Final Project: Breast Cancer Decision Tree

Arif Ahmed, Abhishek Singh, and Kory Duong

ADTA5120

Dr. Anna George

12/10/2019

# Breast Cancer Decision Tree

## Introduction

Diagnosing a patient with breast cancer requires multiple steps. As given by MayoClinic[1], the steps are as follows: breast exam, mammogram, breast ultrasound, removing a sample of breast cells for testing (biopsy) and breast magnetic resonance imaging (MRI). Understanding the result from all of these sources to analyze is very time-consuming. With human errors, these predictions might not be very accurate. Therefore, our team came up with a novel way to detect breast cancer using machine learning. Utilizing the Wisconsin Breast Cancer and IBM SPSS, we created a decision tree module that can assist researchers in automating the process of determining which tissues are malignant and benign. We use the decision tree as a classifier to determine what tissue is malignant or benign. Before we jump into the decision tree, we must first characterize what is considered to be benign or malignant.

## Dataset

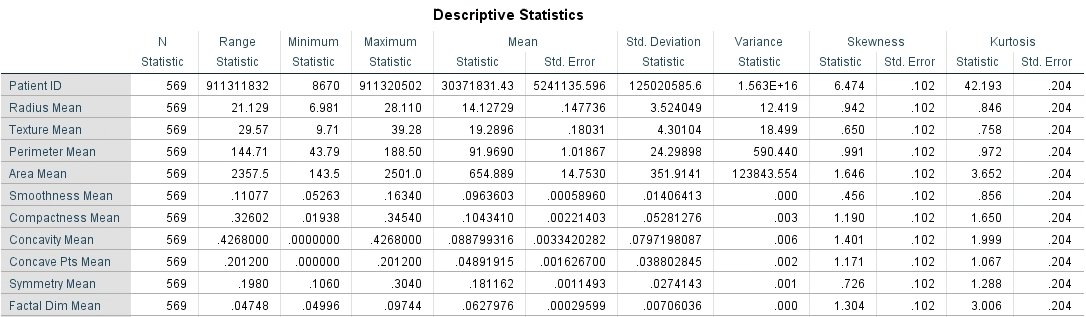
The data consists of 569 cases. The different features that we use for this project are defined as radius, texture, perimeter, area, smoothness, compactness, concavity concave points, symmetry, and fractal dimension. In Figure 1, Figure 2 and Figure 3, we illustrate the descriptive statistics of each variable and their standing in each category (ie. Mean, Standard Deviation [SD] and Worse). 

