormone receptor-positive Azka et al. (2022). Further research revealed that while both sexes experienced equal rates of treatment discontinuation, men showed higher adherence to ET than women. The investigation also highlighted the beneficial effects of ET on breast cancer survivorship outcomes for both male and female patients.

The temporal changes in racial and ethnic differences in breast cancer survival were examined in the current study (Hill et al., 2023). According to the findings, these differences date back to the 1970s, before breast cancer screening and technological advances in treatment were introduced. It was noted that Black women consistently outlived White non-Hispanic women by a margin of 10–12 percentage points.

In the study by Alsumai et al. (2023), the investigators investigated characteristics that predicted clinically node-negative breast cancer patients to have a positive sentinel lymph node biopsy (SLNB). They discovered that a variety of variables, such as the presence of more SLNs, greater tumor sizes, grade I malignancy, the presence of ER-positive/PR-positive/HER2-negative status, the prominence of the axillary lymph nodes on preoperative imaging, and lymphovascular invasion, were independent predictors of SLN positivity. Making informed therapy decisions may be made easier as a result of these findings, which provide insightful information about identifying patients at risk for positive SLNB.

According to American statistics on breast cancer, there were predicted to be 40,290 breast cancer deaths and around 231,840 new cases of invasive breast cancer among women in 2015. While breast cancer death rates significantly decreased between 1989 and 2012, inequities still exist, with black women's mortality rates being 42% higher than those of white women (CA Cancer J Clin 2016;31–42). These racial discrepancies are anticipated to persist, particularly in Southern states, according to rising incidence rates in black women. These results highlight the need to encourage healthy habits and routine mammography screening.

The predictive value of various lymph node characteristics in breast cancer patients involved in Austrian Breast and Colorectal Cancer Study Group (ABCSG) trials was investigated in this study by Tausch et al. (2011). The study, which was based on a cohort of 7,052 breast cancer patients, concentrated on variables like the lymph node ratio (LNR), the number of excised axillary lymph nodes (NRN), and the number of nodes that were involved (NIN). It was discovered that patients who underwent mastectomy and had one to three positive lymph nodes had worse recurrence-free survival (RFS) and overall survival (OS) scores as NIN and LNR increased. With NIN demonstrating high connections, the study sheds light on the significance of lymph node status as a predictive factor in breast cancer.

Mammograms and ultrasound imaging are the principal methods used by radiologists for detecting and diagnosing breast cancer as mentioned in the paper (Dabass et al., 2019). The accuracy with which the segmentation is performed is crucial in establishing whether the tumor is benign or malignant. Hybrid techniques can be used to increase segmentation accuracy. Hybrid techniques may be employed to produce ultrasound and thermography breast cancer images that are less hazardous to various elements than mammography.

**Conclusion:**

Despite the fact that the use of ML approaches to analyze SEER data is still in its infancy, the articles we have looked at suggest potential directions. The domains where ML has initially demonstrated effectiveness using SEER data include modeling of tumor development trajectories, categorization of prognostic signs, prediction of survival outcomes, and better therapy recommendations. There is still a lot of research to be done on the creation and assessment of reliable ML pipelines for cancer registry analytics.

The use of SEER data has also allowed researchers to examine how changes in sexual behavior and breast cancer screening programs have changed incidence over time. This database's current investigations will help reveal epidemiologic trends and knowledge gaps. Priority challenges include tracking incidence and survival disparities, evaluating the efficacy of screening, and linking SEER to other data sources.

This study highlights the importance of having access to high-quality cancer registry data for finding and comprehending patterns and trends in breast cancer in the US population. The SEER database has given us information we may use to develop public health initiatives that can lower deaths and has assisted us in better understanding the epidemiology of breast cancer. As risk factors and treatments change, ongoing observation and thorough study will remain essential.

**Chapter 3**

**Methodology:**

**Software and Tools:**

In the face of large data, choosing proper software becomes an important issue. The selected software needs to be able to process large amounts of data fast and accurately. Our research relies on a set of software tools and libraries that each have an important role of enabling data preparation and initial explorations. These instrumental tools include:

**1. Python:**

Python, an agile language, with high repute in handling data, is our primary choice. Python is our preferred language for several compelling reasons:

User-Friendly Syntax: Python language comes in handy for data analysts and scientists in such a way that its syntax is straightforward and obvious for the users.

Rich Library Ecosystem: The wide range of tools within Python’s extensive library ecosystem makes it possible for us to manipulate and analyze complex datasets efficiently.

Community Support: We have chosen Python because of its open source nature and community oriented development process. This ensures availability of resources, documentation and help from the community that all go together in building a solid case for using python here.

**2. Numpy and Pandas:**

Furthermore, we also deploy the power of the Numpy and Pandas libraries in order to maximize our already enhanced data processing skills. They have strong features such as array operation functions and data analysis capabilities that have simplified analysis of complicated data sets leading to gaining meaning results.

Numpy: To deal with huge datasets, manage arrays and perform numerical tasks effectively, Numpy is required.

Pandas: Pandas, on the other hand, makes it easier for a programmer to deal with raw datasets by offering a user-friendly DataFrame structure which enables quicker data indices, searches, and analyses. It is an essential part of our toolbox when dealing with tasks on structured data.

**3. Scikit-Learn (Sklearn):**

We have a vital component in our toolkit called sklearn that covers all aspects of machine learning. This includes many predictive modeling as well as methods for evaluating models and performance parameters of performance monitoring. Using Sklearn, we can rapidly learn, design, test, and evaluate sophisticated machine learning algorithms that aid us in forecasting and finding patterns in our data.

**4. Jupyter Notebook:**

For our multi-dimensional IDE to support data analysis and develop an interactive and versatile environment, we use Jupyter notebook. We can use Jupyter Notebook as an open-source web application that allows writing and executing code besides, data visualisation, documenting our analysis process and collaboration with other persons. Due to its flexibility, interactivity and simplicity in operation this makes a very useful tool for analysis of data and scientific calculations.

**Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data that might be unexpected.

EDA is an important first step in any data analysis. Understanding where outliers occur and how variables are related can help one design statistical analyses that yield meaningful results.

It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning. The below figure describes high-level overview of our data.

**A screenshot of a computer

Description automatically generated**

*Figure1: overview of Data columns*

**Cleaning the data:**

Data cleaning is a critical step in machine learning (ML) that involves identifying and correcting errors, inconsistencies, and inaccuracies in data. By cleaning data, we can improve its quality, enhance model performance, reduce bias, save time and resources. data cleaning can help in creating better ML models that provide accurate results.

Our dataset doesn't have any missing values, and we decided to get rid of the "status" column because it's not needed.

**Data Visualization:**

**Numerical Variable Plots:**

**Age Distribution:**

A graph with green and black lines

Description automatically generated

*Figure2: Histogram of age*

The histogram for age shows a right-skewed distribution, indicating a younger population of patients with a smaller number of older patients.

The presence of outliers on the upper end of the age range is visible in the boxplot, which suggests that there are a few patients who are significantly older than the average patient population.

**Tumor Size Distribution:**

A graph with a green bar and a white bar

Description automatically generated with medium confidence

*Figure3: Histogram of Tumor size*

Tumor size has a right-skewed distribution, with most tumors being on the smaller end of the scale, and a few larger tumors representing outliers.

The boxplot shows a median tumor size that is relatively low compared to the maximum, highlighting the presence of large tumors in a small number of patients.

**Regional Node Examined Distribution:**

A graph with green and white squares

Description automatically generated with medium confidence

*Figure4: Histogram of Regional node Examined.*

The histogram for the number of regional nodes examined appears to be right-skewed, indicating that most patients had a lower number of nodes examined.

The boxplot reveals outliers, suggesting that in some cases, an unusually high number of nodes were examined.

**Regional Node Positive Distribution:**

A graph with a green bar

Description automatically generated

*Figure5: Histogram of regional node positive Distribution*

The distribution of positive regional nodes is highly skewed to the right, with many patients having few to no positive nodes and a small number having a high number.

Outliers in the boxplot indicate some patients had a very high count of positive nodes, which may be indicative of more advanced disease.

**Survival Months Distribution:**

A graph and a graph

Description automatically generated with medium confidence

*Figure6: Histogram of Survival months distribution*

The distribution of survival months is also right-skewed, suggesting that a large number of patients survived for a shorter period post-diagnosis, with fewer patients surviving for a longer time.

