Predicting The Recurrence of Breast Cancer Using Predictive Models

Tareq Haboukh

Predictive Analytics
Prof. Uzair Ahmad

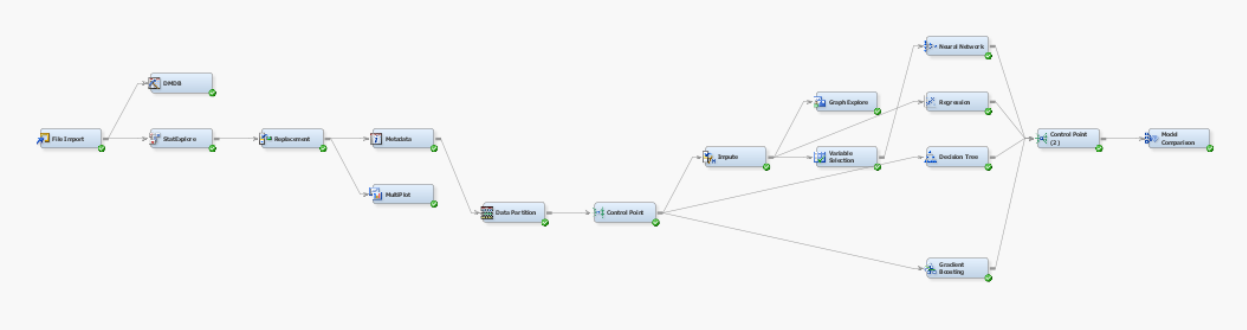
April 17, 2022.

In this report, I'll be working on the Breast Cancer Dataset from the UCI Machine Learning Repository and applying predictive analytics on the dataset to be able to predict whether Cancer is one of the recursive types or not.

This Analysis and the model trained are aimed to help doctors determine if a cancer is recursive with accuracy up to 75.8% and 66.7% precision.

In the analysis, four different models were used (Decision Trees, Gradient Boosting, Regression, Neural Network) and their results were compared using the cross-validation method in the Model Comparison Node. **Gradient Boosting** had the best results and can identify if the cancer is a recurrence event with a 22.9% Misclassification Rate and an Average Squared Error of 9% it performed better on the validation dataset in comparison to the other models.

The analysis was done using SAS Enterprise Miner, the following is the diagram used and it shows the whole process from first, importing the dataset, exploring the attributes, checking for anomalies, splitting the data, imputing missing data, training the predictive models, and finally comparing between them.



The dataset has 10 attributes and 285 records of breast cancer patients, Class attribute was assigned the Target role, and the rest are inputs.

Attributes Information:

Attribute	Type	Values
Class	Binary (T)	no-recurrence-events, recurrence-events.
Age	Ordinal	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
Menopause	Nominal	lt40, ge40, premeno.
Tumor-Size	Ordinal	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
Inv-Nodes	Ordinal	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.
Node-Caps	Binary	yes, no.
Deg-Malig	Nominal	1, 2, 3.
Breast	Nominal	left, right.
Breast-Quad	Nominal	left-up, left-low, right-up, right-low, central.
Irradiat	Binary	yes, no.

Importing and exploring The Dataset to SAS

```
In [1]: # Import Breast Cancer Dataset, add column names, and export txt file.
        import pandas as pd

        BreastCancer = pd.read_csv('breast-cancer.data')
        BreastCancer.columns = ['Class','Age','Menopause','Tumor_Size','Inv_Nodes','Node_Caps','Deg_Malig','Breast','Breast_Quad','Irradiat']

        df = pd.DataFrame(BreastCancer)
        df.to_csv('BreastCancer.txt', index=False)
```

In order to get the "breast-cancer.data" ready for the analysis, the file was loaded, and column names were added using the code illustrated before, then the data was exported as a text file, this step was done using Pandas. then the new file was imported to SAS enterprise miner using the File Import node (Attribute types changed as per the attribute information).

