

Sri Lanka Institute of Information Technology

**Classification of Breast Cancer survival**

**Using**

**Supervised Learning**

Machine Learning

(IT-4060)

Submitted by

|  |  |
| --- | --- |
| **Name** | **IT Number** |
| Nirojan.Y | IT20221928 |
| Lavaniyah.L | IT20202422 |
| Perera G.P.A.H.R | IT20175702 |

Date of Submission: 28.04.2023

**Abstract**

Breast cancer is a serious and prevalent disease that affects millions of women worldwide. It is a type of cancer that starts in the breast tissue and can spread to other parts of the body. Early detection and accurate diagnosis of breast cancer are crucial for successful treatment and management of the disease. Now a days, machine learning plays a major role in order to find the solution for early detection of breast cancer. There are several supervised learning algorithms that can be used for breast cancer survival classification, including logistic regression, decision trees, random forests, and support vector machines. Each algorithm has its strengths and weaknesses and may be more or less appropriate depending on the specific dataset and problem at hand. The process of developing a breast cancer survival classification model using supervised learning typically involves several steps, including data preprocessing, feature selection or extraction, model selection, and evaluation. The goal is to develop a model that can accurately classify new, unlabeled data points as whether they can survive or not, with high precision and recall. Furthermore, Breast cancer survival classification using supervised learning has the potential to significantly improve early detection, leading to better patient outcomes and reduced healthcare costs. However, it is important to note that these models should be used in conjunction with clinical expertise and other diagnostic tools, rather than as a replacement for them.

**Keywords:** Artificial Intelligence, Breast Cancer, Computing Methodologies, Genetic Algorithm, Machine Learning, Support Vector Machine, Decision Tree, Random Forest, Logistics regression

**Table of Contents**

1. Introduction
   1. Problem Statement……...……………………………………...................... 04
2. Methodology
   1. Data Collection…………………………………………………………...... 05
      1. Dataset………………….………………………………………….. 06
      2. Description of Dataset….………………………………………….. 06
   2. Data Mapping and Preprocessing……………………………...................... 08
   3. Supervised Learning Algorithms……………………..……………………. 10
      1. Introduction and Selection of the algorithms
      2. Implementation of Selected algorithms
3. Evaluation……………………….…………………………………………………….. 16
   1. Assessment of the project result…………………………………………… 16
   2. Future work………………………………………………………………....19
4. References……………………………………………………………………………... 20
5. Appendix……………………………………………………………………………….23
   1. Appendix A: Code Listening………………………………………………..23

**1. Introduction**

* 1. **Problem Statement**

Breast cancer is the most prevalent cancer in women globally [1] and accounts for 30% of all female cancers [3]; as a result, 1.5 million women worldwide are diagnosed with the illness each year, and 500,000 of them pass away as a result. While the fatality rate has reduced over the previous 30 years, this condition has become more prevalent. However, it is believed that mammography screening reduces mortality by 20% and improves cancer therapy by 60% [5].

Patients with modest and undetectable malignancy indicators can have diagnostic mammography performed to evaluate aberrant breast cancer tissue. This approach can't be utilized to evaluate places where cancer may be suspected because of the sheer volume of photos. A study found that screenings of women with particularly thick breast tissue missed about 50% of breast tumors [7]. However, within two years of screening, roughly 25% of breast cancer patients receive a negative diagnosis. Therefore, it is essential to diagnose breast cancer as soon as possible [8].

The majority of mammography-based breast cancer screenings are done regularly for all women, often once a year or every two years. The "A fix screening program for everyone" may reduce the efficacy of screening programs and is ineffective in diagnosing cancer on an individual basis [9]. On the other hand, experts contend that taking into account additional risk factors in addition to mammography screening can aid in making a more precise diagnosis of women at risk [11]. Additionally, accurate risk prediction using modeling can assist radiologists in identifying high-risk patients as well as setting up personal screenings for patients and enticing them to join in programs for early detection [12, 13].

Machine learning is a modeling method that has been popular in recent years for its ability to forecast various diseases by extracting information from data and uncovering hidden associations. Some research [14] simply employed demographic risk variables (lifestyle and laboratory data) to predict breast cancer survival, whereas other studies [15] relied on mammographic images or patient biopsy data to make their predictions. Others demonstrated how genetic information may be used to predict breast cancer.

A major challenge in predicting breast cancer survival is the creation of a model for addressing all known risk factors [16]. Without considering other crucial elements, current prediction algorithms could merely analyze mammographic pictures or demographic risk factors. These models, which may be used to identify high-risk women, may also lead to frequent screenings and invasive ultrasound and magnetic resonance imaging (MRI) sampling. Patients could feel the financial and psychological strain.

The problem statement is to develop a supervised learning model that can accurately classify breast cancer survival based on their features. The model will be trained on a dataset of patients' breast tumor characteristics such as tumor size, as well as patient age and family history. According to the model he system will predict whether the patient will survive or not.

1. **Methodology**

The goal of the model is to accurately predict whether the patient will survive or not, based on the features. This can help healthcare professionals to diagnose breast cancer earlier and provide more effective treatment options for patients.

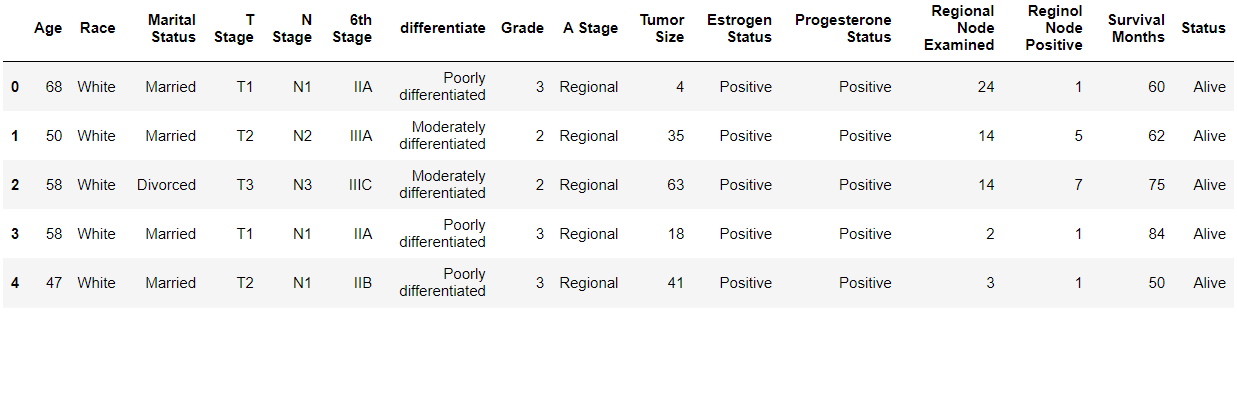
Supervised learning is a popular machine learning technique that can be used to develop predictive models for breast cancer survival classification. This technique involves training a model on a labeled dataset, where each data point is assigned a class label, such as "Alive" or "Dead." The model can then be used to predict the class labels of new, unlabeled data points.