There are outliers in the survival months as shown by the boxplot, indicating some patients have a much longer survival time compared to the majority

**Categorical Variable Plots:**

**Race Distribution:**

A graph with green squares

Description automatically generated

*Figure7: categorical Variable plot of race distribution*

The majority of patients belong to the "White" race category, followed by "Black" and "Other" categories, indicating a dataset possibly reflective of the demographic distribution in the study area.

This could have implications for the generalizability of any predictive models developed from this data to other populations.

**Marital Status Distribution:**A graph of a number of people

Description automatically generated

*Figure8: categorical Variable plot of martial status distribution*

Most patients are married, which could suggest the need for family counseling and support services as part of cancer care.

The number of single and divorced patients also highlights the need for diverse social support structures in patient care.

**T Stage Distribution:**

A graph with different colored squares

Description automatically generated

*Figure9: categorical Variable plot of T stage distribution*

The most common T Stage is T1, followed by T2, suggesting that a significant number of tumors were detected at a smaller size.

Advanced T Stages (T3 and T4) are less common, which might reflect early detection or the nature of the patient population.

**N Stage Distribution:**

N0 is the most common N Stage, meaning many patients did not have spread to regional lymph nodes at the time of diagnosis.

N1, N2, and N3 stages are progressively less common, indicating fewer cases with lymph node involvement.

A graph with green squares

Description automatically generated

*Figure10: categorical Variable plot of N stage distribution*

**Grade Distribution:**

A graph of a number of children

Description automatically generated with medium confidence

*Figure11: categorical Variable plot of Grade distribution*

Most tumors are graded as "Moderately differentiated; Grade II", with "Poorly differentiated; Grade III" being the next most common. The presence of different grades suggests variability in tumor aggressiveness within the patient population.

**Estrogen and Progesterone Status Distribution:**

A graph showing a number of numbers

Description automatically generated

*Figure12: categorical variable plot of Estrogen and progesterone distribution*

A graph with green squares

Description automatically generated

*Figure13: categorical variable plot of Estrogen status*

A high number of tumors are positive for estrogen and progesterone status, which may have implications for treatment as hormone receptor-positive cancers may respond to hormonal therapies.

**Insights from Correlation Matrix:**

A blue and red squares with white text

Description automatically generated

*Figure14: Correlation matrix of Numerical variables*

The heatmap indicates various degrees of correlation between numerical variables. For instance, the number of nodes examined and the number of positive nodes show a positive correlation, which could be expected as more nodes examined increases the chance of finding positive ones.

Certainly, here are the insights for each boxplot comparing "Tumor Size" with selected categorical variables:

**Boxplot Insights:**

**Tumor Size vs. T Stage:**

A diagram of a diagram

Description automatically generated with medium confidence

*Figure15: Box plot of Tumor size vs T stage*

The boxplot for "Tumor Size" across different "T Stages" shows an increasing trend in the median tumor size as the T Stage increases. This is expected as the T Stage directly corresponds to the size and extent of the primary tumor.

There is a noticeable jump in tumor size from T1 to T2, and from T2 to T3. The variation in tumor sizes also increases with the stage, indicated by the larger interquartile ranges (IQR) for higher T Stages.

Outliers are present across all stages but are particularly notable in T2 and T3, suggesting some patients have tumor sizes that are exceptionally large for their stages.

**Tumor Size vs. N Stage:**

A diagram of a box plot

Description automatically generated with medium confidence

*Figure16: Box plot of Tumor size vs N stage*

This boxplot likely reflects the relationship between the size of the primary tumor and the extent of regional lymph node involvement. As "N Stage" increases, indicating more lymph node involvement, there may be a slight trend of increasing tumor size, but this is not as pronounced as with the T Stage.

The overlap in the IQR for the different N Stages suggests that lymph node involvement does not have as direct a correlation with primary tumor size as the T Stage does.

Outliers are present, particularly in the N0 stage, which may indicate cases where the primary tumor is large but has not yet spread to the lymph nodes.

**Tumor Size vs. Grade:**

A graph of a box diagram

Description automatically generated with medium confidence

*Figure17: Box plot of Tumor size vs Grade*

Tumor grade, which is an indication of how much tumor cells differ from normal cells, shows a relationship with tumor size. Higher grades, which indicate more abnormal and possibly more aggressive cells, generally have larger median tumor sizes.

The spread of tumor sizes in Grade III is particularly wide, showing that while the median size is larger, there is considerable variation among patients.

Outliers are evident in all grades, but especially in Grade II and III, indicating some tumors are exceptionally large regardless of their grade.

**Tumor Size vs. Estrogen Status:**

A graph of a diagram

Description automatically generated with medium confidence

*Figure18: Box plot of Tumor size vs Estrogen status*

The median tumor size in estrogen receptor-positive (ER+) cancers appears to be slightly lower than in estrogen receptor-negative (ER-) cancers, which might suggest that ER+ tumors are detected at a smaller size, potentially due to their growth patterns or differences in screening.

There is substantial overlap in the IQRs between ER+ and ER- tumors, indicating a considerable amount of variability within each group.

The presence of outliers in both categories indicates some patients with ER+ status still have very large tumors, and the same is true for ER- status, which emphasizes the heterogeneity of breast cancer.

Each of these insights can be expanded upon with statistical tests to determine if the differences observed are statistically significant. Additionally, clinical expertise can provide context to these observations, shedding light on the biological and treatment-related factors that might explain the patterns seen in the data.

**Insights for Violin Plots of Numerical Variables vs. 6th Stage:**

**Age vs. 6th Stage:**

A diagram of different colored leaves

Description automatically generated

*Figure19: Violin plots of numerical variables vs 6th stage*

The violin plots suggest that the median age tends to be slightly higher in later stages of cancer, indicating a possible correlation between age and cancer progression.

The width of the violins indicates the distribution of patient ages within each stage, with earlier stages showing a broader age range.

**Regional Node Examined vs. 6th Stage:**

A diagram of different colored shapes

Description automatically generated

*Figure20: Violin plots of Regional node examined vs 6th stage*

There is an apparent trend where patients with later stages of cancer have had more regional nodes examined. This could be due to a more aggressive diagnostic approach in advanced stages or might reflect the spreading pattern of the cancer.

The variability within each stage, as shown by the thickness of the violins, suggests that the number of nodes examined is not consistent across all patients within the same stage.

**Regional Node Positive vs. 6th Stage:**

The number of positive regional nodes tends to increase with the cancer stage, which aligns with the expectation that more advanced cancers are likely to have greater lymph node involvement.

A graph of different colored leaves

Description automatically generated

*Figure21: Violin plots of survival months vs 6th stage*

The tails of the violins for later stages indicate the presence of patients with a high number of positive nodes, highlighting cases with significant lymphatic spread.

A graph showing different colors and shapes

Description automatically generated with medium confidence

*Figure22: Violin plots of regional node positive vs 6th stage*

**Survival Months vs. 6th Stage:**

A diagram of a violin plot

Description automatically generated

*Figure23: Violin plots of regional node examined vs 6th stage*

A diagram of a violin plot

Description automatically generated

*Figure24: Violin plots of survival months vs 6th stage*

The distribution of survival months appears to be somewhat uniform across early stages, but for later stages, there is a noticeable decrease in survival, as seen by the shorter length of the violins.

The presence of long tails in the earlier stages suggests that there are survivors who live significantly longer than the median, pointing to successful treatments or less aggressive cancer types.

These insights provide a nuanced understanding of the dataset's features in relation to the stages of cancer. When conducting such an analysis, it's important to consider not only the statistical significance but also the clinical relevance, which could inform treatment decisions and patient care strategies.

**Insights for the Cumulative Distribution Function (CDF) of Age:**

A graph of a number of age

Description automatically generated

*Figure25: cumulative distribution function of Age*

According to the CDF for "Age," around 50% of the patients are under 60 years old. This can be helpful in determining the age at which half of cancer cases manifest.

The bulk of the patients fell into this age group, as indicated by the curve's relative steepness between 40 and 70. Just after the age of 90, the CDF hits about 1 (100%) indicating that nearly all patients are under the age of 90.

