ANALYSING DATA ON BREAST CANCER FOR WOMEN USING MACHINE LEARNING MODEL

P. VAISHNAVI
SCSE, M.Tech Integrated Software Engineering, (Student) VIT University Chennai, Tamil Nadu, India
vaishnavi.pothugunta26@gmail.com

Y.L. SAI CHARITHA
SCSE, M.Tech Integrated Software Engineering, (Student) VIT University Chennai, Tamil Nadu, India
saicharithayallarubailu@gmail.com

T. JAYASRI
SCSE, M.Tech Integrated Software Engineering, (Student) VIT University Chennai, Tamil Nadu, India
jayasrireddy19@gmail.com

***Abstract*--- Nowadays due to the changing lifestyles and the excessive stress among the women folk, there is a rise in the risk of developing Breast Cancer (specifically women between the ages of 16 and 45). In this project titles "Role of metrics in dealing away with cancer" and our specific field of research. Being Cancer in woman (Breast cancer), we intend to bring the role of metrics in Breast Cancer into the lime light in order to be able to be clear about the different areas we can develop on with respect to putting metrics to effective usage in developing better ways to eradicate Breast Cancer and the end the agony among the women folk.**

***Keywords*--- Decision tree, R studio, accuracy, error, Models, Confusion matrix, Visualization.**

I. INTRODUCTION

Breast Cancer being a sensitive topic among the women folk is often not being addressed widely especially in third world countries and the developing countries (Asian and African countries) where the probability of Breast Cancer is about less than a million per year which in comparison to other countries is high. **Malignant growth/ Cancer** is the name given to an assortment of related infections. In such a wide range of cases with malignant growth, a part of our body cells starts to isolate ceaselessly and it also happens to spread into encompassing tissues. Malignancy, it can begin anywhere in human body, which comprises of more than trillions of cells. **Breast Cancer growth** is disease that structures in the cells of the breast. After skin malignant growth, Breast disease is the most widely recognized malignant growth analysed in ladies in the United States. Breast malignant growth can happen in the two people, however it's unquestionably more normal in ladies.

II. OBJECTIVE

- ✓ Identifying the best suitable model with highest accuracy and lowest overall error rates.
- ✓ Understanding the various metrics for all the models.

III. ACKNOWLEDGEMENT

The methodology of this research includes comparing the accuracies of different model of estimation such as decision tree model, extreme boost model, random forest model, neural net model and linear model using a common dataset based off of Breast Cancer and hence conclude on the best model with the highest accuracy of raw data, accuracy of data before and after tuning in. We intend to establish a relation and come to a conclusion on the best model that analyses the data and thereby come obtain clarity in examining and analysing the data effectively to come to a conclusion regarding Breast Cervical Cancer quicker.

IV. LITERATURE REVIEW

[1] In this paper, the evaluated the prognostic and predictive potential of DNA parameters in early and advanced breast cancer. Groups consisted of 150 and 16 breast cancer patients under adjuvant and neoadjuvant therapy respectively, 34 patients with metastatic disease and 35 healthy volunteers, a

