EUNICE CHUA
153775200
BAN210ZBB - FINAL PROJECT

BREAST CANCER DATASET

The analysis of the Breast Cancer Dataset revolved on using machine learning in order to determine the factor/s have influence and predict the probability of breast cancer occurrence. Several algorithms were run to generate the best predictive model (with minimal bias and variance) such as decision trees and regression. Overall, the best predictive model for this dataset is the Probability Tree model garnering the lowest Average Squared Error and the Misclassification Rate. This will be beneficial in determining the probability of symptoms detected as breast cancer earlier so that appropriate medical prognosis will be delivered to the patient. Also, the PA results could raise awareness of the benefits machine learning can do for the health industry. Moreover, the PA model will be beneficial most especially to the medical workers to accurately determine the recurrence of cancer even just with the degree of the malignant tumor.

EUNICE CHUA
153775200
BAN210ZBB - FINAL PROJECT

Exploratory Data Analysis

Using SAS EM and Python, the following EDA techniques were run:

1. Identification of variables and their data types
2. Descriptive statistics
3. Finding Missing values
4. Checking for Unique values
5. Correlation Analysis

Metadata

Table 1 shows the attributes of the dataset. The Breast Cancer Dataset includes 286 rows and 11 columns. There are 2 numeric and 9 categorical types of variables.

Table 1. Metadata

VARIABLE	DESCRIPTION
Class	The target variable of the dataset. Includes two classes: no-recurrence-events and recurrence-events. This indicates the existence of cancer in the patient.
Age	Age of the patient at the time of diagnosis. Grouped into 9 classes with 10 years each class from age 10 to 99 years old.
Menopause	Stage of menopausal of the patient at the time of diagnosis. It has 3 classes: lt40, ge40, premenopausal
Tumor-Size	Size of tumor per 5 millimeters from 0 to 59 millimeters grouped into 12 classes.
Inv-Nodes	Number of extra lymph nodes from 0 to 39, divided into 3 nodes per class with total of 13 classes.
Node-Caps	Presence of lymph nodes on capsule, 2 classes: yes and no.
Deg-Malig	The tumor's grade with 3 having the highest number of abnormal cells.

	There are 3 classes: 1, 2, 3
Breast	Affected breast where the tumor is present. There are 2 classes: left and right
Breast-Quad	Part of the breast if divided into quadrant. There are 5 classes: left-up, left-low, right-up, right-low, central
Irradiat	If patient undergoes radiation therapy treatment. There are 2 classes: yes and no.

Descriptive Statistics

Figure 1 shows that the majority of the patients are aged between 50-59 years old, followed by patients aging 40 to 49 years old and was diagnosed during their early 40s and pre menopausal (Figure 4). Most patients with breast cancer have their left breasts affected in both upper and lower quadrant (Figure 2 and 3). During the diagnosis, the level of malignancy were majorly grade 2 abnormalities with tumor size of about 25-29 millimeters and reported extra lymph nodes of 0-6 (Figure 5, 7 and 6, respectively). More than 200 patients did not undergo radiation therapy treatment (Figure 8).