The curve's beginning and ending shapes, which are flatter at younger ages, steeper in the middle ages, and flatter again towards older ages, imply that the disease is rare in very young individuals and levels off in people who are older.

**Insights for the Density Curve of Tumor Size:**

A graph of a normal curve

Description automatically generated with medium confidence

*Figure26: Density curve of Tumor size*

The 'Tumor Size' density curve has a peak at the lower end of the tumor size range, indicating that most tumors are tiny.

The lengthy tail to the right indicates that, while most tumors are tiny, a considerable number of patients have bigger tumors.The curve is unimodal, which means there is only one major peak in tumor size and no other peaks, implying separate subpopulations within the sample.

The curve's form also indicates that the distribution of tumor sizes is not normal, but right skewed, which is consistent with clinical data in which there are many very small tumors and few very large ones.

**Chapter 4**

**Modeling the Data:**

**K-means Clustering:**

Cluster analysis is a data mining and machine learning technique for grouping comparable objects into clusters. The goal of K-means clustering is to split a set of objects into K clusters in such a way that the total of the squared distances between the items and their assigned cluster mean is minimized.

Pre-processed cluster analysis of the dataset has been taken here. Categorical variables were grouped together and converted into numerical form suitable for k-means, through one-hot encoding. Variables were transformed into z-score by assigning their mean value as 0 and making their standard deviation equal to 1. K-means requires this standardization so that larger scale variables do not influence the distances calculated between variables.

The dataset was divided into clusters using the k-means clustering algorithm. The elbow method was used to find the optimum number of clusters by plotting the within-cluster sum of squares (WCSS) versus the number of clusters. Subsequently, four clusters proved suitable for analysis using this approach.

Used k-means clustering with k=4 and assigned the resulting groupings to the data. Each of the clusters is assigned to the row. Profiling is done by looking at the mean values of the attributes in each cluster and determining the main traits that each one possesses.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cluster1 | Cluster2 | Cluster3 | Cluster4 |
| Average Age | 54.4 | 54.7 | 53.4 | 53.5 |
| Average Tumor Size | 42.5mm | 14.1mm | 30.6mm | 45.5mm |
| Average Number of Regional Nodes Examined | 21.5 | 12.2 | 13.2 | 15.2 |
| Average Number of Regional Nodes Positive | 15.9 | 1.6 | 1.8 | 4.9 |
| Average Survival Months | 63.3 | 74.4 | 72.0 | 70.1 |
| Most Common Race | White | White | White | White |
| Most Common Marital Status | Married (including common law) | Married (including common law) | Married (including common law) | Married (including common law) |
| Most Common T Stage | T2 | T1 | T2 | T2 |
| Most Common N Stage | N3 | N1 | N1 | N2 |
| Most Common 6th Stage | IIIC | IIA | IIB | IIIA |
| Most Common Grade | Moderately differentiated; Grade II | Moderately differentiated; Grade II | Moderately differentiated; Grade II | Moderately differentiated; Grade II |
| Most Common a Stage | Regional | Regional | Regional | Regional |
| Estrogen Status | Positive | Positive | Positive | Positive |
| Progesterone Status | Positive | Positive | Positive | Positive |

**Cluster1: Advanced Disease, Poorer Prognosis:**

The first group has a biggest mean tumor size and positive lymph-nodes more often associated with advanced disease. Stage IIIC is the highest and more often than not it corresponds to a more advanced form of cancer. Patients in this group probably are in the advanced stages of breast cancer and are poor prognosis. The patients with the worst prognosis had aggressive treatment indicating the presence of more severe symptoms or later stage diagnosis.

**Cluster2: Early Detection, Better Prognosis:**

This is the group with small tumors and the least number of positive nodes. This is stage IIA, meaning that it shows early detection. The first cluster is probably associated with early stages of cancer diagnosis and therefore provides better prognosis than in the second cluster. The last one and a half years that made up the highest average of survival months were supported by this. They would have possibly been aided by routine monitoring followed by prompt intervention.

**Cluster3: Intermediate Group with Moderate Outcomes:**

The second group consists of patients with moderate size tumors and few positive lymph nodes. These individuals are believed to be at the latter stage of cancer progression but not yet fully established stage (IIB). Here, these groups may show a mixed prognosis. They have a less evolved disease compared to Cluster1 but not as progressed at the early stage of Cluster2. With more months of survival more than those in Cluster1 and less than that of Cluster2, it could be interpreted as moderate outcome.

**Cluster4: Higher Risk with Regional Spread:**

The tumor average size in this cluster is relatively high, with more than thirty percent nodes with metastases. The stage is often referred to as the worst in which the cancer has progressed and affected the neighboring lymph nodes. The spread of cancer to regional lymph nodes makes patients in this group at higher risk. This makes their treatment more complicated as it may involve combined surgery, chemotherapy, and radiotherapy.

**Regression Model Building:**

After examining carefully and eventually comprehending all the columns and research questions, this information can be used in making a prediction model. Machine learning and statistics utilize algorithms/mathematical models to project future outcomes by historical data. Organizations are able to study through various patterns and trends in addition to probable relations with this system, some of them can be seen only briefly.

Predictive models have numerous applications including predicting consumer behavior, fraud detection, improving marketing efforts and trend forecasting. This study utilizes predictive modeling to model tumor sizes of breast cancer patients using various properties. These variables may include race, marital status, T stage, N stage, 6th stage, grade, A stage, estrogen status, and progesterone status. Therefore, we must talk about what we mean by “outcome variable” such as the tumor size which is one of the most important indicators for patient’s condition reflected in the set of data. Regression models are suitable for predicting a continuous response like tumor size which makes this a perfect subject for regression analysis. Review of these different issues will reveal their impact on the size of the tumors, which would help proper diagnoses and treatment as well as prognosis of recovery.

**Pre-Processing:**

However, in this case, we used a dataset with qualitative parameters including Age, Sex, Site, Diagnostic Confirmation, Histology, Race and Alive. Data was preprocessed with label encoding prior to analysis. This method employs numbers as representatives of each one of the categories in the categorical data, thus enabling the use of a particular class of machine learning algorithms – those using exclusively numerical inputs.

We had however, opted to delete it only unlike the first move when we got rid of all of these columns. Such consideration was driven by the significance and the consequences of status on the data and model accuracy. Then, we trimmed the dataset by excluding the ‘Status’ column and divided it for a training and testing part. From the available data, we chose to use 60% to develop our machine learning models with the remaining 40% for testing purposes. The importance of such a division cannot be underestimated when it comes to machine learning practice, since it helps reduce issues associated with an increased risk for over-fitting. On the other hand, when a model over fits the training data, then it means that the result is not predictable for the new data. It is significant to train on one subset and test on another so as to determine if the developed robust machine learning model has been created or not and comparing with others sets of samples to check its performance level.

Studying breast cancer with the help of SEER Research Plus Data, we have ensured appropriate preparation of and splitting up of data according to established machine learning guidelines. We, therefore, train our model on a wider set of data and test to determine whether our model possesses the capability of generalizing and being true about another part which is left out.

**Model Training:**

The dataset included a set of pre-processed inputs and continuous target. variable. The dataset was randomly split into training and testing sets in an 80: 20 ratios. We trained four regression models on the training set and tested them on the testing set. The MSE was applied as one of the performance measurement methods for these models. MAE, and R2.

**Linear Regression:**

However, our first model was simply a linear regression. It holds that there exists a straightforward linear connection of the input characteristics with the resultant. It works by selecting the coefficients of the input variables that are meant to minimize the difference between the projected results and the real observed values. For evaluating this model’s efficiency, MSE, MAE and R2 were used on the test set.

|  |  |  |
| --- | --- | --- |
| Metric | Training Data | Test Data |
| MSE | 103.929922 | 117.971443 |
| MAE | 6.885314 | 6.942174 |
| MAPE | 0.311699 | 0.319845 |

R-squared and Adjusted R-squared: The r-squared value of 0.761 means there is an explanation for about 76.1% variation in the dependent variable model. This is quite large, indicating that the model fits well with the data. This interpretation is supported by Adjusted R-squared which is very close to the R-squared at 0.760. Adjusted R-squared becomes especially crucial since it accounts that fact that the number of predictors in the model may result in some kind of over-fitting. In our example, however, the two values are very close to each other, implying that there is no much risk.