multiple logistic regression models are constructed and had been discriminating between individuals who are patients and healthy ones. [2] This paper shows the purpose of this study was to evaluate easy and helpful data classification and data analysis for a regular recorded dataset consisting the records of women with breast cancer and network classification for determining the breast cancer response to promote the detection in many useful ways. [3] Among women, breast cancer is a leading cause of death. Breast cancer with these kind of risk predictions can be informed through screening and with preventative actions. The paper explains about Breast Cancer Risk Prediction Tool (BCRAT) an implementation of the Gail model. [4] A number of models have been developed for assessing these risks with varying degrees of validation. With further improving knowledge of how to integrate risk factors together and to gradually integrate further genetic variants into these mentioned models, also we are confident we would be able to discriminate with farther great accuracy which specific women are most likely to develop breast cancer. [5] Racial and ethnic minorities as well as other vulnerable populations experience disparate cancer-related health outcomes. [6] This paper shows the purpose of this study was to evaluate easy and helpful data classification and data analysis for a regular recorded dataset consisting the records of women with breast cancer and network classification for determining the breast cancer response to promote the detection in many useful ways. The article's literature depicts that the outcome disparities are related to patient, individual and health individual service-factors. Lack of insurance, fear of testing, delay in seeking care, and unfavourable tumour characteristics all contribute to disparities at the patient level.[7] This study can help in understanding the importance of screening program for cancer which is free in the improvement of earlier breast cancer detection, service, treatments and survival in poor urban places. [8] Among women, breast cancer is a leading cause of death. Breast cancer with these kind of risk predictions can be informed through screening and with preventative actions. Likewise, prospective data that are being collected to determine whether Patient Navigators effects treatment and does it influence appointment adherence and as well as the underlying many reasons for barriers to proper interventions in this underserved urban minority population.[9] A full-time patient navigator supported patients using the care management model. The unequal weight of cancer is highlighted among lower socioeconomic status and racial/ethnic minority women who suffer higher mortality from breast cancer compared with their more affluent non-Hispanic white counterparts. [10] The aim of this study was to optimize learning algorithm. In this context, we happen to use the well-known programming technique and models to select the best perfect features and parametric model values of the machine learning classifiers. The present study depicts that the programming can function automatically and find its suitable model by combining both feature pre-processing in models and methods classifier algorithms. [11] This paper reviewed and examined the resources and methods used in compiling the estimates and nation wise cancer incidence and briefly explains the key conclusions by cancer site and in 20 major areas in the world. [12] In screening, for breast cancer its efficacy is dependent on the interpretations done by radiologist, rather the fact that it has occurred in the proved value for mammography in screening. The values in such interpretations are not very well understood. [13] This study aimed to evaluate whether radionics can improve the diagnostic performance of mammography compared with that obtained by experienced radiologist. [14] The performed experiments are on the datasets of mammographic Image Analysis Society (MIAS) and (DDSM) Digital Database for Screening Mammography. On MIAS, there is 92 percent sensitivity reached at 1.94 false positive per image (FPI) and On DDSM 93.84 percent at 2.21 FPI. The framework has achieved a better performance compared with other algorithms. [15] From data science perspectives, data mining technology is used to uncover the disease according to some parameters like BMI, age and sugar routine database. There has been deployment for those technologies and resulted in giving the best and considerable results that may help much for breast cancer aid. In this paper, datasets are collected and examined according to 10 predictors. Detailed information about performance metrics. [16] This breast cancer can recur anytime in the survivors of breast cancer, however basically it returns in the initial three to five years after the treatment. This paper reviewed and examined the resources and methods used in compiling the estimates and nation wise cancer incidence and briefly explains the key conclusions by cancer site and in 20 major areas in the world. [17] The goal of this project is to find one or more methods to solve the problem. The suitable models are selected using performance metrics and accuracy such as the area under the curve and Precision-Recall covering area and prediction accuracy. [18] Need to develop an automated system based on metrics and machine learning, begin and malignant tumours are classified into LRO, BNK, MLP, and SNO. [19] In the analysis of the open-access dataset, the proposed model has a distinctive feature in classifying breast cancer based on the performance metrics. Then, it was classification which could be associative, On CBAR product-based software developed produced very successful predictions in the detection of breast cancer classification. [20] This study aimed to evaluate whether radionics can improve the diagnostic performance of mammography compared with that obtained by experienced radiologist. Analyse absolute concentration and structural changes of metabolites in different brain regions using very best neuroimaging and featured technology, and could brief the correlation between them, if exists.

V. PROPOSED METHODOLOGY

The proposed methodology of this research includes identifying the best suitable model with highest accuracy and overall lower error rates among all the other models.