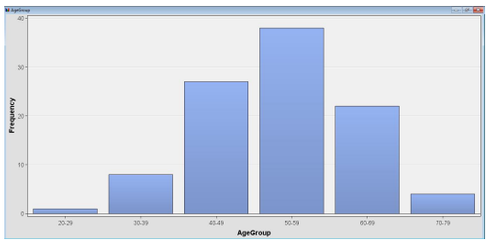


Figure 1. Frequency Distribution of Age

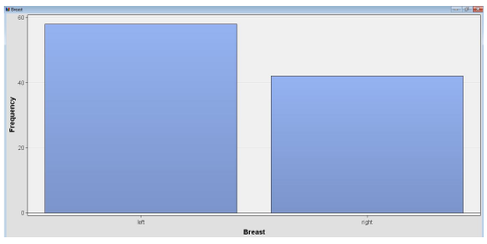


Figure 2. Frequency Distribution of affected Breast

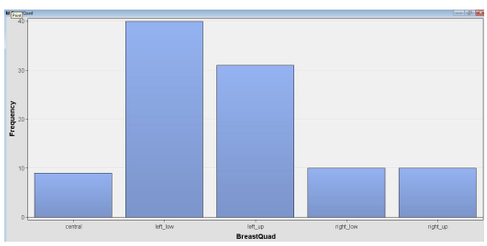


Figure 3. Frequency Distribution of affected Breast Quadrant

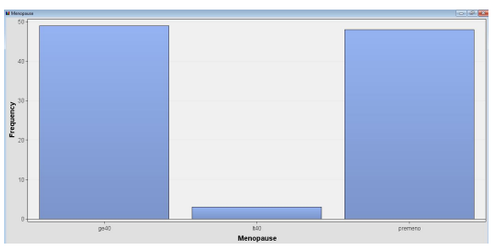


Figure 4. Frequency Distribution of Menopausal Stage

EUNICE CHUA
153775200
BAN210ZBB - FINAL PROJECT

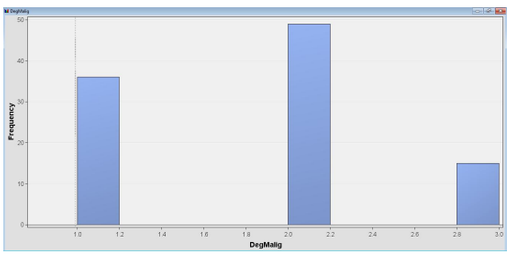


Figure 5. Frequency Distribution of Degree of Malignant during the diagnosing stage

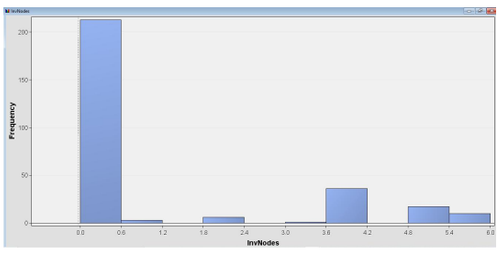


Figure 6. Frequency Distribution of InvNodes

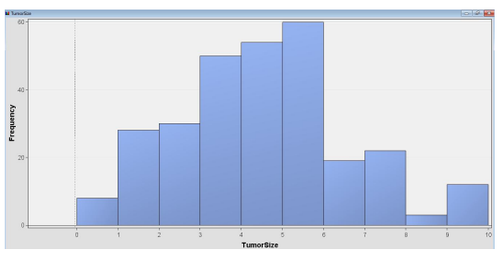


Figure 7. Frequency Distribution of Tumor Size

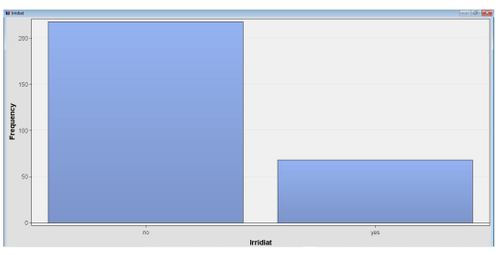


Figure 8. Frequency Distribution of Patients undergoing Radiation Therapy

Missing and Unique Values

Figures 9 and 10 present the missing and values of the dataset, respectively. There are no missing values in the data set. There are 2 unique values each for the following variables: class, breast, and irridiate. Also, there are 3 unique values for menopause, nodescap, and degree of malignant. The age group and breast quadrant have 6 unique values each. 7 unique values for invnodes and 11 unique values for the size of the tumor.

```
PatientID     0
Class         0
AgeGroup      0
Menopause     0
TumorSize     0
InvNodes      0
NodesCap      0
DegMalig      0
Breast        0
BreastQuad    0
Irridiat      0
dtype: int64
```