The classification task can be framed as a binary classification problem, where the output variable is a binary value indicating whether the patient can survive or not. A variety of supervised learning algorithms can be used to build the model, such as logistic regression, decision trees, random forests, support vector machines, and neural networks.

The performance of the model will be evaluated using metrics such as accuracy, precision, recall, and F1-score, and the model can be fine-tuned by adjusting hyperparameters and feature selection techniques to optimize performance.

* 1. **Data Collection**

To classify breast cancer survival using supervised learning, system will need a dataset that includes information about patients' breast cancer diagnosis and other relevant variables such as age, Progesterone Status, and Marital Status. The dataset can be obtained from various sources such as clinical trials, medical institutions, and research studies.

Once got the dataset, supervised learning process begins. This involves splitting the dataset into two parts: a training set and a testing set. The training set is used to train the machine learning model, while the testing set is used to evaluate its performance.

*First five rows of collected dataset.*

Here are the steps, which take to classify breast cancer using supervised learning:

**Data preparation:** Clean and preprocess the data, removing any missing or irrelevant data, and normalize the data.

**Feature selection**: Identify the most relevant features that contribute to breast cancer survival prediction. The techniques such as correlation analysis and principal component analysis (PCA) to identify the most important features.

**Choose a supervised learning algorithm:** There are many supervised learning algorithms to choose from, such as logistic regression, support vector machines (SVM), and random forests. And choose an algorithm that best fits the data and problem.

**Train the model:** Use the training set to train the machine learning model. This involves feeding the algorithm with the labeled data and adjusting the model's parameters until it accurately classifies the training data.

**Evaluate the model:** Use the testing set to evaluate the performance of the model. Metrics which can be use such as accuracy, precision, recall, and F1-score to evaluate the performance of the model.

Overall, classification of breast cancer survival using supervised learning is a complex process that requires careful data preparation, feature selection, algorithm selection, training, and evaluation. With the use of these steps, the accurate and reliable breast cancer survival classification model will be created.

* + 1. **Dataset**

[**https://www.kaggle.com/datasets/reihanenamdari/breast-cancer?resource=download**](https://www.kaggle.com/datasets/reihanenamdari/breast-cancer?resource=download)

* + 1. **Description of Dataset**

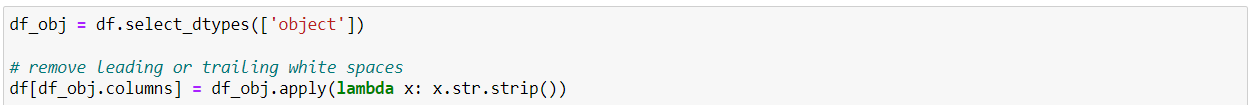
1. Patient Age: The age of the patient at the time of diagnosis.
2. Race: The race of the patient, which can be White, Black, Hispanic, Asian/Pacific Islander, or American Indian/Alaska Native.
3. Marital Status: The marital status of the patient, which can be Married, Single, Divorced, Widowed, or Unknown.\
4. T Stage: The T stage of the tumor, which indicates the size of the primary tumor and whether it has invaded nearby tissue. It is classified into 7 categories ranging from T0 (no evidence of tumor) to T4 (tumor of any size with direct extension to chest wall or skin).
5. N Stage: The N stage of the tumor, which indicates the presence and extent of lymph node involvement. It is classified into 4 categories ranging from N0 (no regional lymph node metastasis) to N3 (metastasis in ipsilateral supraclavicular lymph nodes).
6. 6th Stage: The stage of the cancer is based on the 6th edition of the AJCC staging system. It is classified into 4 categories ranging from Stage 0 (carcinoma in situ) to Stage III (locally advanced cancer).
7. Differentiate: The histological grade of the tumor, which indicates how abnormal the cells appear under a microscope. It is classified into 3 categories: well, differentiated (low grade), moderately differentiated (intermediate grade), and poorly differentiated (high grade).
8. A Stage: The stage of the cancer based on the TNM staging system. It combines information about the size and extent of the primary tumor (T), the presence and extent of lymph node involvement (N), and the presence of distant metastases (M). It is classified into 4 categories ranging from Stage 0 (carcinoma in situ) to Stage IV (metastatic cancer).
9. Estrogen Status: The status of the tumor with respect to estrogen receptor expression. It is classified as Positive or Negative.
10. Progesterone Status: The status of the tumor with respect to progesterone receptor expression. It is classified as Positive or Negative.
11. Grade: The grade of the tumor, which indicates how abnormal the cells appear under a microscope. It is classified into 3 categories: Grade I (well differentiated), Grade II (moderately differentiated), and Grade III (poorly differentiated).
12. Survival Months: The number of months from the date of diagnosis to death or last follow-up.
13. Status: The status of the patient at the end of the study, which can be Alive or Dead.

|  |  |
| --- | --- |
| Categorical data | Numerical data |
| * + - * Race       * Marital status       * T stage       * N stage       * 6th stage       * Differentiate       * Grade       * A stage       * Estrogen Status       * Progesterone Status       * Status | * Age * Tumor Size * Size * Regional Node Examined * Survival Months |

* 1. **Data Mapping and Preprocessing**

**Data collection:** Collecting a dataset of breast cancer cases. This dataset should be representative of the population being studied.

**Data preprocessing:** Preprocessing the dataset to remove any missing or erroneous data, and to transform the data into a format suitable for machine learning algorithms. This may involve standardization, normalization, or feature scaling.