The used machine learning models are

- ✓ ***Decision Tree model***: An algorithm is considered as Model of computation that to be basically a decision tree, i.e., a sequence of branching operations based on comparisons of some quantities, the comparisons are being assigned as unit computational cost.
- ✓ ***Random Forest Model***: It is an ensemble learning method for classification, Regression and other tasks that operate by constructing a multitude of decision Tree at training time.
- ✓ ***SVM (Support Vector Machine Model)***: (SVMs or support-vector networks) support-vector machines are supervised learning models with associated learning algorithms.
- ✓ ***Linear Model:*** It describes response in a continuous variable as a function of one or more predictor variables. They also can help you to understand and predict the behaviour of complex systems or could analyses experimental, financial, and biological data from sources.

VI. IMPLEMENTATION

Research includes dataset "**Breast Cancer Data Set**" Year of release and subsequent updates: 1988, 2000 and 2019.
Database URL: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer. The above attached dataset is provided by the Oncology Institute which is one of three domains. This data set comprises of 201 and 85 instances of both the respective classes. The instances are categorised using 9 attributes, includes of linear and nominal.
Data Set Characteristics: Multivariate
Attribute Characteristics: Categorical
Attribute Information:
1. Class: no-recurrence-events, recurrence-events
2. age: with 9 classes (10-19), (20-29), (30-39), (40-49), (50-59), (60-69), (70-79), (80- 89), (90-99).
3. Menopause: lt40, ge40, premenopausal.
4. tumour size: with 9 classes 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
5. inv nodes: with 13 classes 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
6. node-caps: (type) yes, no.
7. degree-malignancy: 1, 2, 3.
8. Breast: left, right.
9. breast-quad: left-up, left-low, right-up, right-low, central.
10. Irradiate: yes, no.

About Dataset:

Characteristics of Data Set:	Multivariate	No. of Instances :	286
Characteristics of Attribute:	Categorical	No. of Attributes :	9
Tasks Associated:	Classification	Area :	Life

Software used: **R Studio (Programming tool)**

R code for the loading dataset as data frame

Loading the dataset as a data frame. The current working directory is in the same directory where we stored the dataset .

```
> uciurl <- "https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data"
> download.file(url=uciurl, destfile="breast-cancer-wisconsin.data", method="curl")
# read the data
> data <- read.table("breast-cancer-wisconsin.data", na.strings = "?", sep=",")
> str(data)
## 'data.frame': 699 obs. of 11 variables:
## >>$-V1 :
int 1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561 1033078 1033
078 ...
```

R code for putting the names in columns and Summarize data

```
names(data) <- c("ClumpThickness",
"UniformityCellSize",
"UniformityCellShape",
"MarginalAdhesion",
"SingleEpithelialCellSize",
"BareNuclei",
"BlandChromatin",
"NormalNucleoli",
"Mitoses",
"Class")
data$Class <- factor(data$Class, levels=c(2,4), labels=c("benign", "malignant"))
print(summary(data))
## ClumpThickness UniformityCellSize UniformityCellShape MarginalAdhesion
## Min. : 1.000 Min. : 1.000 Min. : 1.000 Min. : 1.000
## 1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 1.000 1st Qu.: 1.000
## Median : 4.000 Median : 1.000 Median : 1.000 Median : 1.000
## Mean : 4.418 Mean : 3.134 Mean : 3.207 Mean : 2.807
## 3rd Qu.: 6.000 3rd Qu.: 5.000 3rd Qu.: 5.000 3rd Qu.: 4.000
## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000
```

```
##

## SingleEpithelialCellSize BareNuclei BlandChromatin
## Min. : 1.000 Min. : 1.000 Min. : 1.000
## 1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 2.000
## Median : 2.000 Median : 1.000 Median : 3.000
## Mean : 3.216 Mean : 3.545 Mean : 3.438
## 3rd Qu.: 4.000 3rd Qu.: 6.000 3rd Qu.: 5.000
## Max. :10.000 Max. :10.000 Max. :10.000
## NA's :16

## NormalNucleoli Mitoses Class
## Min. : 1.000 Min. : 1.000 benign :458
## 1st Qu.: 1.000 1st Qu.: 1.000 malignant:241
## Median : 1.000 Median : 1.000
## Mean : 2.867 Mean : 1.589
## 3rd Qu.: 4.000 3rd Qu.: 1.000
## Max. :10.000 Max. :10.000
```