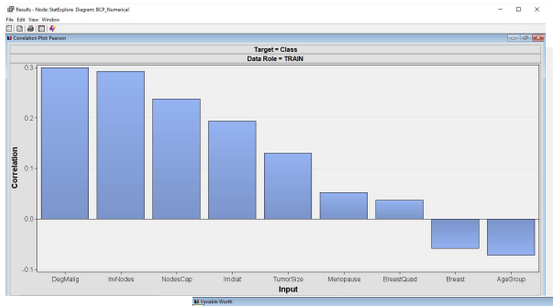
Figure 9. Missing Values

```
PatientID     286
Class           2
AgeGroup        6
Menopause       3
TumorSize      11
InvNodes        7
NodesCap        3
DegMalig        3
Breast          2
BreastQuad      6
Irridiat        2
dtype: int64
```

Figure 10. Unique Values

Correlation Analysis

The correlation of the variables with the class variable is shown in Figure 11, the generated correlation plot by the Stat Explore node. It shows that the there is a negative relationship between the breast affected and age group with the class. Nodescap and Irridiate have positive relationship with class. This is also supported with the p-values with less than 0.05. Thus, the variables that are highly influential to the class variables are degree of malignant, InvNodes, NodesCap, TumorSize, and Irridiat.



Input	Chi-Square	Df	Prob
DegMalig	31.6950	2	<.0001
InvNodes	28.7997	6	<.0001
NodesCap	22.6917	2	<.0001
TumorSize	17.9157	10	0.0564
Irridiat	10.7542	1	0.0010
BreastQuad	5.8711	5	0.3190
AgeGroup	3.9977	5	0.5497
Breast	0.9836	1	0.3213
Menopause	0.7915	2	0.6732

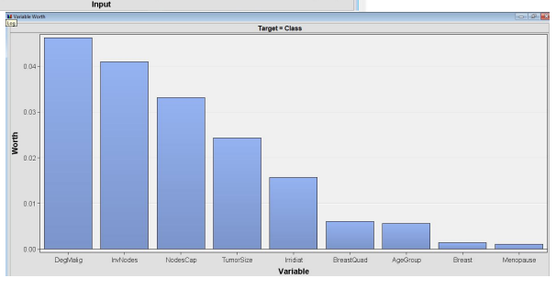


Figure 11. Correlation Analysis

Predictive Analytics Models

Figure 12 shows the process flow of the PA for the breast cancer dataset. From the importation of the file, data partition node was added to divide the dataset for training, validation, and test sets using 40:30:30 ratio. There are several models available for running categorical variables, the analysis will be focusing on decision trees and logistic regression. Annex B shows the properties assigned for each model. For the decision trees, the assessment measured assigned was average squared error. Since the dataset is composed of categorical variables, logistic regression is used for this analysis incorporating the stepwise method. Model

Comparison node was added in order to determine the optimal results for generalization among the models.