Graphical user interface, text, application

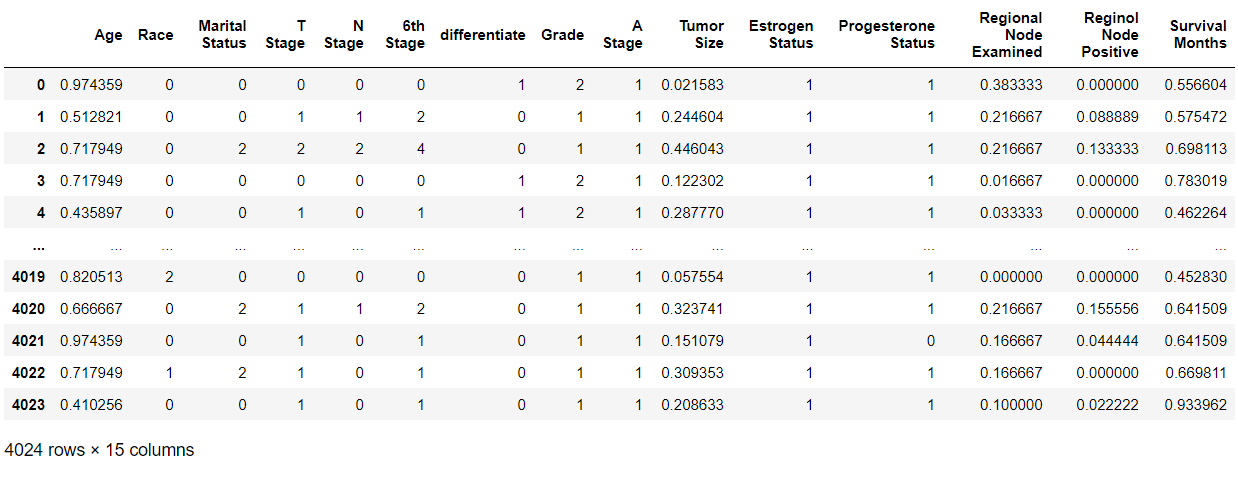
Description automatically generated

Text

Description automatically generated

Graphical user interface, text

Description automatically generated



*collected dataset after data preprocessing.*

**Feature selection:** Selecting the most important features from the dataset to be used in the classification model. This can be done using statistical methods or machine learning algorithms.

**Model training:** Using a supervised learning algorithm, such as logistic regression, decision tree, random forest, or support vector machine, to train a model on the preprocessed dataset.

**Model evaluation:** Evaluating the performance of the model using metrics such as accuracy, precision, recall, and F1 score, on a validation set or through cross-validation.

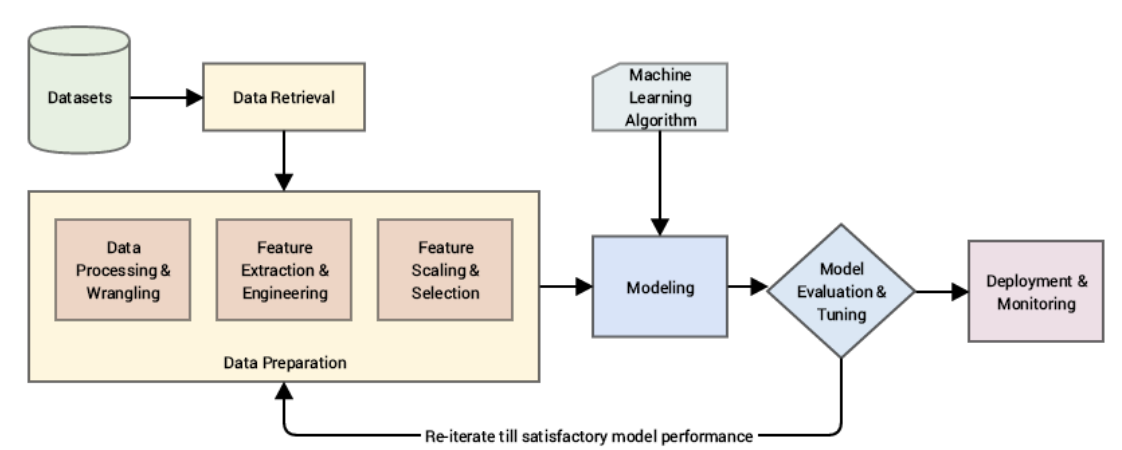
**Model testing:** Testing the final model on a separate testing set to evaluate its generalization performance.

To preprocess the data, it is important to first clean the data by removing any missing or erroneous data points. Then, the data can be transformed into a format suitable for machine learning algorithms. This may involve standardization, normalization, or feature scaling to make the features more comparable to each other.

Feature selection is also an important step in breast cancer classification. Selecting the most important features helps to reduce the dimensionality of the dataset and improve the accuracy of the model. Feature selection can be done using statistical methods such as correlation analysis or using machine learning algorithms such as recursive feature elimination.

Once the data is preprocessed and the features are selected, the supervised learning algorithm can be used to train a model on the preprocessed dataset. The performance of the model can be evaluated using metrics such as accuracy, precision, recall, and F1 score. These metrics help to determine the effectiveness of the model in classifying breast cancer cases.

Finally, the trained model can be tested on a separate testing set to evaluate its generalization performance. If the model performs well on the testing set, it can be used to classify surviving level of breast cancer cases in clinical settings.



*Figure 2.2.1 Overall Diagram*

* 1. **Supervised Learning Algorithms**

Supervised learning algorithms are a type of machine learning algorithm in which the computer is trained using labeled data. In supervised learning, the algorithm is provided with a set of inputs and corresponding outputs, known as the training set. The algorithm uses this training set to learn the mapping between inputs and outputs, and then applies this mapping to new, unseen inputs in order to make predictions or classifications.

Following are some supervised learning algorithms related for our topic:

**Logistic Regression**: a classification algorithm that uses logistic function to model the probability of a binary output.

**Decision Trees:** a decision-making algorithm that creates a tree-like model of decisions and their possible consequences.

**Random Forest:** an ensemble learning algorithm that uses multiple decision trees to improve the accuracy of predictions.

**Support Vector Machines (SVMs):** a binary classification algorithm that finds the hyperplane that maximizes the margin between two classes.

* + 1. **Introduction and Selection of the algorithms**
* **Logistic Regression**

Logistic regression is a commonly used machine learning algorithm for binary classification problems, such as breast cancer survival detection. The goal of logistic regression is to predict a binary outcome (e.g., in our case the detected patient will survive or not) based on one or more predictor variables (e.g., age, marital status, etc.).

To apply logistic regression for breast cancer survival range detection, first need to collect a dataset that includes information on patients with breast cancer in different stages, along with relevant predictor variables. This dataset could be used to train a logistic regression model, which would then be used to predict the probability of breast cancer detected patients based on their predictor variables.

One example of such a dataset is the Breast Cancer Wisconsin (Diagnostic) Dataset, which contains information on various characteristics of breast cancer tumors, along with a binary label indicating whether the tumor is malignant or benign.

To train a logistic regression model on this dataset, first split the data into training and test sets, and then use the training data to fit a logistic regression model. The model would then be evaluated on the test data to determine its performance.

In practice, it is important to carefully choose and preprocess the predictor variables used in the model, as well as to evaluate the model's performance using appropriate metrics such as accuracy, precision, recall, and f1-score. It is also important to consider factors such as class imbalance and feature selection in order to build an accurate and robust model.