R code to Split dataset

Splitting the dataset into a training (70%) and a validation set (30%). To compare later different models or the same models trained with different parameters, we use the same training and validation set. Since we are splitting them randomly, we set a seed so that we maintain the same split throughout our experiments.

```
set.seed(1234)
ind <- sample(2, nrow(data), replace=TRUE, prob=c(0.7, 0.3))
trainData <- data[ind==1,]
validationData <- data[ind==2,]
```

R code to Train Dataset

```
library(rpart)
library(rpart.plot)
library(party)
```

Producing a decision tree by training the induction algorithm on the train dataset.

```
tree = rpart(Class ~ ., data=trainData, method="class")
print(tree)
```

```
## n= 485
## node), split, n, loss, yval, (yprob)
## * denotes terminal node
## 1) root 485 171 benign (0.647422680 0.352577320)
## 2) UniformityCellSize< 2.5 292 9 benign (0.969178082 0.030821918)
## 4) BareNuclei< 4.5 280 1 benign (0.996428571 0.003571429) *
## 5) BareNuclei>=4.5 12 4 malignant (0.333333333 0.666666667) *
## 3) UniformityCellSize>=2.5 193 31 malignant (0.160621762 0.839378238)
## 6) UniformityCellShape< 2.5 15 3 benign (0.800000000 0.200000000) *
## 7) UniformityCellShape>=2.5 178 19 malignant (0.106741573 0.893258427)
## 14) UniformityCellSize< 4.5 45 14 malignant (0.311111111 0.688888889)
## 28) BareNuclei< 2.5 10 2 benign (0.800000000 0.200000000) *
## 29) BareNuclei>=2.5 35 6 malignant (0.171428571 0.828571429) *
## 15) UniformityCellSize>=4.5 133 5 malignant (0.037593985 0.962406015)
```

Visual Representation

```
plot(tree)
text(tree)
```

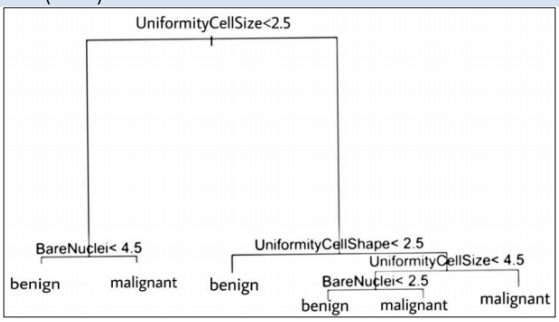

Figure 1 - Decision tree by training the induction algorithm on the train dataset

Decision tree #1

```
rpart.plot(tree, extra = 104, nn = TRUE)
```

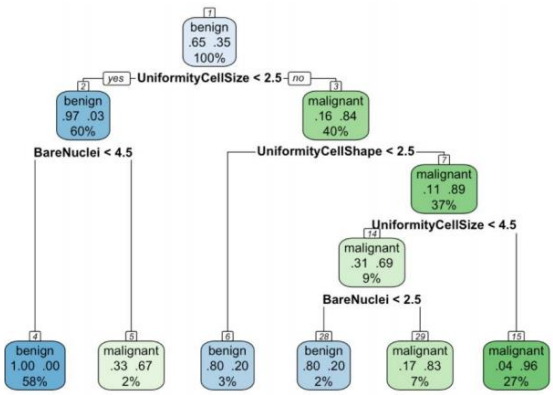

Figure 2 - #1 decision tree, class, tree
Ref: MALIGNANT refers to YES (presence of Breast Cancer) & BENIGN refers to NO Breast Cancer