Class Variable Summary Statistics

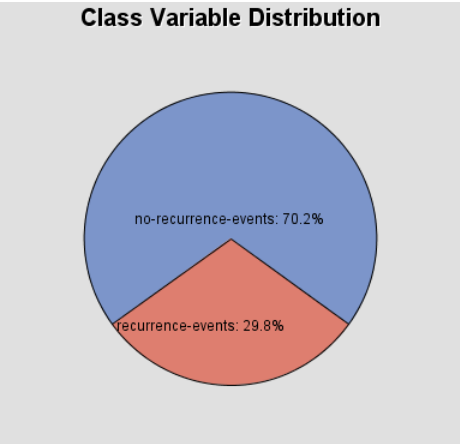
Variable	Label	Type	Number of Levels	Missing
Age		C	6	0
Breast		C	2	0
Breast_Quad		C	6	0
Class		C	2	0
Deg_Malig		N	3	0
Inv_Nodes		C	7	0
Irradiat		C	2	0
Menopause		C	3	0
Node_Caps		C	3	0
Tumor_Size		C	11	0

To check if there were any missing data the DMDB node was used, and it showed no missing values, all attributes are the character type except Deg_Malig, and the Number of suggests further investigation, Node_Caps is supposed to be binary, but the summary statistics show 3 levels.

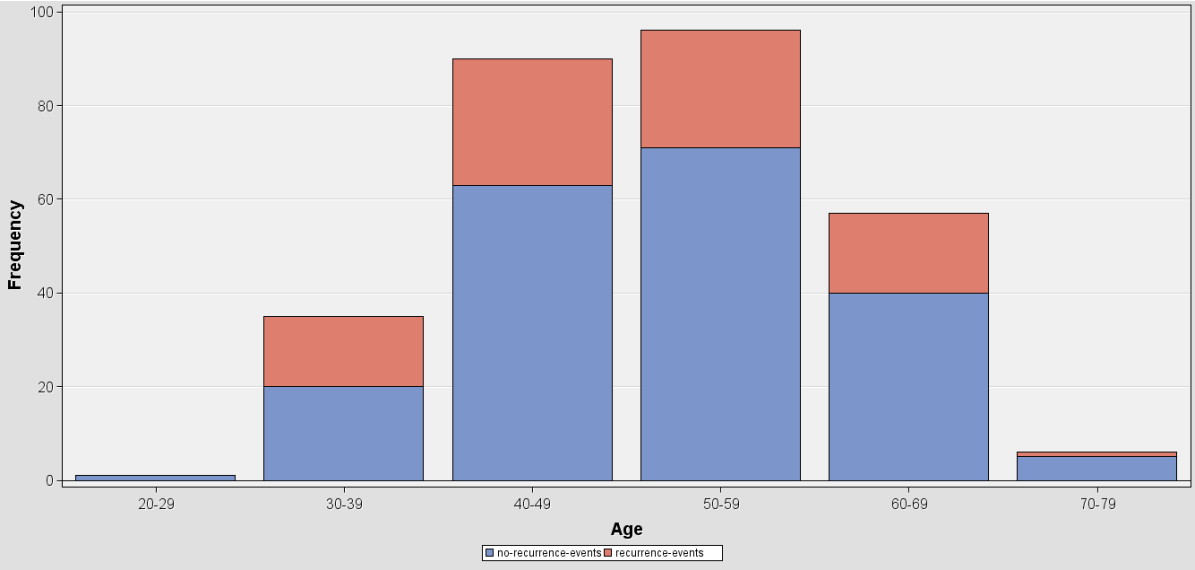
Breast_Quad	left_low
Breast_Quad	left_up
Breast_Quad	right_up
Breast_Quad	right_low
Breast_Quad	central
Breast_Quad	?
Node_Caps	no
Node_Caps	yes
Node_Caps	?

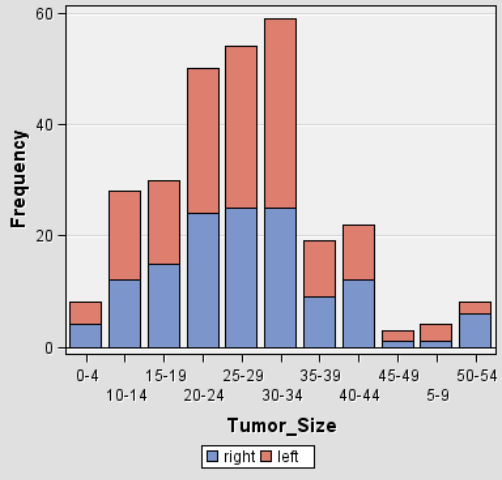
To be sure the node StatExplore can be used to list the levels of each attribute. It shows that both Breast_Quad and Node_Caps has a "?" level which indicates missing data, this issue can be fixed by using the Replacement node to replace "?" values with blank values so SAS miner can treat them as missing value and not an extra level. The missing values will be imputed later on.