* **SVM**

Support Vector Machines (SVM) is a popular machine learning algorithm that can be used for binary classification problems such as breast cancer survival detection. SVM works by finding a hyperplane that separates the two classes (i.e., whether the tumor is malignant or benign cells to predict the detected patient will survive or not) with maximum margin. The algorithm tries to maximize the distance between the hyperplane and the closest points from each class, which are called support vectors.

To use SVM for breast cancer survival detection, we first need to prepare the data by extracting features from the medical images (such as mammograms) that are indicative of cancerous cells. Commonly used features include Age, marital status, and Race. We then train the SVM model on a labeled dataset.

Once the SVM model is trained, we can use it to predict whether a new patient will survive or not. To do this, we extract the same features from the new patient and feed them into the SVM model, which outputs a binary classification result.

* **Decision Tree**

A decision tree for breast cancer survival detection can be created using a dataset of patient information and outcomes. Here are the steps to create a decision tree:

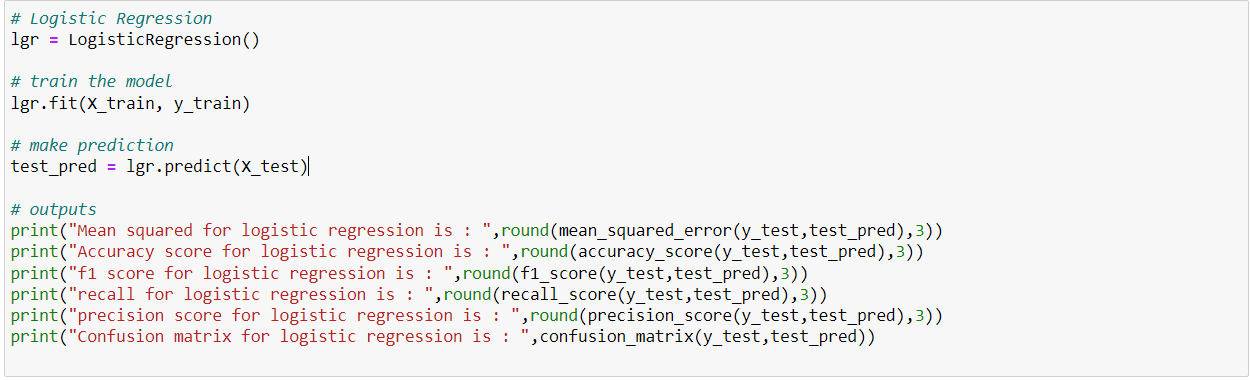
* **Random Forest**

Random Forest is a commonly used machine learning algorithm that can be applied to various classification and regression problems, including predicting the survival of patients with breast cancer.

To use Random Forest for this purpose, system will need a dataset that includes information about patients with breast cancer, such as age, tumor size, tumor stage, hormone receptor status, and other relevant clinical factors. The dataset should also include information about whether the patient survived or not.

Once got the dataset, by using Random Forest to create a predictive model that can be used to classify new patients as either survivors or non-survivors based on their clinical factors.

* + 1. **Implementation of Selection Algorithms**
* **Logistic Regression**

****

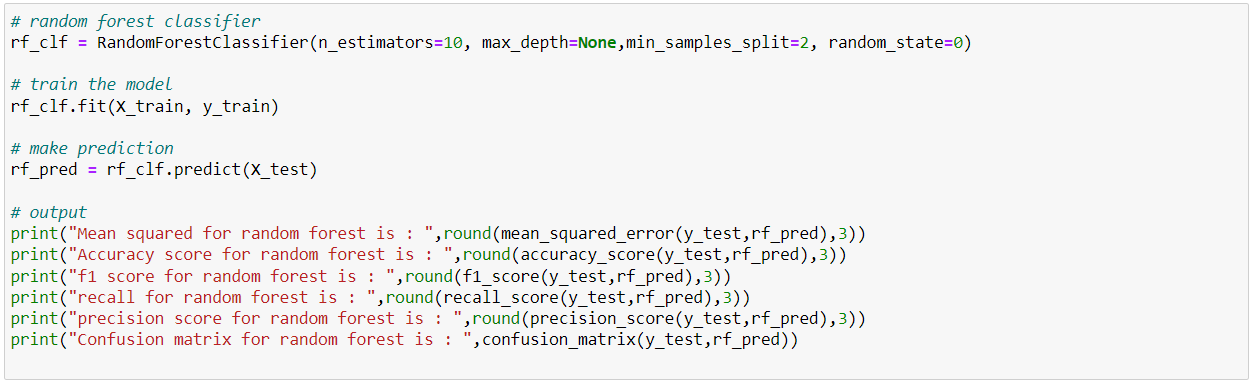
* **A picture containing text

  Description automatically generatedSVM**
* **Decision Tree**

**Text

Description automatically generated with medium confidence**

* **Random Forest**

****

1. **Evaluation**

To evaluate a model for detecting the survival of patients with breast cancer, there are several commonly used evaluation metrics that can be used. These include:

***Accuracy:*** This is a measure of how often the model correctly predicts the survival status of patients. It is calculated as the number of correct predictions divided by the total number of predictions.

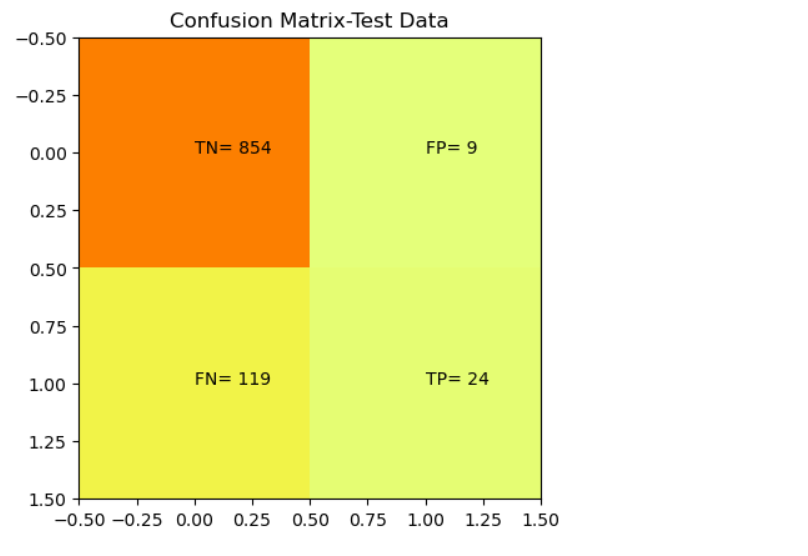
***Precision:*** This measures the proportion of true positive predictions among all the positive predictions. In the context of breast cancer survival, this would mean the proportion of patients who are correctly predicted to survive among all patients predicted to survive.