R code to display parameters

```
rpart.control()
## $minsplit
## [1] 20
##
## $minbucket
## [1] 7
##
## $cp
## [1] 0.01
## $maxcompete
## [1] 4
## $maxsurrogate
## [1] 5
## $usesurrogate
## [1] 2
## $surrogatestyle
## [1] 0
## $maxdepth
## [1] 30
## $xval
## [1] 10
```

Decision tree #2

```
tree_with_params1 =
rpart(Class ~ ., data1=trainData1, method="class", minsplit = 1,
minbucket = 1, cp = -1)
rpart.plot(tree_with_params, extra = 104, nn = TRUE)
```

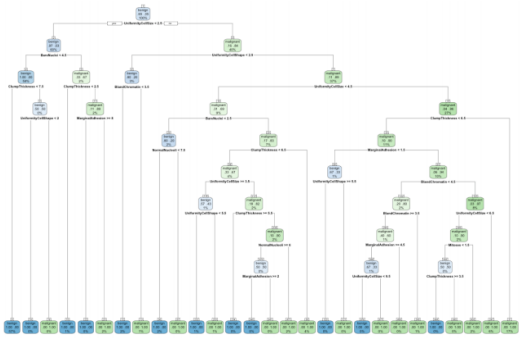

Figure 3 - #2 Decision tree, class, tree with params

Decision tree #3

```
library(party)
ctree = ctree(Class ~ ., data=trainData)
print(ctree)
```

```
##
## Conditional inference tree with 7 terminal nodes
##
## Response: Class
## Inputs: ClumpThickness, UniformityCellSize, UniformityCellShape, MarginalAdhe
sion, SingleEpithelialCellSize, BareNuclei, BlandChromatin, NormalNucleoli, Mitos
es
## Number of observations: 485
##
## 1) UniformityCellSize <= 2; criterion = 1, statistic = 323.402
## 2) BareNuclei <= 4; criterion = 1, statistic = 169.767
## 3) SingleEpithelialCellSize <= 2; criterion = 1, statistic = 29.834
## 4)* weights = 264
## 3) SingleEpithelialCellSize > 2
## 5)* weights = 16
## 2) BareNuclei > 4
```

```
## 6)* weights = 12
## 1) UniformityCellSize > 2
## 7) BareNuclei <= 3; criterion = 1, statistic = 42.361
## 8) UniformityCellSize <= 4; criterion = 1, statistic = 23.999
## 9)* weights = 27
## 8) UniformityCellSize > 4
## 10)* weights = 20
## 7) BareNuclei > 3
## 11) BlandChromatin <= 4; criterion = 0.996, statistic = 12.225
## 12)* weights = 50
## 11) BlandChromatin > 4
## 13)* weights = 96
```

```
plot(ctree, type="simple")
```

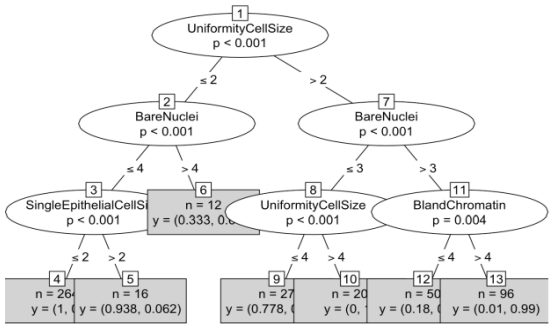

Figure 4 - #3 Decision tree, response, ctree

EVALUATION OF DECISION TREES

Printing the results in the validation set, creating a function that get as input a tree model, the validation data and the type of the tree and printing results in the console.

```
evaluation <- function(model, data, atype) {
 cat("\nConfusion matrix:\n")
 prediction = predict(model, data, type=atype)
 xtab = table(prediction, data$Class)
 print(xtab)
 cat("\nEvaluation:\n\n")
 accuracy = sum(prediction == data$Class)/length(data$Class)
 precision = xtab[1,1]/sum(xtab[,1])
 recall = xtab[1,1]/sum(xtab[1,])
 f = 2 * (precision * recall) / (precision + recall)
 cat(paste("Accuracy:\t", format(accuracy, digits=2), "\n",sep=" "))
 cat(paste("Precision:\t", format(precision, digits=2), "\n",sep=" "))
 cat(paste("Recall:\t\t", format(recall, digits=2), "\n",sep=" "))
 cat(paste("F-measure:\t", format(f, digits=2), "\n",sep=" "))}
evaluation(tree, validationData, "class")
evaluation(entTree, validationData, "class")
evaluation(tree_with_params, validationData, "class")
evaluation(ctree, validationData, "response")
```