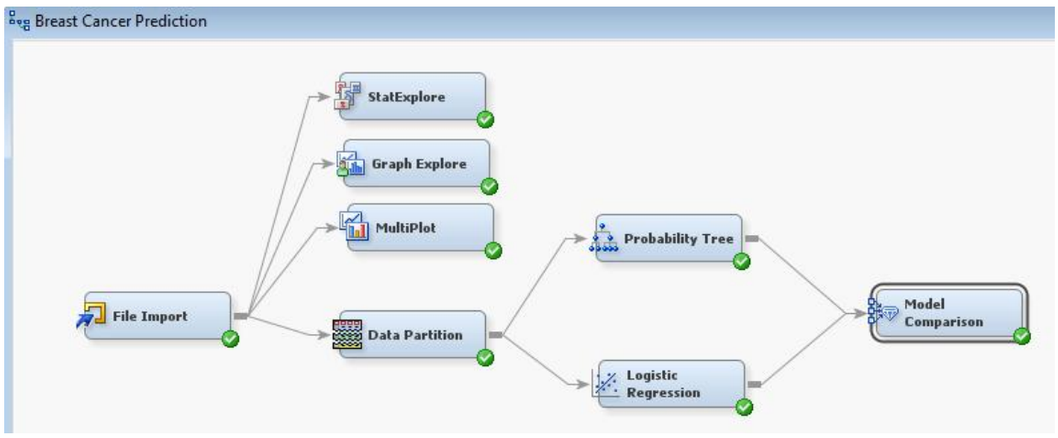


Figure 12. PA Process Flow

The Probability Tree generated misclassification rate for the training dataset of 0.265, validation set of 0.259 and test set of 0.318. The average square error was 0.182 for training, 0.178 for validation, and 0.199 in the test. The results also presented the classification of the train and validation sets through the confusion matrix. Seen in the Table 2 is the summary of the analysis for this matrix. The Precision and recall rates of the validation set has increased by 2% and 10%, respectively. This indicates that the False classification is minimized after the training of the model. Accuracy Rate is at 74% and F1 Score of 59%. Thus, the model is fairly accurate with minimal and acceptable error.

The Tree in Figure 13 also showed that the degree of malignant, grades 1 to 2 are predominantly have about 20% recurrence of the cancer, however, for grade 3 malignant more than 50% probability of cancer recurrence.

Fit Statistics

Target=Class Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	113.000	85.000	88.000
MISC	Misclassification Rate	0.265	0.259	0.318
MAX	Maximum Absolute Error	0.805	0.805	0.805
SSE	Sum of Squared Errors	41.111	30.279	35.109
ASE	Average Squared Error	0.182	0.178	0.199
RASE	Root Average Squared Error	0.427	0.422	0.447
DIV	Divisor for ASE	226.000	170.000	176.000
DFT	Total Degrees of Freedom	113.000	.	.

Event Classification Table

Data Role=TRAIN Target=Class Target Label=' '

False Negative	True Negative	False Positive	True Positive
16	66	14	17

Data Role=VALIDATE Target=Class Target Label=' '

False Negative	True Negative	False Positive	True Positive
10	47	12	16

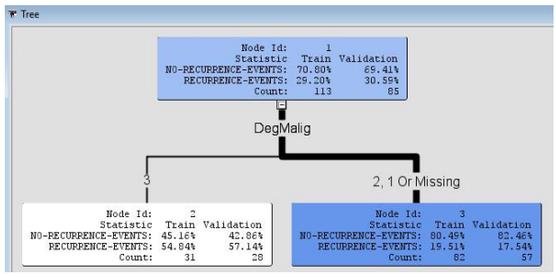


Figure 13. Probability Tree Results

Table 2. Analysis of Confusion Matrix for Probability Tree Model

	TRAIN	VALIDATION
PRECISION	55%	57%
RECALL	52%	62%
ACCURACY	73%	74%
F MEASURE	53%	59%

Figure 14 shows the results of the Logistic Regression model for the breast cancer dataset. The reported misclassification rates are 0.265 for train set, 0.259 for validation set, and 0.318 for the test set. Meanwhile, the average squared errors were 0.181 for train, 0.178 for validation, and 0.199 for the test set. Table 3 presents the calculated measurements for the confusion matrix analysis for the logistic regression model. Precision Rate increased from 55% to 57% for the train and validation tests. Recall rate has increased as well by 10% from 52% to 62%. This implies that the erroneous classification of negative recurrence was minimized by 62%. Accuracy rate showed an increase by 1% and F1 score by 6%.

The logistic regression also showed that the variable with significant correlation with the class variable is the degree of malignant as indicated on the p-value of less than 0.05.