***Recall:*** This measures the proportion of true positive predictions among all actual positive cases. In the context of breast cancer survival, this would mean the proportion of patients who are correctly predicted to survive among all patients who actually survived.

***F1-score:*** This is the harmonic mean of precision and recall and provides a combined measure of both metrics. It is useful when precision and recall have different levels of importance for the problem being solved.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ***Accuracy*** | ***Precision*** | ***Recall*** | ***F1-score*** |
| ***logistic regression*** | 0.873 | 0.727 | 0.168 | 0.273 |
| ***decision tree*** | 0.779 | 0.263 | 0.308 | 0.284 |
| ***random forest*** | 0.851 | 0.443 | 0.189 | 0.265 |
| ***support vector machine*** | 0.866 | 0.667 | 0.112 | 0.192 |

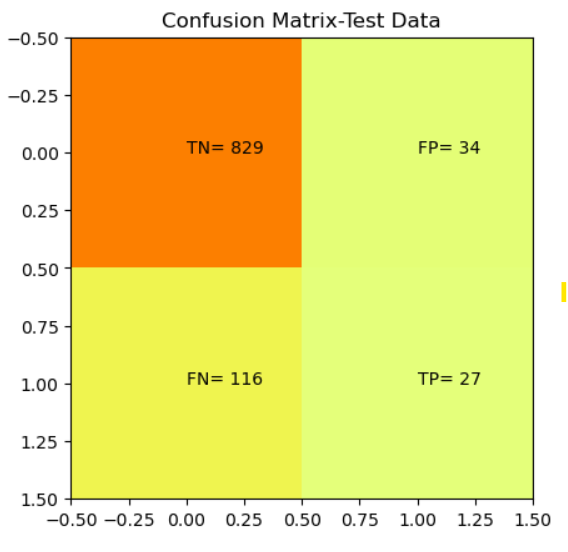
*Comparison of the algorithms*

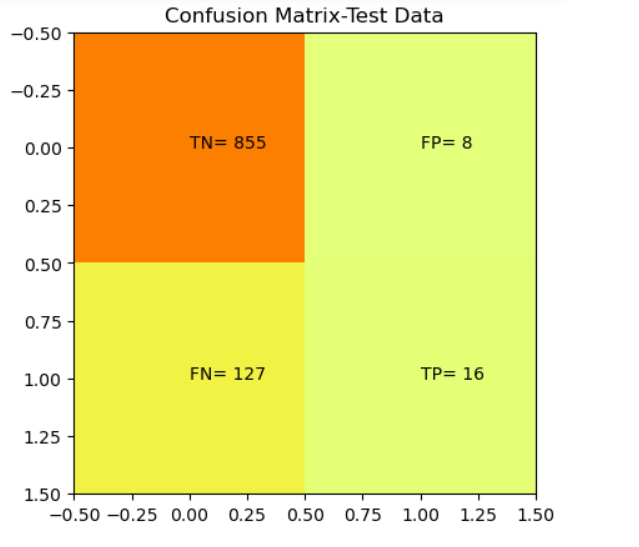


Chart, treemap chart

Description automatically generated*Confusion matrix for test data – logistic regression*

*Confusion matrix for test data – decision tree*



*Confusion matrix for test data – random forest*

*Confusion matrix for test data – support vector machine*

1. **Future work**

Worldwide, the number of instances of breast cancer is increasing at an alarming rate. Worldwide, death rates are increasing and are much higher in industrialized nations. BC must be diagnosed and treated early to increase long-term survival. DL models are used to diagnose breast cancer, which considerably aids medical professionals in reaching a choice.

Notably, the information utilized was a November 2017 update from the NCI's SEER Program. If one is available, an updated dataset would be helpful in confirming the efficacy of classification systems for breast cancer detection. A fresh dataset could receive new features. Further research is required on the impact of feature selection on the predictive assessments of the approaches. To explore prediction approaches and improve accuracy, deep learning with huge data may employ additional machine learning algorithms.

The analysis of the data shows that the combination of multidimensional data with various feature selection, classification, and dimensionality reduction approaches might offer advantageous tools for inference in this field. It is necessary to conduct additional studies in this area to improve the effectiveness of classification algorithms and enable them to make predictions on a wider range of factors. In order to attain high accuracy, we want to parametrize our categorization systems. We are investigating several datasets and the potential application of Machine Learning methods to characterize Breast Cancer. We aim to maximize accuracy while lowering mistake rates.