F-statistic and Prob (F-statistic): This represents the overall importance (or significance) of the model and is measured by the F- statistic. This means that the chance relationship between the variables is highly probable because the F-statistic value of 424.6 is significant for p > .05 and the p value is small. Also supporting this point is a relatively low value of Prob (F-statistic), which approaches zero. This indicates the possibility of observing such a high a F-statistic, despite the model being insignificant with respect power explanations. Closeness to zero (high confidence) implies a very low probability of the model not possessing predictive power. Cumulatively, these statistics show that your simple linear regression fits well while being statistically important with much variation accounted for. percentage of the total variation in the dependent variable.

**Ridge regression:**

The Ridge Regression model, serving as our second model, exhibits the following performance characteristics: MSE for the train data, MSE value is 103.980524 and test data, MSE values is 118.786386. This indicates that these estimates have considerable amounts of predictive error, particularly in the test data. The mistake may not be insignificant, but it shows the extent to which the model is acceptable regarding the predicted variables.

MAE These values are the MAE values and each one is the arithmetical average of an absolute value of the wrong prediction even when it is not positive. These values are like what is associated with MSE. That implies that there was a moderately correct prediction of the real situation from such a given model and not that exceptional. The difference in predicted and actual values is represented by the MAPE values of 31.2% and 32.1%. The stated points confirm the fact that the model is reliable, however the margins of those errors are still significant.

**Metric Training test data:**

|  |  |  |
| --- | --- | --- |
| MSE | 103.980524 | 118.786386 |
| MAE | 6.878477 | 6.946656 |
| MAPE | 0.312149 | 0.321306 |

|  |  |  |
| --- | --- | --- |
| **Elastic Net Regression:** | | |
| Metric | Training Data | Test Data |
| MSE | 321.848619 | 367.560315 |
| MAE | 13.080054 | 13.721555 |
| MAPE |  | 0.731947 |

|  |  |  |
| --- | --- | --- |
| **Random Forest Regression:** | | |
| Metric | Training Data | Test Data |
| MSE | 102.660362 | 119.880106 |
| MAE | 6.866862 | 6.927029 |
| MAPE | 0.312569 | 0.316764 |

These errors are however higher for each of the new models as compared with the prior ones with respect to training and testing data. As MSE and MAE go up, it means that in an average sense, the model’s prediction is farther away from the actual values. High discrepancies in percentage deviations from actual values are demonstrated by MAPE scores of > 69% for all datasets. This thus portrays the possibility of a scenario whereby Elastic Net Regression is not the most appropriate model for this dataset or for the said dataset.

**Random Forest Regression:**

The above-mentioned model exhibits better metrics for performance. The MSE and MAE are small in comparison, indicating that the forecasts conform with the real ones to a larger extent. Similarly, the MAPE values are also quite low at about 31% showing that the forecast is made in a smaller percentage compared to the Elastic Net regression model. It is worth noting that the metrics are consistent across both training and test sets, which illustrates good generalizability and stability model performance.

According to the random feature importance graph, T Stage (T2, T3) is one of the crucial factors determining outcome. These comprise the size and metastatic aspect of the tumor that has a bearing on cancer outlook. Some other critical variables that are relevant here include regional node positive as well as Nodes Examined. It is less important due to some demographic features such as race and marital status. This implies that in prediction, it is more of a clinical characteristic as opposed to a demographic one.

A graph with a bar and text

Description automatically generated with medium confidence

*Figure27: Random feature Importance graph*

**Model comparison:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | MSE (Training Data) | MSE (Test Data) | MAE (Training Data) | MAE (Test Data) | MAPE (Training Data) | MAPE (Test Data) |
| Simple Linear Regression | 103.93 | 117.97 | 6.89 | 6.94 | 0.312 | 0.320 |
| Ridge Regression | 103.98 | 118.79 | 6.88 | 6.95 | 0.312 | 0.321 |
| Elastic Net Regression | 321.85 | 367.56 | 13.08 | 13.72 | 0.696 | 0.732 |
| Random Forest Regression | 102.66 | 119.88 | 6.87 | 6.93 | 0.313 | 0.317 |

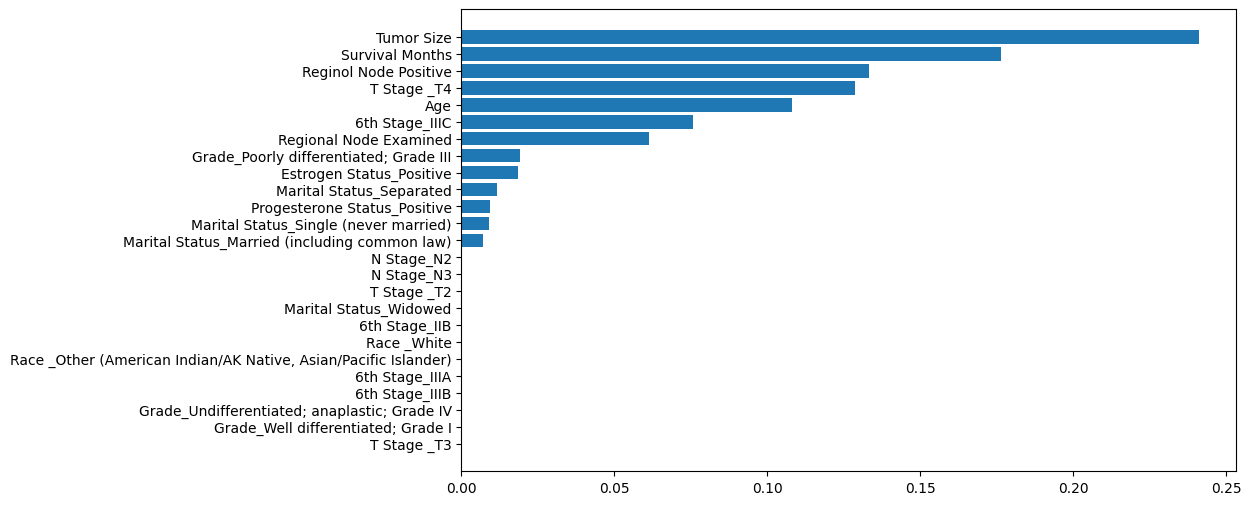
In this course of our capstone project, we performed a comparative study using four different regression-based models for forecasting the value of our target variable. Finally, upon thorough assessment, random forest regression had less MSE than other models while training. However, in testing, the MSE was slightly greater than those for the SLR model and the RR model. Nonetheless, both the third and fourth exhibited meaningful similarity of measurements in all respects.

This meant that Ridge Regulation effect on this data set could be negligible in its effect with regularization. Nevertheless, Elastic Net Regression’s model was characterized with higher errors in terms of both training and test data. On the other hand, the weaker the perfect model fits, it means Elastic Net will not work well with our data peculiarities. However, the MAE and MAPE for all the models were nearly identical. This demonstrates that the mean error is possibly equal in all those models considered above implying that none of them outperforms the others in these aspects. In conclusion, the random forest regression had an opportunity to learn but simple linear regression and ridge regression were comparable and competitive while testing processes gave the added advantage on simplicity and interpretability. The Elastic Net Regression does not live up to expectations. This may indicate that in this instance model complexity is not directly linked with predictive accuracy, so another option may be more appropriate.

**Decision Tree Classifier:**

A decision tree is a non-parametric supervised learning algorithm that can be used to perform classification and regression problems. It is a hierarchical tree with a root node, branches, internal nodes, and leaf nodes. Decision trees are used for classification and regression applications, resulting in simple models.

Accuracy of the Decision tree model is 97%. Below plot shows the significance of variables. Tumor size feature has the highest significance.