Fit Statistics

Target=Class Target Label=' '

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	129.258	.	.
ASE	Average Squared Error	0.181	0.178	0.199
AVERR	Average Error Function	0.545	0.539	0.584
DFE	Degrees of Freedom for Error	110.000	.	.
DFM	Model Degrees of Freedom	3.000	.	.
DFT	Total Degrees of Freedom	113.000	.	.
DIV	Divisor for ASE	226.000	170.000	176.000
ERR	Error Function	123.258	91.551	102.860
FPE	Final Prediction Error	0.191	.	.
MAX	Maximum Absolute Error	0.839	0.839	0.839
MSE	Mean Square Error	0.186	0.178	0.199
NOBS	Sum of Frequencies	113.000	85.000	88.000
_NW_	Number of Estimate Weights	3.000	.	.
RASE	Root Average Sum of Squares	0.426	0.422	0.446
RFPE	Root Final Prediction Error	0.437	.	.
RMSE	Root Mean Squared Error	0.432	0.422	0.446
SBC	Schwarz's Bayesian Criterion	137.440	.	.
SSE	Sum of Squared Errors	40.997	30.334	34.973
SUMW	Sum of Case Weights Times Freq	226.000	170.000	176.000
MISC	Misclassification Rate	0.265	0.259	0.318

Event Classification Table

Data Role=TRAIN Target=Class Target Label=' '

False Negative	True Negative	False Positive	True Positive
16	66	14	17

Data Role=VALIDATE Target=Class Target Label=' '

False Negative	True Negative	False Positive	True Positive
10	47	12	16

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	DegMalig	2	1	13.8544		0.0010

Figure 14. Logistic Regression Results

Table 3. Analysis of Confusion Matrix for Logistic Regression Model

	TRAIN	**VALIDATION**
PRECISION	55%	57%
RECALL	52%	62%
ACCURACY	73%	74%
F MEASURE	53%	59%

The model comparison node showed that the best model for the prediction of breast cancer is the Probability Tree as shown in Figure 15. There is a minimal difference on the results between the two models. Both generated the same rates in confusion matrix analysis. The difference lied on the Average Squared Error for both the train and validation datasets by several decimal places such as ASE for the train set is 0.18191 for the probability tree model while 0.18140 for the logistic regression model, and for the ASE for the validation set 0.17811 in probability tree model and 0.17844 in logistic regression model. Therefore, considering the statistics of misclassifiation and average squared error, the best model to predict the breast cancer recurrence is the Probability Tree model.

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree	Probability Tree	0.25882	0.18191	0.26549	0.17811
	Reg	Logistic Regression	0.25882	0.18140	0.26549	0.17844

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree	Probability Tree	TRAIN	Class		16	66	14	17
Tree	Probability Tree	VALIDATE	Class		10	47	12	16
Reg	Logistic Regression	TRAIN	Class		16	66	14	17
Reg	Logistic Regression	VALIDATE	Class		10	47	12	16

Figure 15. Model Comparison Results

EUNICE CHUA
153775200
BAN210ZBB - FINAL PROJECT

ANNEX A

PROPERTIES OF THE MODELS

Logistic Regression

Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	

Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No
Correlation	No
Statistics	No
Suppress Output	No
Details	No
Design Matrix	No

Probability Tree

Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25

Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importa	
Observation Based Importa	No
Number Single Var Importar	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustm	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Interactive Sample	
Create Sample	Default
Sample Method	Random
Sample Size	10000
Sample Seed	12345
Performance	Disk
Score	
Variable Selection	Yes
Leaf Role	Segment
Report	
Precision	4
Tree Precision	4
Class Target Node Color	Percent Correctly Classified
Interval Target Node Color	Average
Node Text	

HP Forest

Tree Options	
Maximum Number of Trees	50
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each S	0.6
Number of Obs in Each Sam	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Con	.
Significance Level	0.05
Max Categories in Split Sea	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest Percentage of Obs	1.0E-5
Smallest Number of Obs in N	1
Split Size	.
Use as Modeling Node	Yes
Score	
Variable Selection	Yes
Variable Importance Metho	Loss Reduction
Number of Variables to Con	25
Cutoff Fraction	0.01

I, EUNICE M. CHUA, declare that the attached assignment is my own work in accordance with the Seneca Academic Policy. I have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. I have not distributed my work to other students