1. **References**
2. Zhang X, Shengli SU, Hongchao WA. Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. Big Data Research*.*2019;5(1):2019005. doi: 10.11959/j.issn.2096-0271.2019005. [[CrossRef](https://doi.org/10.11959%2Fj.issn.2096-0271.2019005" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+Big+Data+Research&title=+Intelligent+diagnosis+model+and+method+of+palpation+imaging+breast+cancer+based+on+data+mining&author=X+Zhang&author=SU+Shengli&author=WA+Hongchao&volume=5&issue=1&publication_year=+2019&pages=2019005&doi=10.11959/j.issn.2096-0271.2019005&)]
3. Chen SI, Tseng HT, Hsieh CC. Evaluating the impact of soy compounds on breast cancer using the data mining approach. Food & function*.*2020;11(5):4561–70. doi: 10.1039/C9FO00976K. [[PubMed](https://pubmed.ncbi.nlm.nih.gov/32400770)] [[CrossRef](https://doi.org/10.1039%2FC9FO00976K" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+Food+&+function&title=+Evaluating+the+impact+of+soy+compounds+on+breast+cancer+using+the+data+mining+approach&author=SI+Chen&author=HT+Tseng&author=CC+Hsieh&volume=11&issue=5&publication_year=+2020&pages=4561-70&pmid=32400770&doi=10.1039/C9FO00976K&)]
4. Aavula R, Bhramaramba R, Ramula US. A Comprehensive Study on Data Mining Techniques used in Bioinformatics for Breast Cancer Prognosis. Journal of Innovation in Computer Science and Engineering*.*2019;9(1):34–9. [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+Journal+of+Innovation+in+Computer+Science+and+Engineering&title=+A+Comprehensive+Study+on+Data+Mining+Techniques+used+in+Bioinformatics+for+Breast+Cancer+Prognosis&author=R+Aavula&author=R+Bhramaramba&author=US+Ramula&volume=9&issue=1&publication_year=+2019&pages=34-9&)]
5. Kaushik D, Kaur K. Application of Data Mining for high accuracy prediction of breast tissue biopsy results. 2016 Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC); Moscow, Russia: IEEE; 2016. p. 40-5. doi: 10.1109/DIPDMWC.2016.7529361. [[CrossRef](https://doi.org/10.1109%2FDIPDMWC.2016.7529361" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar?q=Kaushik+D+Kaur+K+Application+of+Data+Mining+for+high+accuracy+prediction+of+breast+tissue+biopsy+results.+2016+Third+International+Conference+on+Digital+Information+Processing,+Data+Mining,+and+Wireless+Communications+(DIPDMWC);+Moscow,+Russia:+IEEE;+2016.+p.+40-5+10.1109/DIPDMWC.2016.7529361+)]
6. Mokhtar SA, Elsayad A. Predicting the severity of breast masses with data mining methods. ArXiv preprint arXiv:1305*.*7057 2013 doi: 10.48550/ARxIV.1305.7057. [[CrossRef](https://doi.org/10.48550%2FARxIV.1305.7057" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+ArXiv+preprint+arXiv:1305&title=+Predicting+the+severity+of+breast+masses+with+data+mining+methods&author=SA+Mokhtar&author=A+Elsayad&publication_year=7057+2013&doi=10.48550/ARxIV.1305.7057&)]
7. Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology*.*2018;12(2):119–26.
8. Fan J, Wu Y, Yuan M, Page D, Liu J, Ong IM, Peissig P, Burnside E. Structure-leveraged methods in breast cancer risk prediction. The Journal of Machine Learning Research*.*2016;17(1):2956–70. [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5446896/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/28559747)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+The+Journal+of+Machine+Learning+Research&title=+Structure-leveraged+methods+in+breast+cancer+risk+prediction&author=J+Fan&author=Y+Wu&author=M+Yuan&author=D+Page&author=J+Liu&volume=17&issue=1&publication_year=+2016&pages=2956-70&)]
9. Burnside ES, Liu J, Wu Y, Onitilo AA, McCarty CA, Page CD, et al. Comparing Mammography Abnormality Features to Genetic Variants in the Prediction of Breast Cancer in Women Recommended for Breast Biopsy. Acad Radiol*.*2016;23(1):62–9. doi: 10.1016/j.acra.2015.09.007. [ [PMC Free Article](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4684977) ] [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4684977/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/26514439)] [[CrossRef](https://doi.org/10.1016%2Fj.acra.2015.09.007" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+Acad+Radiol&title=+Comparing+Mammography+Abnormality+Features+to+Genetic+Variants+in+the+Prediction+of+Breast+Cancer+in+Women+Recommended+for+Breast+Biopsy&author=ES+Burnside&author=J+Liu&author=Y+Wu&author=AA+Onitilo&author=CA+McCarty&volume=23&issue=1&publication_year=+2016&pages=62-9&pmid=26514439&doi=10.1016/j.acra.2015.09.007&)]
10. Stephens K. New Mammogram Measures of Breast Cancer Risk Could Revolutionize Screening. AXIS Imaging News*.*2020 [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+AXIS+Imaging+News&title=+New+Mammogram+Measures+of+Breast+Cancer+Risk+Could+Revolutionize+Screening&author=K+Stephens&publication_year=+2020&)]
11. Feld SI, Fan J, Yuan M, Wu Y, Woo KM, Alexandridis R, Burnside ES. Utility of Genetic Testing in Addition to Mammography for Determining Risk of Breast Cancer Depends on Patient Age. AMIA Jt Summits Transl Sci Proc*.*2018;2017:81–90. [ [PMC Free Article](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961791) ] [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961791/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/29888046)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+AMIA+Jt+Summits+Transl+Sci+Proc&title=+Utility+of+Genetic+Testing+in+Addition+to+Mammography+for+Determining+Risk+of+Breast+Cancer+Depends+on+Patient+Age&author=SI+Feld&author=J+Fan&author=M+Yuan&author=Y+Wu&author=KM+Woo&volume=2017&publication_year=+2018&pages=81-90&pmid=29888046&)]
12. Guan Y, Nehl E, Pencea I, Condit CM, Escoffery C, Bellcross CA, McBride CM. Willingness to decrease mammogram frequency among women at low risk for hereditary breast cancer. Sci Rep*.*2019;9(1):9599. doi: 10.1038/s41598-019-45967-6. [ [PMC Free Article](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6610104) ] [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6610104/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/31270367)] [[CrossRef](https://doi.org/10.1038%2Fs41598-019-45967-6" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+Sci+Rep&title=+Willingness+to+decrease+mammogram+frequency+among+women+at+low+risk+for+hereditary+breast+cancer&author=Y+Guan&author=E+Nehl&author=I+Pencea&author=CM+Condit&author=C+Escoffery&volume=9&issue=1&publication_year=+2019&pages=9599&pmid=31270367&doi=10.1038/s41598-019-45967-6&)]
13. American Cancer Society. Cancer facts & figures 2018*.* Atlanta: American Cancer Society; 2018. [[Google Scholar](https://scholar.google.com/scholar_lookup?title=+Cancer+facts+&+figures+2018&publication_year=2018&)]
14. Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T, Dive C. Progress and prospects of early detection in lung cancer. Open Biol*.*2017;7(9):170070. doi: 10.1098/rsob.170070. [ [PMC Free Article](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5627048) ] [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5627048/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/28878044)] [[CrossRef](https://doi.org/10.1098%2Frsob.170070" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+Open+Biol&title=+Progress+and+prospects+of+early+detection+in+lung+cancer&author=S+Blandin+Knight&author=PA+Crosbie&author=H+Balata&author=J+Chudziak&author=T+Hussell&volume=7&issue=9&publication_year=+2017&pages=170070&pmid=28878044&doi=10.1098/rsob.170070&)]
15. Ghani MU, Alam TM, Jaskani FH. Comparison of classification models for early prediction of breast cancer. 2019 International Conference on Innovative Computing (ICIC); Lahore, Pakistan: IEEE; 2019. p. 1-6. doi: 10.1109/ICIC48496.2019.8966691. [[CrossRef](https://doi.org/10.1109%2FICIC48496.2019.8966691" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar?q=Ghani+MU+Alam+TM+Jaskani+FH++Comparison+of+classification+models+for+early+prediction+of+breast+cancer.+2019+International+Conference+on+Innovative+Computing+(ICIC);+Lahore,+Pakistan:+IEEE;+2019.+p.+1-6+10.1109/ICIC48496.2019.8966691+)]
16. Ferreira P, Fonseca NA, Dutra I, Woods R, Burnside E. Predicting malignancy from mammography findings and image-guided core biopsies. International Journal of Data Mining and Bioinformatics*.*2015;11(3):257–76. doi: 10.1504/IJDMB.2015.067319. [ [PMC Free Article](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4764253) ] [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4764253/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/26333262)] [[CrossRef](https://doi.org/10.1504%2FIJDMB.2015.067319" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+International+Journal+of+Data+Mining+and+Bioinformatics&title=+Predicting+malignancy+from+mammography+findings+and+image-guided+core+biopsies&author=P+Ferreira&author=NA+Fonseca&author=I+Dutra&author=R+Woods&author=E+Burnside&volume=11&issue=3&publication_year=+2015&pages=257-76&pmid=26333262&doi=10.1504/IJDMB.2015.067319&)]
17. Brédart A, Kop JL, Antoniou AC, Cunningham AP, De Pauw A, et al. Clinicians’ use of breast cancer risk assessment tools according to their perceived importance of breast cancer risk factors: an international survey. J Community Genet*.*2019;10(1):61–71. doi: 10.1007/s12687-018-0362-8. [ [PMC Free Article](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC6325038) ] [[PMC free article](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6325038/)] [[PubMed](https://pubmed.ncbi.nlm.nih.gov/29508368)] [[CrossRef](https://doi.org/10.1007%2Fs12687-018-0362-8" \t "_blank)] [[Google Scholar](https://scholar.google.com/scholar_lookup?journal=+J+Community+Genet&title=+Clinicians%E2%80%99+use+of+breast+cancer+risk+assessment+tools+according+to+their+perceived+importance+of+breast+cancer+risk+factors:+an+international+survey&author=A+Br%C3%A9dart&author=JL+Kop&author=AC+Antoniou&author=AP+Cunningham&author=A+De+Pauw&volume=10&issue=1&publication_year=+2019&pages=61-71&pmid=29508368&doi=10.1007/s12687-018-0362-8&)]
18. **Contribution**