#1 class

```
## Confusion matrix:
```

prediction	benign	malignant
benign	138	6
malignant	6	64

```
## Evaluation:
## Accuracy: 0.94
## Precision: 0.96
## Recall: 0.96
## F-measure: 0.96
```

#2 class

```
## Confusion matrix:
```

prediction	benign	malignant
benign	138	5
malignant	6	65

```
## Evaluation:
## Accuracy: 0.95
## Precision: 0.96
## Recall: 0.97
## F-measure: 0.96
```

#3 class

```
## Confusion matrix:
```

prediction	benign	malignant
benign	139	12
malignant	5	58

Evaluation:

Accuracy: 0.92

Precision: 0.97

Recall: 0.92

F-measure: 0.94

#4 response

Confusion matrix:

prediction	benign	malignant
benign	138	6
malignant	6	64

Evaluation:

Accuracy: 0.94

Precision: 0.96

Recall: 0.96

F-measure: 0.96

VII. RESULTS

- *OVERALL ERROR RATES & ACCURACY*

Total number of nodes: 8

proportion	Decision tree	ADA boost	Random forest	SVM	Neural network	linear
DT1(20 %)	91.4%	94.2 %	90.6%	18.8 %	92.4 %	91.6 %
DT2(15 %)	53.2%	72.8 %	82.8%	12.3 %	54.2 %	83.9 %
DT3(15 %)	48.8%	57.4 %	66.4%	8.55 %	49.8 %	67.4 %
DT4(42 %)	74.3%	56.8 %	45.6%	31.0 %	75.4 %	46.7 %
DT5(20 %)	91.0%	94.2 %	90.5%	18.7 %	92.4 %	91.6 %
DT6(22 %)	53.2%	72.8 %	72.9%	14.5 %	44.2 %	83.9 %
DT7(18 %)	38.1%	57.4 %	66.4%	8.55 %	49.2 %	67%
DT8(45 %)	4.3%	6.8 %	5.6%	1.08 %	5.4%	6.7 %
100% - TOTAL ACCURACY						**82%**

- *EXECUTING DECISION TREE PARTITIONS SEPERATELY*

proportion	Decision tree	ADA boost	Random forest	SVM	Neural network	linear
DT1(20 %)	91.4%	87.6 %	87.5%	94.2 %	90.6 %	18.8 %
DT2(15 %)	53.2%	77.6 %	79.7%	72.8 %	82.8 %	12.3 %
DT3(15 %)	48.8%	54.5 %	77.4%	57.4 %	66.4 %	8.55 %
DT4(42 %)	74.3%	61.2 %	82.5%	56.8 %	45.6 %	31.0 %
DT5(20 %)	74.8%	76.6 %	82.3%	80.0 %	57.7 %	6.04 %
DT6(22 %)	78.4%	78.4 %	83.5%	73.0 %	58.0 %	6.24 %
DT7(18 %)	51.0%	50.0 %	74.0%	61.0 %	67.0 %	2.6%
DT8(45 %)	53.0%	50.0 %	90.0%	53.0 %	57.0 %	1.0%
100% - TOTAL ACCURACY						**86.4 %**

- *RESULT #1 - EXECUTING DECISION TREE PARTITIONS SEPERATELY*

Number of instances (samples)	**286**
Number of attributes (columns)	**9**

For the above characteristics of the dataset, Decision tree partitions were performed separately and individual values were obtained for each node and each model. The Overall Accuracy before and after Separate Training are obtained and hence the difference is computed along with the nature of rate of change.