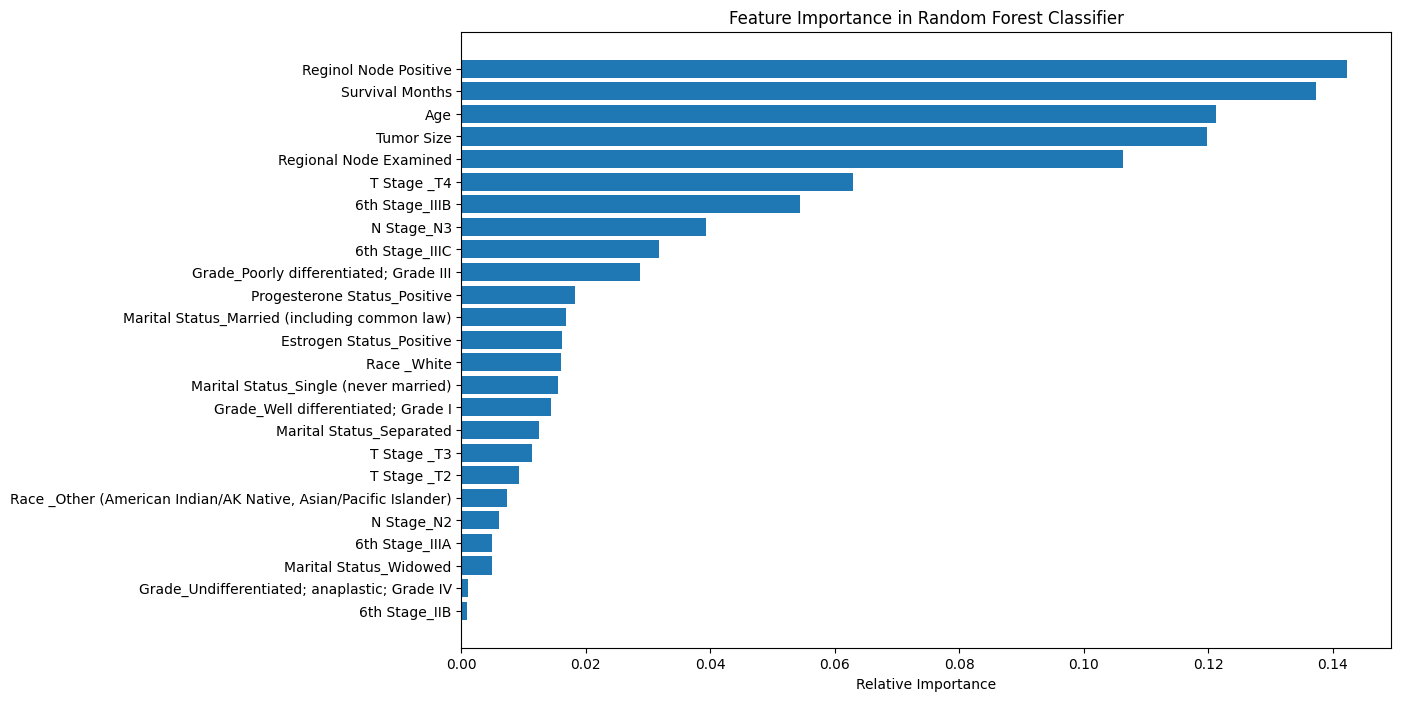


*Figure28: significance of variables in Decision tree model*

**Random Forest Classifier:**

The random forest algorithm is an ensemble learning strategy that combines several classifiers to improve the performance of a model. Random Forest is a decision tree-based supervised machine-learning technique. Random Forest is a classification and regression algorithm.

The accuracy of the Random Forest model is 98%. Below plot shows the significance of variables. Regional node positive variable has the highest significance. By using Random Forest classifier, the importance of each variable for the modeling has been plotted.



*Figure29: Significance of variables in Random forest classifier*

**Conclusion:**

Following an exhaustive examination of the four models, it is evident that the Random Forest Regression model best predicts new, unseen data. Though this model has a slightly larger MSE in test data, it demonstrates the ability to fit complicated patterns without overfitting that is important for generalization on new datasets. For example, Simple Linear Regression and Ridge Regression have a comparable performance with better interpretability but still fail to measure up to a highly nonlinear model like the random forests in the context of capturing complex interactions in our data set. It is therefore recommended that for better confidence in its predictive abilities when deploying the model to predict un-seen data, Random Forest Regression should be used.

In conclusion, the cluster analysis has successfully identified groups of patients with similar characteristics. These groups have the potential to guide clinical decisions and healthcare strategies. Future analyses could expand on this work by incorporating additional patient data, exploring other clustering algorithms, and validating the findings with external datasets.  
  
**Limitations:**

Several variables like ethnicity and marital status have high rates of a single category (e.g. White, Married), making comparisons difficult. More diverse samples are needed. Expanding to multi-center data could help. Tumor grade information relies on pathology interpretations which can be subjective. Standardizing grading methodology across patients could be beneficial.

**Future Work:**

* Incorporate genomic data, gene expression profiles and biomarkers to better predict prognosis and treatment responses.
* Collect longitudinal follow-up data over many years to enable survival analyses and understand long-term outcomes.
* Integrate detailed treatment data such as specific chemo or radiation regimens to analyze treatment efficacy.
* Enhance dataset with imaging data, pathology slides, and digital pathology quantitative image analyses to improve risk stratification models.
* Link patient data with psychosocial questionnaires to study impact on emotional health and quality of life.
* Develop personalized prediction tools combining clinical, molecular, and imaging data to guide individualized treatment plans.

**References:**

A;, Hill DA;Prossnitz ER;Royce M;Nibbe. “Temporal Trends in Breast Cancer Survival by Race and Ethnicity: A Population-Based Cohort Study.” *PloS One*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/31647839/. Accessed 8 Dec. 2023.

Ali, Azka, et al. “Endocrine Adherence in Male versus Female Breast Cancer: A Seer-Medicare Review - Breast Cancer Research and Treatment.” *SpringerLink*, Springer US, 10 Feb. 2022, link.springer.com/article/10.1007/s10549-022-06536-0.

Alsumai TS;Alhazzaa N;Alshamrani A;Assiri S;Alhefdhi A; “Factors Predicting Positive Sentinel Lymph Node Biopsy in Clinically Node-Negative Breast Cancer.” *Breast Cancer (Dove Medical Press)*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/36237483/. Accessed 8 Dec. 2023.

*Breast Cancer Statistics, 2013 - Desantis - 2014 - CA: A Cancer Journal ...*, acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21203. Accessed 8 Dec. 2023.

“Breast Cancer.” *World Health Organization*, World Health Organization, www.who.int/news-room/fact-sheets/detail/breast-cancer. Accessed 8 Dec. 2023.

EJ;, Watkins. “Overview of Breast Cancer.” *JAAPA : Official Journal of the American Academy of Physician Assistants*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/31513033/. Accessed 8 Dec. 2023.

H;, Momenimovahed Z;Salehiniya. “Epidemiological Characteristics of and Risk Factors for Breast Cancer in the World.” *Breast Cancer (Dove Medical Press)*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/31040712/. Accessed 8 Dec. 2023.

LA;, Zagami P;Carey. “Triple Negative Breast Cancer: Pitfalls and Progress.” *NPJ Breast Cancer*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/35987766/. Accessed 8 Dec. 2023.

Lakhtakia, Ritu. “A Brief History of Breast Cancer: Part I: Surgical Domination Reinvented.” *Sultan Qaboos University Medical Journal*, U.S. National Library of Medicine, May 2014, www.ncbi.nlm.nih.gov/pmc/articles/PMC3997531/.

*Seer Breast Cancer Data | IEEE DataPort*, ieee-dataport.org/open-access/seer-breast-cancer-data. Accessed 8 Dec. 2023.

*Segmentation Techniques for Breast Cancer Imaging Modalities-A Review ...*, www.semanticscholar.org/paper/Segmentation-Techniques-for-Breast-Cancer-Imaging-Dabass-Arora/fa13b10ef664763f1a5c0ea0bdb8603d30b17be4. Accessed 8 Dec. 2023.

Soran, Atilla, et al. “Lymph Node Ratio (LNR): Predicting Prognosis after Neoadjuvant Chemotherapy (NAC) in Breast Cancer Patients.” *European Journal of Breast Health*, U.S. National Library of Medicine, 1 Oct. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6776131/.

“Stages of Breast Cancer: Understand Breast Cancer Staging.” *Stages of Breast Cancer | Understand Breast Cancer Staging | American Cancer Society*, www.cancer.org/cancer/types/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html. Accessed 8 Dec. 2023.

U;, Shah SM;Khan RA;Arif S;Sajid. “Artificial Intelligence for Breast Cancer Analysis: Trends & Directions.” *Computers in Biology and Medicine*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/35016100/. Accessed 8 Dec. 2023.