|  |  |
| --- | --- |
| IT Number | Contribution |
| IT20221928 | * Dataset collection * Encode the categorical columns. (preprocessing) * Train the model using Decision tree classification. * Train the model using Random Forest classification. * Report generation (abstract, data collection, evaluation) |
| IT20202422 | * Replace null values (preprocessing) * Train the model using Support vector machine classification. * Plot the data. * Report generation (problem statement, data mapping and preprocessing, future work) |
| IT20175702 | * Normalize numerical columns (preprocessing) * Plot the data. * Train the model using Logistic regression classification. * Report generation (methodology, description of dataset, selection of algorithm) |

1. **Appendix**

**7.1 Appendix A: Code Listening**

import pandas as pd

import numpy as np

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

from sklearn.preprocessing import LabelEncoder ,MinMaxScaler

from sklearn.metrics import mean\_squared\_error , accuracy\_score , confusion\_matrix ,recall\_score, f1\_score ,precision\_score

from sklearn.linear\_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.svm import SVC

import matplotlib.pyplot as plt

import seaborn as sns

# load the data into a pandas dataframe

df = pd.read\_csv("breast\_cancer\_data.csv")

# first five data

df.head()

# select all column with data type object

df\_obj = df.select\_dtypes(['object'])

# remove leading or trailing white spaces

df[df\_obj.columns] = df\_obj.apply(lambda x: x.str.strip())

# Replacing null value columns (text) with most used value

a1 = df.select\_dtypes(include=['object']).isnull().sum()

for i in a1.index:

    b1 = df[i].value\_counts().index.tolist()

    df[i] = df[i].fillna(b1[0])

# Replacing null value columns (int, float) with mean value

a1 = df.select\_dtypes(include=['integer','float']).isnull().sum()

for i in a1.index:

    df[i] = df[i].fillna(df[i].mean())

# assign target column

y = df.Status

#column to be ignored

drop\_cols = ['Status']

# assign feature columns

x = df.drop(drop\_cols,axis=1)

# plot all the column

plt.figure(figsize = (25,25))

plt.subplot(6,2,1)

sns.countplot(x = 'Race', palette='Set2', data = df)

plt.subplot(6,2,2)

sns.countplot(x = 'Marital Status', palette='Set2', data = df)

plt.subplot(6,2,3)

sns.countplot(x = 'T Stage ', palette='Set2', data = df)

plt.subplot(6,2,4)

sns.countplot(x = 'N Stage', palette='Set2', data = df)

plt.subplot(6,2,5)

sns.countplot(x = '6th Stage', palette='Set2', data = df)

plt.subplot(6,2,6)

sns.countplot(x = 'differentiate', palette='Set2', data = df)

plt.subplot(6,2,7)

sns.countplot(x = 'A Stage', palette='Set2', data = df)

plt.subplot(6,2,8)

sns.countplot(x = 'Estrogen Status', palette='Set2', data = df)

plt.subplot(6,2,9)

sns.countplot(x = 'Progesterone Status', palette='Set2', data = df)

plt.subplot(6,2,9)

sns.countplot(x = 'Grade', palette='Set2', data = df)

plt.subplot(6,2,11)

sns.countplot(x = 'Status', palette='Set2', data = df)

# plot the column against target column

plt.figure(figsize = (20,20))

plt.subplot(3,2,1)

sns.countplot(x = 'Status', hue= 'Race', palette='Set2', data = df)

plt.subplot(3,2,2)

sns.countplot(x = 'Status', hue= 'Marital Status', palette='Set2', data = df)

plt.subplot(3,2,3)

sns.countplot(x = 'Status', hue= 'differentiate', palette='Set2', data = df)

plt.subplot(3,2,4)

sns.countplot(x = 'Status', hue= 'Grade', palette='Set2', data = df)

plt.subplot(3,2,5)

sns.countplot(x = 'Status', hue= 'T Stage ', palette='Set2', data = df)

plt.subplot(3,2,6)

sns.countplot(x = 'Status', hue= 'N Stage', palette='Set2', data = df)

# encoding for Race Variable

race\_encoding = {"White":0,"Black":1,"Other":2}

x["Race"] = x["Race"].replace(race\_encoding)

# encoding for Marital Status Variable

marital\_status\_encoding = {"Married":0,"Single":1,"Divorced":2,"Widowed":3,"Separated":4}

x["Marital Status"] = x["Marital Status"].replace(marital\_status\_encoding)

# encoding for T Stage Variable

t\_stage\_encoding = {"T1":0,"T2":1,"T3":2,"T4":3}

x["T Stage "] = x["T Stage "].replace(t\_stage\_encoding)

# encoding for N Stage Variable

n\_stage\_encoding = {"N1":0,"N2":1,"N3":2}

x["N Stage"] = x["N Stage"].replace(n\_stage\_encoding)