Overall accuracy before separate training:	**82%**
Overall accuracy After separate training:	**86.4%**

Nature of change of Overall accuracy:	**Increase (^)**
Rate of change of Overall accuracy:	**4.1%**

- *TUNING THE DECISION TREES*

Proportion	**Model**	**Model Tuning**	**Accuracy after tuning (%)**	**Overall Accuracy**
DT1 (25%)	Decision Tree	Min Bucket	99.00	33.90
DT2 (19%)	Random Forest	Number of Variables	82.60	7.450
DT3 (17%)	Random Forest	Number of Variables	78.50	5.987
DT4 (13%)	Random Forest	Number of Variables	81.30	06.80
DT5 (11%)	Random Forest	Number of Variables	85.70	04.98
DT6 (09%)	SVM	Laplacian (laplacedot)	93.50	02.87
DT7 (05%)	Random Forest	Number of Variables	83.00	09.75
100% - total accuracy				**91.497%**

- *RESULT #2 – AFTER TUNING THE DECISION TREES*

Number of instances (samples)	**286**
Number of attributes (columns)	**9**

For above mentioned characteristics of the dataset, Decision tree were tuned and individual values were obtained for each node and each model. The Overall Accuracy before and after Tuning are obtained and hence the difference is computed along with the nature of rate of change.

Overall accuracy before separate tuning:	**86.4%**
Overall accuracy After separate tuning:	**91.497%**
Nature of change of Overall accuracy:	**Increase (^)**
Rate of change of Overall accuracy:	**4.2%**

VIII. CONCLUSION

- **SVM** is the best model for measuring the dataset because it had the highest accuracy rates and lowest overall error rates among all the other models.

The values are as follows:

General Accuracy-	82.00%
Accuracy before Tuning (after executing decision tree partitions separately)-	86.40%
Accuracy after Tuning-	93.50%
Overall Accuracy-	02.87%

- Accuracy of raw data set is 82%
- Accuracy before tuning the data was 86.4% (By separating leaf nodes)
- Accuracy after tuning the data is 91.497%

REFERENCES

[1] Panagopoulou, M., Karaglani, M., Balgkouranidou, I. et al. *Circulating cell-free DNA in breast cancer: size profiling, levels, and methylation patterns lead to prognostic and predictive classifiers*. Oncogene 38, 3387–3401 (2019). https://www.nature.com/articles/s41388-018-0660-y#citeas

[2] *Determination of Breast Cancer Response to Bevacizumab Therapy Using Contrast- Enhanced Ultrasound and Artificial Neural Networks Kenneth Hoyt*, PhD, Jason M. Warram, BS, Heidi Umphrey, MD, Lin Belt, RDMS, Mark E. Lockhart, MD, Michelle L. Robbin, MD, and Kurt R. Zinn, DVM, PhD, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3122922/

[3] Stark GF, Hart GR, Nartowt BJ, Deng J (2019) *Predicting breast cancer risk using personal health data and machine learning models*. PLoS ONE 14(12): e0226765. https://doi.org/10.1371/journal.pone.0226765

[4] Evans, D.G.R., Howell, *A. Breast cancer risk-assessment models. Breast Cancer Res 9*, 213 (2007). https://doi.org/10.1186/bcr1750

[5] *Metrics for evaluating patient navigation during cancer diagnosis and treatment*: crafting a policy-relevant research agenda for patient navigation in cancer care B. Ashleigh Guadagnolo, MD, MPH,1 Daniel Dohan, PhD,2 Peter Raich, MD,3 and For the ACS Patient Navigation Leadership Summit.

[6] Blackman DJ, Masi CM. *Racial and ethnic disparities in breast cancer mortality*: are we doing enough to address the root causes? J Clin Oncol. 2006;24(14):2170–8.