**Codes:**

**Code1(Exploratory Data Analysis):**

import pandas as pd

**# Load the data from the Excel file**

file\_path = 'cap cancer.xlsx' # Replace with your file path

data = pd.read\_excel(file\_path)

**# Initial data inspection (not included in the actual code, but assumed as part of best practices)**

data.info()

data.head()

**# Remove the 'Status' column**

data.drop('Status', axis=1, inplace=True)

**Univariate Analysis for Numerical Variables:**

import matplotlib.pyplot as plt

import seaborn as sns

# Set the aesthetic style of the plots

sns.set\_style("whitegrid")

# Function to create histogram and boxplot for numerical variables

def plot\_numerical(column, data):

fig, ax = plt.subplots(1, 2, figsize=(18, 4)

# Histogram

sns.histplot(data[column], bins=30, ax=ax[0], kde=True)

ax[0].set\_title(f'Histogram of {column}')

# Boxplot

sns.boxplot(x=data[column], ax=ax[1])

ax[1].set\_title(f'Boxplot of {column}'

plt.tight\_layout()

plt.show()

# Iterate over the numerical columns to create plots

numerical\_cols = data.select\_dtypes(include=['int64', 'float64']).columns.tolist()

for col in numerical\_cols:

plot\_numerical(col, data)

**Univariate Analysis for Categorical Variables:**

# Function to create bar plot for categorical variables

def plot\_categorical(column, data):

fig, ax = plt.subplots(figsize=(10, 6))

# Count plot

sns.countplot(y=column, data=data, order = data[column].value\_counts().index)

ax.set\_title(f'Count Plot of {column}')

plt.tight\_layout()

plt.show()

# Iterate over the categorical columns to create plots

categorical\_cols = data.select\_dtypes(include=['object']).columns.tolist()

for col in categorical\_cols:

plot\_categorical(col, data)

### **Correlation Matrix:**

# Compute the correlation matrix for numerical variables

correlation\_matrix = data[numerical\_cols].corr()

# Plot the heatmap for the correlation matrix

plt.figure(figsize=(10, 8))

sns.heatmap(correlation\_matrix, annot=True, fmt=".2f", cmap="coolwarm", cbar\_kws={'shrink': .5})

plt.title('Correlation Matrix of Numerical Variables')

plt.tight\_layout()

plt.show()

### **Pairwise Relationship Plots:**

# Function to create pairplot for selected variables with 'Tumor Size'

def plot\_pairplot(data, target\_variable, features):

# Select columns for pairplot

pairplot\_data = data[features + [target\_variable]]

# Pairplot

sns.pairplot(pairplot\_data, kind='scatter', diag\_kind='kde', corner=True)

g.fig.suptitle(f'Pairwise Relationships with {target\_variable}', y=1.08)

plt.show()

# Selecting a few variables that might influence 'Tumor Size'

selected\_features = ['Age', 'Regional Node Examined', 'Reginol Node Positive', 'Survival Months']

plot\_pairplot(data, 'Tumor Size', selected\_features)

**Boxplots for Categorical Variables vs. Tumor Size:**

# Function to create boxplot for categorical variables against 'Tumor Size'

def plot\_boxplot\_categorical\_vs\_target(column, target, data):

fig, ax = plt.subplots(figsize=(10, 6))

# Boxplot

sns.boxplot(x=column, y=target, data=data)

ax.set\_title(f'Boxplot of {target} vs {column}')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

# Generate and save boxplots for the selected categorical variables against 'Tumor Size'

categorical\_vars\_for\_boxplot = ['T Stage ', 'N Stage', 'Grade', 'Estrogen Status']

for var in categorical\_vars\_for\_boxplot:

plot\_boxplot\_categorical\_vs\_target(var, 'Tumor Size', data)

**Violin Plots for Numerical Variables vs. 6th Stage:**

# Function to create violin plot for categorical variables against '6th Stage'

def plot\_violinplot\_categorical\_vs\_stage(column, stage, data):

fig, ax = plt.subplots(figsize=(12, 8))

# Violin plot

sns.violinplot(x=stage, y=column, data=data)

ax.set\_title(f'Violin Plot of {column} vs {stage}')

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

# Generate and save violin plots for numerical variables against '6th Stage'

for num\_var in numerical\_cols:

if num\_var != 'Tumor Size': # We exclude 'Tumor Size' since it's the target for regression

plot\_violinplot\_categorical\_vs\_stage(num\_var, '6th Stage', data)

**Code2(K-Means Clustering):**

import pandas as pd

# Load the Excel file into a pandas DataFrame to examine its contents

file\_path = 'cap cancer.xlsx'

data = pd.read\_excel(file\_path)

# Display the first few rows of the dataframe to understand its structure

data.head()

# Remove the "Status" column from the dataset

data\_without\_status = data.drop('Status', axis=1)

# Check the data types of the remaining columns

data\_types = data\_without\_status.dtypes

data\_types

from sklearn.preprocessing import OneHotEncoder, StandardScaler

from sklearn.compose import ColumnTransformer

from sklearn.pipeline import Pipeline

# Identify categorical columns (non-numeric columns)

categorical\_cols = data\_without\_status.columns[data\_without\_status.dtypes == 'object']

# Create the preprocessing pipeline for categorical data

# We use OneHotEncoder to convert categorical variables into one-hot numeric array

categorical\_transformer = Pipeline(steps=[

('onehot', OneHotEncoder(handle\_unknown='ignore'))

])

# Create the preprocessing pipeline for numerical data

# StandardScaler will standardize the features by removing the mean and scaling to unit variance

numerical\_cols = data\_without\_status.columns[data\_without\_status.dtypes != 'object']

numerical\_transformer = Pipeline(steps=[

('scaler', StandardScaler())

])

# Combine the numerical and categorical transformations

preprocessor = ColumnTransformer(

transformers=[

('num', numerical\_transformer, numerical\_cols),

('cat', categorical\_transformer, categorical\_cols)

])

# Fit and transform the preprocessor on the data without the 'Status' column

data\_preprocessed = preprocessor.fit\_transform(data\_without\_status)

# The result is a numpy array, so we need to convert it to a DataFrame

# Get feature names for the one-hot encoded columns

ohe\_feature\_names = preprocessor.named\_transformers\_['cat'].named\_steps['onehot'].get\_feature\_names\_out(categorical\_cols)

# Combine with numerical column names

feature\_names = list(numerical\_cols) + list(ohe\_feature\_names)

# Convert to DataFrame

data\_preprocessed\_df = pd.DataFrame(data\_preprocessed, columns=feature\_names)

data\_preprocessed\_df.head()

from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

# Determine the optimal number of clusters by using the elbow method

wcss = []

for i in range(1, 11):

kmeans = KMeans(n\_clusters=i, init='k-means++', max\_iter=300, n\_init=10, random\_state=0)

kmeans.fit(data\_preprocessed\_df)

wcss.append(kmeans.inertia\_)

# Plot the elbow graph

plt.figure(figsize=(10, 5))

plt.plot(range(1, 11), wcss, marker='o')

plt.title('Elbow Method')

plt.xlabel('Number of clusters')

plt.ylabel('WCSS') # within cluster sum of squares

plt.xticks(range(1, 11))

plt.grid(True)

plt.show()

# It seems there was a reset, so I'll need to re-import the necessary libraries and data.

import pandas as pd

from sklearn.preprocessing import OneHotEncoder, StandardScaler

from sklearn.compose import ColumnTransformer

from sklearn.pipeline import Pipeline

from sklearn.cluster import KMeans

# Load the data again

file\_path = 'cap cancer.xlsx'

data = pd.read\_excel(file\_path)

# Remove the "Status" column from the dataset

data\_without\_status = data.drop('Status', axis=1)

# Identify categorical columns (non-numeric columns)

categorical\_cols = data\_without\_status.columns[data\_without\_status.dtypes == 'object']

# Create the preprocessing pipeline for categorical data

categorical\_transformer = Pipeline(steps=[

('onehot', OneHotEncoder(handle\_unknown='ignore'))

])

# Create the preprocessing pipeline for numerical data

numerical\_cols = data\_without\_status.columns[data\_without\_status.dtypes != 'object']

numerical\_transformer = Pipeline(steps=[

('scaler', StandardScaler())

])