Moving on, the attribute Class has 200 records as no-recurrence-events and 85 records as recurrence-events. This difference in proportions might become an issue while training the predictive models.



Also, it is interesting to note that the class variable is represented evenly across most of the age groups, roughly 30% of records were recurrence-events for each age group except the 20-29 group as it had only one record. Meaning that the 70-30 Class distribution is also seen in the Age attribute as shown below, and it follows a normal distribution.

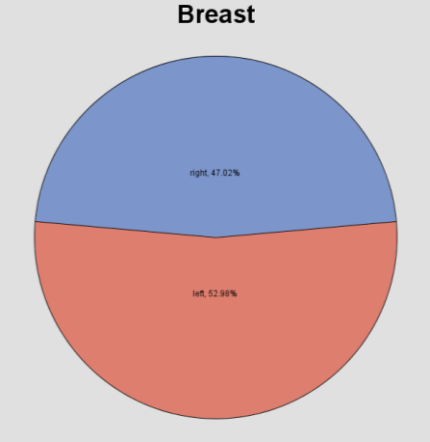


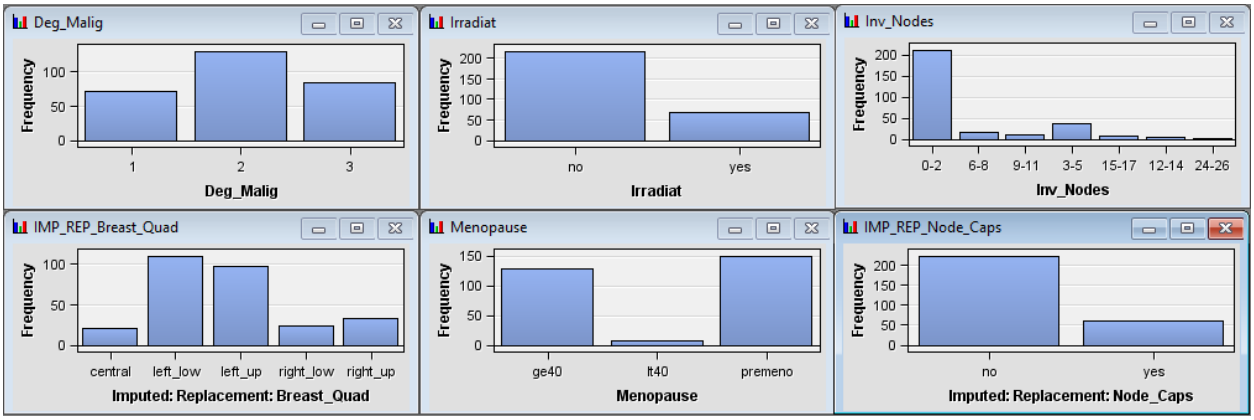


As for Tumor_size 57% of the cases were between bin 20-24 and bin 30-34 with 163 records. the Tumor _size attribute followed a normal distribution just like the age attribute.

*5-9 bin is in the wrong order which gives the impression of a right-skewed distribution.

151 data points for the Breast attribute are left at 52% and the rest 134 are right 48%, it seems like it is more likely for Breast to be left in the mentioned Tumor_Size interval from size 20-34.