[7] Oluwole SF, Ali AO, Adu A, Blane BP, Barlow B, Oropeza R, et al. *Impact of a cancer screening program on breast cancer stage at diagnosis in a medically underserved urban community*. J Am Coll Surg. 2003;196(2):180–8. [Google Scholar]

[8] Gabram SG, Lund MJ, Gardner J, Hatchett N, Bumpers HL, Okoli J, et al. *Effects of an outreach and internal navigation program on breast cancer diagnosis in an urban cancer center with a large African-American population*. Cancer. 2008;113(3):602–7. [Google Scholar]

[9] Battaglia TA, Roloff K, Posner MA, Freund KM. *Improving follow-up to abnormal breast cancer screening in an urban population. A patient navigation intervention Cancer*. 2007;109(2 Suppl):359–67. [Google Scholar]

[10] Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, Mohammed Faisal Nagi, "*Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms*", Journal of Healthcare Engineering, vol. 2019, Article ID 4253641, 11 pages, 2019. https://doi.org/10.1155/2019/4253641

[11] Ferlay, I. Soerjomataram, R. Dikshit et al., "*Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012*," International Journal of Cancer, vol. 136, no. 5, pp. 359–389, 2014.View at: Google Scholar

[12] J. G. Elmore, C. K. Wells, C. H. Lee, D. H. Howard, and A. R. Feinstein, "*Variability in radiologists' interpretations of mammograms*," New England Journal of Medicine, vol. 331, no. 22, pp. 1493–1499, 1994.View at: Publisher Site Google Scholar

[13] N. Mao, P. Yin, Q. Wang et al., "Added value of radiomics on mammography for breast cancer diagnosis: a feasibility study," Journal of the American College of Radiology, vol. 16, no. 4, pp. 485–491, 2019.View at: Google Scholar

[14] H. Wang, J. Feng, Q. Bu et al., "*Breast mass detection in digital mammogram based on gestalt psychology*," Journal of Healthcare Engineering, vol. 2018, Article ID 4015613, 13 pages, 2018.View at: Google Scholar

[15] Classification of Breast Cancer Using Data Mining Farah Sardouka *, Dr. Adil Deniz Durub , Dr. Oğuz Bayatc https://core.ac.uk/download/pdf/235050738.pdf

[16] Umesh D R ; B Ramachandra '*Association rule mining based predicting breast cancer recurrence on SEER breast cancer data'*

[17] https://www.researchgate.net/publication/335079111_Breast_Cancer_diagnosis_using_machine_learning_classification_methods_using_Hadoop

[18] Breast Cancer Predictive Analytics Using Supervised Machine Learning Techniques https://www.researchgate.net/publication/338778878_Breast_Cancer_Predictive_Analytics_Using_Supervised_Machine_Learning_Techniques

[19] Arslan, a, Tunç, Z, Balıkçı Çiçek, İ, Çolak, C. "*a novel interpretable web-based tool on the associative classification methods: an application on breast cancer dataset*". The Journal of Cognitive Systems 5 (2020): 33-40 https://dergipark.org.tr/en/pub/jcs/issue/55836/770164

[20] Tong, T., Lu, H., Zong, J. et al. *Chemotherapy-related cognitive impairment in patients with breast cancer based on MRS and DTI analysis. Breast Cancer 27*, 893–902 (2020). https://doi.org/10.1007/s12282-020-01094-z

Sources

Introduction www.cancer.gov, www.mayoclinic.org

Dataset https://archive.ics.uci.edu

Decision tree www.researchgate.net

Scatter Plot http://www.alcula.com/calculators/statistics/scatter-plot/ https://chartio.com/learn/charts/what-is-a-scatter-plot/#:~:text=A%20scatter%20plot%20(aka%20scatter,to%20obser ve%20relationships%20between%20variables

Box Plot http://www.alcula.com/calculators/statistics/box-plot/ https://en.wikipedia.org/wiki/Box_plot

Dataset – Creators : Matjaz Zwitter & Milan Soklic (physicians), Institute of Oncology University Medical Center Ljubljana, Yugoslavia.