# Combine the numerical and categorical transformations

preprocessor = ColumnTransformer(

transformers=[

('num', numerical\_transformer, numerical\_cols),

('cat', categorical\_transformer, categorical\_cols)

])

# Fit and transform the preprocessor on the data without the 'Status' column

data\_preprocessed = preprocessor.fit\_transform(data\_without\_status)

# The result is a numpy array, so we need to convert it to a DataFrame

# Get feature names for the one-hot encoded columns

ohe\_feature\_names = preprocessor.named\_transformers\_['cat'].named\_steps['onehot'].get\_feature\_names\_out(categorical\_cols)

# Combine with numerical column names

feature\_names = list(numerical\_cols) + list(ohe\_feature\_names)

# Convert to DataFrame

data\_preprocessed\_df = pd.DataFrame(data\_preprocessed, columns=feature\_names)

# Perform k-means clustering with k = 4

k = 4

kmeans = KMeans(n\_clusters=k, init='k-means++', max\_iter=300, n\_init=10, random\_state=0)

kmeans.fit(data\_preprocessed\_df)

# Assign the cluster labels to the original data (before preprocessing)

data\_with\_clusters = data\_without\_status.assign(Cluster=kmeans.labels\_)

# Display the first few entries with the cluster assignments

data\_with\_clusters.head()

**Classification:**

**Code3(Decision Tree and Random Forest):**

from sklearn.tree import DecisionTreeClassifier, plot\_tree

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification\_report, confusion\_matrix, accuracy\_score, precision\_score, recall\_score

import numpy as np

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import OneHotEncoder

import pandas as pd

df=pd.read\_csv('/content/SEER Breast Cancer Dataset .csv')

df.head()

df.info()

df.nunique()

df.drop(['Status','Unnamed: 3'],axis=1,inplace=True)

df.head()

df.isna().sum()

categorical\_cols = df.select\_dtypes(include=['object']).columns.drop('A Stage')

categorical\_cols

encoder = OneHotEncoder(sparse=False,drop='first')

encoded\_data = pd.DataFrame(encoder.fit\_transform(df[categorical\_cols]))

encoded\_data.columns = encoder.get\_feature\_names\_out(categorical\_cols)

encoded\_data.index = df.index

final\_data = df.drop(categorical\_cols, axis=1)

final\_data = pd.concat([final\_data, encoded\_data], axis=1)

final\_data

# Splitting the dataset into features (X) and target variable (y)

X = final\_data.drop('A Stage', axis=1)

y = final\_data['A Stage']

# Splitting the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

X.columns

**Decision Tree:**

dt\_classifier = DecisionTreeClassifier(random\_state=42)

# Fit the model on the training data

dt\_classifier.fit(X\_train, y\_train)

# Predictions

y\_pred\_train = dt\_classifier.predict(X\_train)

y\_pred\_test = dt\_classifier.predict(X\_test)

# Classification Report and Confusion Matrix

print("Training Data:",classification\_report(y\_train, y\_pred\_train))

print("Testing Data:",classification\_report(y\_test, y\_pred\_test))

print(confusion\_matrix(y\_test, y\_pred\_test))

# Feature Importance Plot

plt.figure(figsize=(10, 6))

feature\_importances = dt\_classifier.feature\_importances\_

indices = np.argsort(feature\_importances)

plt.barh(range(len(indices)), feature\_importances[indices])

plt.yticks(range(len(indices)), X\_train.columns[indices])

plt.show()

**Random Forest Model:**

# Initialize the Random Forest Classifier

rf\_classifier = RandomForestClassifier(random\_state=42)

# Fit the Random Forest model

rf\_classifier.fit(X\_train, y\_train)

# Predictions on training and testing data

y\_train\_pred\_rf = rf\_classifier.predict(X\_train)

y\_test\_pred\_rf = rf\_classifier.predict(X\_test)

# Classification Report and Confusion Matrix for training data

print("Random Forest - Training Data: Classification Report")

print(classification\_report(y\_train, y\_train\_pred\_rf))

print("Confusion Matrix:")

print(confusion\_matrix(y\_train, y\_train\_pred\_rf))

# Classification Report and Confusion Matrix for testing data

print("\nRandom Forest - Testing Data: Classification Report")

print(classification\_report(y\_test, y\_test\_pred\_rf))

print("Confusion Matrix:")

print(confusion\_matrix(y\_test, y\_test\_pred\_rf))

# Feature Importance Plot

plt.figure(figsize=(12, 8))

plt.title("Feature Importance in Random Forest Classifier")

feature\_importances\_rf = rf\_classifier.feature\_importances\_

indices\_rf = np.argsort(feature\_importances\_rf)

plt.barh(range(len(indices\_rf)), feature\_importances\_rf[indices\_rf], align='center')

plt.yticks(range(len(indices\_rf)), [X\_train.columns[i] for i in indices\_rf])

plt.xlabel('Relative Importance')

plt.show()

**Code 4(Regression):**

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import OneHotEncoder

import statsmodels.api as sm

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn import metrics

import scipy.stats as stats

from sklearn.linear\_model import Ridge, Lasso, ElasticNet

from sklearn.tree import DecisionTreeRegressor, plot\_tree

from sklearn.model\_selection import GridSearchCV

from sklearn.ensemble import RandomForestRegressor

sns.set()

df=pd.read\_csv('/content/SEER Breast Cancer Dataset .csv')

df.head()

df.info()

df.nunique()

df.drop(['Status','Unnamed: 3'],axis=1,inplace=True)

df.head()

df.isna().sum()

categorical\_cols = df.select\_dtypes(include=['object']).columns

categorical\_cols

encoder = OneHotEncoder(sparse=False,drop='first')

encoded\_data = pd.DataFrame(encoder.fit\_transform(df[categorical\_cols]))

encoded\_data.columns = encoder.get\_feature\_names\_out(categorical\_cols)

encoded\_data.index = df.index

final\_data = df.drop(categorical\_cols, axis=1)

final\_data = pd.concat([final\_data, encoded\_data], axis=1)

final\_data

def calculate\_metrics(y\_true, y\_pred):

print("MSE",metrics.mean\_squared\_error(y\_true, y\_pred))

print("MAE",metrics.mean\_absolute\_error(y\_true, y\_pred))

print("MAPE",metrics.mean\_absolute\_percentage\_error(y\_true, y\_pred))

X = final\_data.drop('Tumor Size', axis=1)

y = final\_data['Tumor Size']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Adding a constant to the model (intercept)

X\_train\_sm = sm.add\_constant(X\_train)

# Building the linear regression model using statsmodels

model = sm.OLS(y\_train, X\_train\_sm)

results = model.fit()

# Displaying the summary of the regression model

results.summary()

print("Performance on Training Data")

calculate\_metrics(y\_train,results.predict(X\_train\_sm))

print("-----------\nPerformance on Test Data")

calculate\_metrics(y\_test,results.predict(sm.add\_constant(X\_test)))

**Lets try removing two most significant variable and check the results:**

X\_train\_new\_reduced = X\_train.drop(['Grade\_Undifferentiated; anaplastic; Grade IV', 'A Stage\_Regional'], axis=1)

X\_test\_new\_reduced = X\_test.drop(['Grade\_Undifferentiated; anaplastic; Grade IV', 'A Stage\_Regional'], axis=1)

# Adding a constant to the model (intercept) for the reduced dataset

X\_train\_new\_reduced\_sm = sm.add\_constant(X\_train\_new\_reduced)

# Building the new linear regression model using statsmodels

model\_new\_reduced = sm.OLS(y\_train, X\_train\_new\_reduced\_sm)

results\_new\_reduced = model\_new\_reduced.fit()

# Displaying the summary of the new regression model

results\_new\_reduced.summary()

print("Performance on Training Data")

calculate\_metrics(y\_train,results\_new\_reduced.predict(X\_train\_new\_reduced\_sm))

print("-----------\nPerformance on Test Data")

calculate\_metrics(y\_test,results\_new\_reduced.predict(sm.add\_constant(X\_test\_new\_reduced)))

### **Lets try removing all the non-significant variables and check the results:**

# Identifying significant variables (p < 0.05) for removal

p\_values=results\_new\_reduced.pvalues

significant\_vars = p\_values[p\_values < 0.05].index[1:]