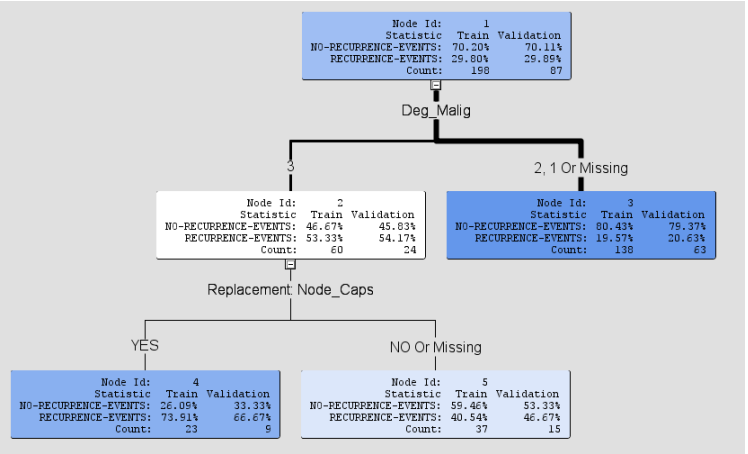


The rest of the attributes do not show any outliers or anomalies, but it is worth mentioning that in Inv_Nodes most of the cases were in the 0-2 bin, and left_low, left up make up 72% of the data in the Breast_Quad attribute.

The Predictive Models

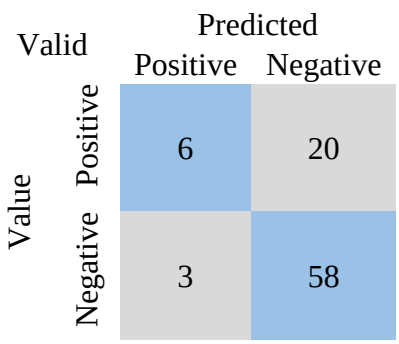
To begin with the analysis the dataset needs to be split into two parts, 70% for the training dataset (200 Records) and 30% for the Validation dataset (85 Records) in order to train and validate the predictive models.

The **Decision Tree** node resulted in a 2-level tree, first Split with the attribute Deg_Malig into two leaves, "1,2 or missing" leaf and "3" leaf. The second split was using the attribute Node_Caps and it generated two leaves "Yes","No", or "Missing".



The Decision Tree model managed to classify the Class of the validation dataset with 18% Average Squared Error (ASE) and 26% Misclassification rate (MISC).

Fit Statistics ▲	Statistics Label	Train	Validation
ASE	Average Squared Error	0.177127	0.185765
DFT	Total Degrees of Freedom	198	.
DIV	Divisor for ASE	396	174
MAX	Maximum Absolute Error	0.804348	0.804348
MISC	Misclassification Rate	0.242424	0.264368
NOBS	Sum of Frequencies	198	87
RASE	Root Average Squared Error	0.420864	0.431005
SSE	Sum of Squared Errors	70.14219	32.32311

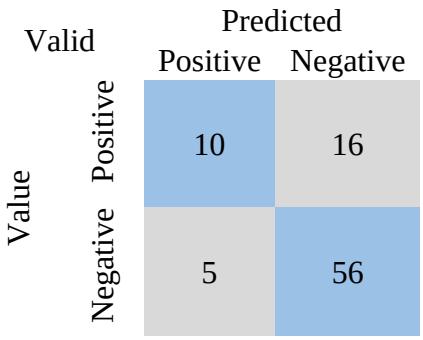


The second model to be trained is the **Gradient Boosting** model. When the node is run the model assigned The Tumor_size attribute as the most important factor followed by Inv_Nodes and Age.

Variable Name	Importance	Validation Importance
Tumor_Size	1	0.852261
Inv_Nodes	0.858694	0.920246
Age	0.835385	0.933171
REP_Node_Caps	0.775261	0.819998
Deg_Malig	0.732332	1
REP_Breast_Quad	0.647945	0.534094
Irradiat	0.543048	0.803626
Menopause	0.519976	0.609469
Breast	0.504107	0.306866

In terms of the performance of the model, the ASE is at 19% and MISC at 22.9% for the validation dataset which is better than the previous Decision Tree due to the MISC rate.