# encoding for 6th Stage Variable

label\_encoder\_x= LabelEncoder()

x["6th Stage"]=label\_encoder\_x.fit\_transform(x["6th Stage"])

# encoding for differentiate Variable

x["differentiate"]=label\_encoder\_x.fit\_transform(x["differentiate"])

# encoding for Grade Variable

x["Grade"]=label\_encoder\_x.fit\_transform(x["Grade"])

# encoding for A Stage Variable

a\_stage\_encoding = {"Distant":0,"Regional":1}

x["A Stage"] = x["A Stage"].replace(a\_stage\_encoding)

# encoding for Estrogen Status Variable

estrogen\_stage\_encoding = {"Negative":0,"Positive":1}

x["Estrogen Status"] = x["Estrogen Status"].replace(estrogen\_stage\_encoding)

# encoding for Progesterone Status Variable

progesterone\_status\_encoding = {"Negative":0,"Positive":1}

x["Progesterone Status"] = x["Progesterone Status"].replace(progesterone\_status\_encoding)

# encoding for Status Variable

status\_encoding = {"Alive":0,"Dead":1}

y = y.replace(status\_encoding)

# Normalize numerical columns

scaler = MinMaxScaler()

# Define the numerical columns

numerical\_cols = ['Age', 'Tumor Size', 'Regional Node Examined', 'Reginol Node Positive','Survival Months']

for col in numerical\_cols:

    encoded\_col = scaler.fit\_transform(df[[col]])

    x[col] = encoded\_col

x

# remove headers

x = x.iloc[:,:-1].values

# split data into traing and testing

X\_train, X\_test, y\_train, y\_test = train\_test\_split(x, y,test\_size=.25, random\_state=42)

# Logistic Regression

lgr = LogisticRegression()

# train the model

lgr.fit(X\_train, y\_train)

# make prediction

test\_pred = lgr.predict(X\_test)

# outputs

print("Mean squared for logistic regression is : ",round(mean\_squared\_error(y\_test,test\_pred),3))

print("Accuracy score for logistic regression is : ",round(accuracy\_score(y\_test,test\_pred),3))

print("f1 score for logistic regression is : ",round(f1\_score(y\_test,test\_pred),3))

print("recall for logistic regression is : ",round(recall\_score(y\_test,test\_pred),3))

print("precision score for logistic regression is : ",round(precision\_score(y\_test,test\_pred),3))

print("Confusion matrix for logistic regression is : ",confusion\_matrix(y\_test,test\_pred))

# visualize the confusion matrix

plt.clf()

cm = confusion\_matrix(y\_test,test\_pred)

plt.imshow(cm,interpolation='nearest',cmap=plt.cm.Wistia)

plt.title('Confusion Matrix-Test Data')

s = [['TN','FP'],['FN','TP']]

for k in range(2):

    for l in range(2):

        plt.text(l,k,str(s[k][l])+"= "+str(cm[k][l]))

plt.show()

# decision tree classification

dc\_clf = DecisionTreeClassifier()

# train with decision tree

dc\_clf.fit(X\_train, y\_train)

# make prediction

dc\_pred = dc\_clf.predict(X\_test)

# outputs

print("Mean squared for decision tree is : ",round(mean\_squared\_error(y\_test,dc\_pred),3))

print("Accuracy score for decision tree is : ",round(accuracy\_score(y\_test,dc\_pred),3))

print("f1 score for decision tree is : ",round(f1\_score(y\_test,dc\_pred),3))

print("recall for decision tree is : ",round(recall\_score(y\_test,dc\_pred),3))

print("precision score for decision tree is : ",round(precision\_score(y\_test,dc\_pred),3))

print("Confusion matrix for decision tree is : ",confusion\_matrix(y\_test,dc\_pred))

# visualize the confusion matrix

plt.clf()

cm = confusion\_matrix(y\_test,dc\_pred)

plt.imshow(cm,interpolation='nearest',cmap=plt.cm.Wistia)

plt.title('Confusion Matrix-Test Data')

s = [['TN','FP'],['FN','TP']]

for k in range(2):

    for l in range(2):

        plt.text(l,k,str(s[k][l])+"= "+str(cm[k][l]))

plt.show()

# random forest classifier

rf\_clf = RandomForestClassifier(n\_estimators=10, max\_depth=None,min\_samples\_split=2, random\_state=0)

# train the model

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

# make prediction

rf\_pred = rf\_clf.predict(X\_test)

# output

print("Mean squared for random forest is : ",round(mean\_squared\_error(y\_test,rf\_pred),3))

print("Accuracy score for random forest is : ",round(accuracy\_score(y\_test,rf\_pred),3))

print("f1 score for random forest is : ",round(f1\_score(y\_test,rf\_pred),3))

print("recall for random forest is : ",round(recall\_score(y\_test,rf\_pred),3))

print("precision score for random forest is : ",round(precision\_score(y\_test,rf\_pred),3))

print("Confusion matrix for random forest is : ",confusion\_matrix(y\_test,rf\_pred))

# visualize the confusion matrix

plt.clf()

cm = confusion\_matrix(y\_test,rf\_pred)

plt.imshow(cm,interpolation='nearest',cmap=plt.cm.Wistia)

plt.title('Confusion Matrix-Test Data')

s = [['TN','FP'],['FN','TP']]

for k in range(2):

    for l in range(2):

        plt.text(l,k,str(s[k][l])+"= "+str(cm[k][l]))

plt.show()

# support vector machine

svm\_clf = SVC()

# train the model

svm\_clf.fit(X\_train, y\_train)

# make prediction

svm\_pred = svm\_clf.predict(X\_test)

# outputs

print("Mean squared for support vector machine is : ",round(mean\_squared\_error(y\_test,svm\_pred),3))

print("Accuracy score for support vector machine is : ",round(accuracy\_score(y\_test,svm\_pred),3))

print("f1 score for support vector machine is : ",round(f1\_score(y\_test,svm\_pred),3))

print("recall for support vector machine is : ",round(recall\_score(y\_test,svm\_pred),3))

print("precision score for support vector machine is : ",round(precision\_score(y\_test,svm\_pred),3))

print("Confusion matrix for support vector machine is : ",confusion\_matrix(y\_test,svm\_pred))

# visualize the confusion matrix

plt.clf()

cm = confusion\_matrix(y\_test,svm\_pred)

plt.imshow(cm,interpolation='nearest',cmap=plt.cm.Wistia)

plt.title('Confusion Matrix-Test Data')

s = [['TN','FP'],['FN','TP']]

for k in range(2):

    for l in range(2):

        plt.text(l,k,str(s[k][l])+"= "+str(cm[k][l]))

plt.show()

**7.1 Appendix B: Medical report of Lavaniyah**