# Dropping significant variables from the training and testing sets

X\_train\_no\_sig = X\_train[significant\_vars]

X\_test\_no\_sig = X\_test[significant\_vars]

# Rebuilding the model without significant variables

X\_train\_no\_sig\_sm = sm.add\_constant(X\_train\_no\_sig)

model\_no\_sig = sm.OLS(y\_train, X\_train\_no\_sig\_sm)

results\_no\_sig = model\_no\_sig.fit()

# Displaying the summary of the model without significant variables

results\_no\_sig.summary()

print("Performance on Training Data")

calculate\_metrics(y\_train,results\_no\_sig.predict(X\_train\_no\_sig\_sm))

print("-----------\nPerformance on Test Data")

calculate\_metrics(y\_test,results\_no\_sig.predict(sm.add\_constant(X\_test\_no\_sig)))

significant\_vars

### **Validating the Linear Regression assumptions:**

### **1) Mean of residuals should be 0**

### residuals = results\_no\_sig.resid

### np.mean(residuals)

### **2) Linearity and Homoscedasticity**

### # Predictions on the training data

### y\_train\_pred\_final = results\_no\_sig.predict(X\_train\_no\_sig\_sm)

### # Calculating residuals

### residuals\_final = y\_train - y\_train\_pred\_final

### # Checking linearity and homoscedasticity

### plt.figure(figsize=(10, 6))

### sns.scatterplot(x=y\_train\_pred\_final, y=residuals\_final)

### plt.title('Residuals vs Predicted Values')

### plt.xlabel('Predicted Values')

### plt.ylabel('Residuals')

### plt.axhline(y=0, color='red', linestyle='--')

### **3) Normality of error terms:**

### stats.probplot(residuals\_final, dist="norm", plot=plt)

### plt.title('Q-Q Plot of Residuals')

### plt.tight\_layout()

### plt.show()

### **Ridge Regression:**

### ridge\_reg = Ridge()

### ridge\_reg.fit(X\_train, y\_train)

### y\_train\_pred\_ridge = ridge\_reg.predict(X\_train)

### y\_test\_pred\_ridge = ridge\_reg.predict(X\_test)

### print("Ridge Regression - Training Data")

### calculate\_metrics(y\_train, y\_train\_pred\_ridge)

### print("\n--------------------------------------\n")

### print("Ridge Regression - Testing Data")

### calculate\_metrics(y\_test, y\_test\_pred\_ridge)

### **Lasso Regression:** lasso\_reg = Lasso()

### lasso\_reg.fit(X\_train, y\_train)

### y\_train\_pred\_lasso = lasso\_reg.predict(X\_train)

### y\_test\_pred\_lasso = lasso\_reg.predict(X\_test)

### print("\nLasso Regression - Training Data")

### calculate\_metrics(y\_train, y\_train\_pred\_lasso)

### print("\n--------------------------------------\n")

### print("\nLasso Regression - Testing Data")

### calculate\_metrics(y\_test, y\_test\_pred\_lasso)

### **Elastic Net Regression:**

### elastic\_net\_reg = ElasticNet()

### elastic\_net\_reg.fit(X\_train, y\_train)

### y\_train\_pred\_elastic\_net = elastic\_net\_reg.predict(X\_train)

### y\_test\_pred\_elastic\_net = elastic\_net\_reg.predict(X\_test)

### print("\nElastic Net Regression - Training Data")

### calculate\_metrics(y\_train, y\_train\_pred\_elastic\_net)

### print("\n--------------------------------------\n")

### print("\nElastic Net Regression - Testing Data")

### calculate\_metrics(y\_test, y\_test\_pred\_elastic\_net)

### **Decision Tree:**

### dtree = DecisionTreeRegressor(random\_state=1)

### dtree.fit(X\_train, y\_train)

### y\_train\_pred\_dtree = dtree.predict(X\_train)

### y\_test\_pred\_dtree = dtree.predict(X\_test)

### # Performance on Training Data

### print("Decision Tree Regression - Training Data")

### calculate\_metrics(y\_train, y\_train\_pred\_dtree)

### print("\n--------------------------------------\n")

### # Performance on Testing Data

### print("Decision Tree Regression - Testing Data")

### calculate\_metrics(y\_test, y\_test\_pred\_dtree)

**Hyperparameter Tuning:**

### # Decision Tree Regressor

### dt\_reg = DecisionTreeRegressor(random\_state=42)

### # Parameters for tuning

### params = {

### 'max\_depth': [3, 5, 7, 10],

### 'min\_samples\_split': [2, 5, 10],

### 'min\_samples\_leaf': [1, 2, 4]

### }

### # GridSearch for parameter tuning

### grid\_search = GridSearchCV(dt\_reg, params, cv=5, scoring='neg\_mean\_squared\_error', n\_jobs=-1)

grid\_search.fit(X\_train, y\_train)

### # Best model after tuning

### dt\_best = grid\_search.best\_estimator\_

### # Predictions

### y\_train\_pred\_dt = dt\_best.predict(X\_train)

### y\_test\_pred\_dt = dt\_best.predict(X\_test)

### # Performance on Training Data

### print("Decision Tree Regression - Training Data")

### calculate\_metrics(y\_train, y\_train\_pred\_dt)

### print("\n--------------------------------------\n")

### # Performance on Testing Data

### print("Decision Tree Regression - Testing Data")

### calculate\_metrics(y\_test, y\_test\_pred\_dt)

### # Feature Importance Plot

### plt.figure(figsize=(8, 4))

### plt.title("Feature Importance in Decision Tree Regressor")

### importances = dt\_best.feature\_importances\_

### indices = np.argsort(importances)

### features = X\_train.columns

### plt.barh(range(len(indices)), importances[indices], align='center')

### plt.yticks(range(len(indices)), [features[i] for i in indices])

### plt.xlabel('Relative Importance')

### plt.show()

### **Random Forest:**

### rf = RandomForestRegressor(random\_state=1, oob\_score=True)

### rf.fit(X\_train, y\_train)

### y\_train\_pred\_rf = rf.predict(X\_train)

### y\_test\_pred\_rf = rf.predict(X\_test)

### # Performance on Training Data

### print("Random Forest Regression - Training Data")

### calculate\_metrics(y\_train, y\_train\_pred\_rf)

### print("\n--------------------------------------\n")

### # Performance on Testing Data

### print("RandOM Forest Regression - Testing Data")

### calculate\_metrics(y\_test, y\_test\_pred\_rf)

### ​

### # Initialize Random Forest Regressor

### rf\_reg = RandomForestRegressor(random\_state=42)

### # Parameters for GridSearchCV tuning

### rf\_params = {

### 'n\_estimators': [50,100, 200,300],

### 'max\_depth': [3, 5, 7,10,15],

### 'min\_samples\_split': [2, 5],

### 'min\_samples\_leaf': [1, 2]

### }

### # GridSearchCV for hyperparameter tuning

### rf\_grid\_search = GridSearchCV(rf\_reg, rf\_params, cv=5, scoring='neg\_mean\_squared\_error', n\_jobs=-1)

### rf\_grid\_search.fit(X\_train, y\_train) # Assuming X\_train\_final and y\_train\_new are defined

### # Best model after tuning

### rf\_best = rf\_grid\_search.best\_estimator\_

### # Predictions on training and testing data

### y\_train\_pred\_rf = rf\_best.predict(X\_train)

### y\_test\_pred\_rf = rf\_best.predict(X\_test)

### # Performance on Training Data

### print("Random Forest Regression - Training Data")

### calculate\_metrics(y\_train, y\_train\_pred\_rf)

### print("\n--------------------------------------\n")

### # Performance on Testing Data

### print("Random Forest Regression - Testing Data")

### calculate\_metrics(y\_test, y\_test\_pred\_rf)

### # Feature Importance Plot

### plt.figure(figsize=(10, 6))

### plt.title("Feature Importance in Random Forest Regressor")

### rf\_importances = rf\_best.feature\_importances\_

### rf\_indices = np.argsort(rf\_importances)

### rf\_features = X\_train.columns

### plt.barh(range(len(rf\_indices)), rf\_importances[rf\_indices], align='center')

### plt.yticks(range(len(rf\_indices)), [rf\_features[i] for i in rf\_indices])

### plt.xlabel('Relative Importance')

### plt.show()