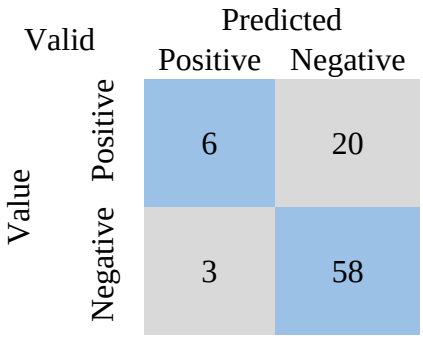
Fit Statistics ▲	Statistics Label	Train	Validation
ASE	Average Squared Error	0.092355	0.197025
DFT	Total Degrees of Freedom	198	.
DIV	Divisor for ASE	396	174
MAX	Maximum Absolute Error	0.856289	0.908853
MISC	Misclassification Rate	0.09596	0.229885
NOBS	Sum of Frequencies	198	87
RASE	Root Average Squared Error	0.3039	0.443875
SSE	Sum of Squared Errors	36.57261	34.28229
SUMW	Sum of Case Weights Times Freq	396	174



Before running the Regression and the Neural Network nodes, first, the missing values needed to be imputed using the Surrogate Tree option in the Impute node, this step is done now because the previous models are forgiving when it comes to missing values.

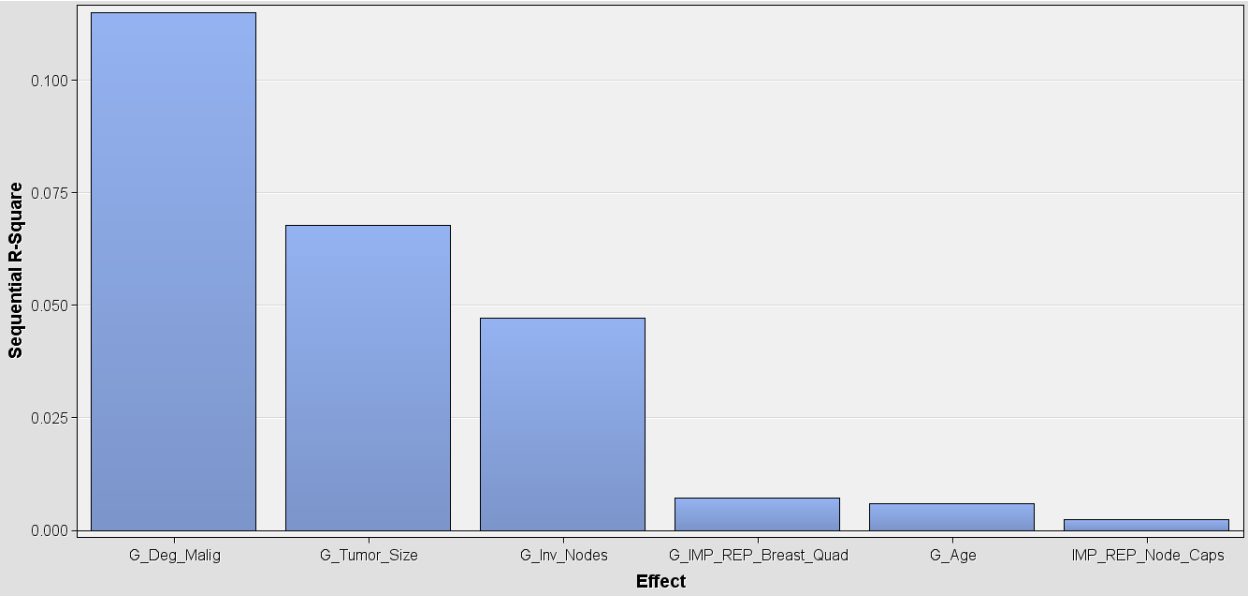
The **Regression** Model had 18% ASE and 26% MISC for the Validation dataset. Which is similar to the Decision Tree's performance.

Fit Statistics ▲	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	219.6158	.
ASE	Average Squared Error	0.176302	0.184275
AVERR	Average Error Function	0.534383	0.552723
DFE	Degrees of Freedom for Error	194	.
DFM	Model Degrees of Freedom	4	.
DFT	Total Degrees of Freedom	198	.
DIV	Divisor for ASE	396	174
ERR	Error Function	211.6158	96.17378
FPE	Final Prediction Error	0.183572	.
MAX	Maximum Absolute Error	0.841443	0.841443
MISC	Misclassification Rate	0.242424	0.264368
MSE	Mean Square Error	0.179937	0.184275
NOBS	Sum of Frequencies	198	87
_NW_	Number of Estimate Weights	4	.
RASE	Root Average Sum of Squares	0.419883	0.429273
RFPE	Root Final Prediction Error	0.428453	.
RMSE	Root Mean Squared Error	0.424189	0.429273
SBC	Schwarz's Bayesian Criterion	232.7689	.
SSE	Sum of Squared Errors	69.81541	32.06387
SUMW	Sum of Case Weights Times Freq	396	174



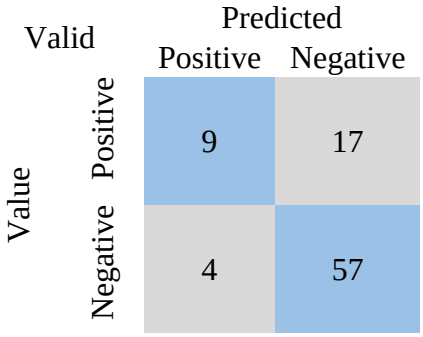
Before training the **Neural Network** node, first, we need to select the most influential attributes using the Variable Selection Node. Where it picked the following attributes based on the R-Square value, the attributes are in order of significance.

Deg_Malig, Tumor_Size, Inv_Nodes, Breast_Quad, and Node_Caps.



After attribute reduction is done, the Neural Network node resulted in 19% ASE and 24% MISC for the Validation Dataset.

Fit Statistics ▲	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	412.1488	.
ASE	Average Squared Error	0.144445	0.19651
AVERR	Average Error Function	0.43977	0.634516
DFE	Degrees of Freedom for Error	79	.
DFM	Model Degrees of Freedom	119	.
DFT	Total Degrees of Freedom	198	.
DIV	Divisor for ASE	396	174
ERR	Error Function	174.1488	110.4058
FPE	Final Prediction Error	0.579607	.
MAX	Maximum Absolute Error	0.936528	0.995576
MISC	Misclassification Rate	0.20202	0.241379
MSE	Mean Squared Error	0.362026	0.19651
NOBS	Sum of Frequencies	198	87
_NW_	Number of Estimated Weights	119	.
RASE	Root Average Squared Error	0.380059	0.443294
RFPE	Root Final Prediction Error	0.761319	.
RMSE	Root Mean Squared Error	0.601686	0.443294
SBC	Schwarz's Bayesian Criterion	803.4526	.
SSE	Sum of Squared Errors	57.20006	34.19274
SUMW	Sum of Case Weights Times Freq	396	174
WRONG	Number of Wrong Classifications	40	21



In order to compare the four predictive models in cross-validation, the Node Model Comparison is used. And it picked **Gradient Boosting** as the best model. The selection table is set to validation and the selection statistic is Average Profit/Loss.

It performed better on the Training and Validation dataset with 9.2% and 19.7% respectively for ASE and 9.5% and 22.9% for MISC.

Selected Model	Model Description	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Gradient Boosting	0.092355	0.09596	0.197025	0.229885
	Neural Network	0.144445	0.20202	0.19651	0.241379
	Regression	0.176302	0.242424	0.184275	0.264368
	Decision Tree	0.177127	0.242424	0.185765	0.264368

By looking at the confusion matrix below, Gradient Boosting had the highest True Negatives and True Positives 138, 41 and lowest False Negatives and False Positives 18, 1 for the training dataset.

And for the validation dataset had the highest True Positive 10 and Lowest False Negative 16 making it the best model for this specific case.

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Boost	Gradient Boosting	TRAIN	Class		18	138	1	41
Boost	Gradient Boosting	VALIDATE	Class		16	56	5	10
Tree	Decision Tree	TRAIN	Class		42	133	6	17
Tree	Decision Tree	VALIDATE	Class		20	58	3	6
Reg	Regression	TRAIN	Class		42	133	6	17
Reg	Regression	VALIDATE	Class		20	58	3	6
Neural	Neural Network	TRAIN	Class		32	131	8	27
Neural	Neural Network	VALIDATE	Class		17	57	4	9

With this, the analysis is concluded.

I, Tareq Haboukh, 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.