Figure 1. Descriptive statistics of the Features

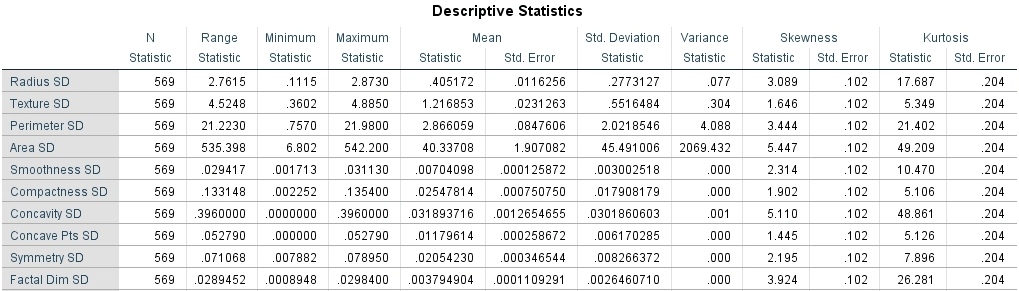


Figure 2. Descriptive statistics of the Features

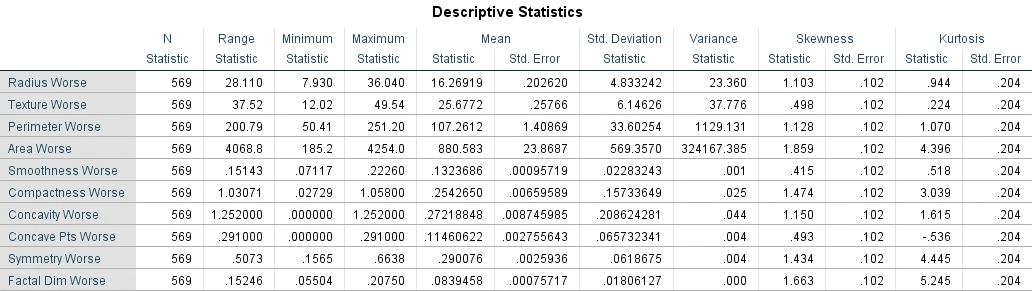


Figure 3. Descriptive statistics of the Features

The above descriptive statistics gives us a very high-level spread of our data and allows us to see where any anomalies might be in the data by measurements such as standard deviation.

#### Anomalous Data

While looking for anomalies in the data, we notice that there were a few zeroes. They were counted as significant because they were very close to the other values. Looking at the dataset one observation we have is that one of the rows has missing data. We decided to delete the missing value row in the data which has only one column as shown in Figure 4.

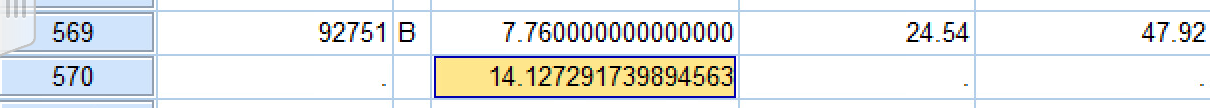


Figure 4. Showing the row with missing data.

Looking at the feature set, another observation is that the use of PatientID is not very helpful. Using PatientID would result in two problems. 1. The model might be biased based on a particular patient. Hence the training accuracy would be very high but the test accuracy will be bad as we would not see the same patient again in the test set. 2. The model would try to use PatientID as a feature and would not be able to find the best feature.

Ideally, we would like to have a very large dataset to have a conclusive experiment. Something in the range of 100,000 data points can give us more confidence in predicting malignant or benign tissue.

## Model Criteria

In Machine Learning on a very high level, there are two types of models, supervised and unsupervised. In a supervised model, we need training data to learn parameters. A decision tree is a supervised learning method. As pointed out in [2] “Decision tree algorithms are primarily composed of training data, test data, a heuristic evaluation function, and a stopping criterion function.” The decision tree algorithm at every step is trying to find the information gain of each variable. The variable which has the maximum information gain is used to find the node of the tree. The various decision tree algorithms that we have tried in this paper are: QUEST, CHAID, and CRT.

Decision tree used to predict the type of cancer should be very accurate. We should avoid misclassification and the model should minimize false negatives. Based on that we have the following criteria:

* Estimated risk cap should be less than 0.05
* The combined accuracy to predict should be at least more than 93%
* The individual accuracy should not be less than 90%
* The difference in test and train accuracy should not be more than 10%
  + This will help us understand if the model is over-fitting on the training set
* We should find a model to minimize false negatives.

## Model Methodology

#### QUEST

As given in IBM SPSS definition in [5], QUEST—or Quick, Unbiased, Efficient Statistical Tree—is a binary classification method for building decision trees. The input variables in QUEST can be continuous too, but the output class should be categorical.

As in [5], **Strengths**. Like CHAID, but unlike C&R Tree, QUEST uses statistical tests to decide whether or not an input field is used. It also separates the issues of input selection and splitting, applying different criteria to each.

#### CRT

CRT or Classification and Regression (C&R) Tree node, starts to slit the tree based on the input field that reduces the impurity index from the split to the maximum. All the splits are binary and the CRT gives the functionality of pruning.

#### CHAID

CHAID or Chi-squared Automatic Interaction Detection is a decision tree algorithm that uses chi-square statistics to find the best split. As in [5], CHAID will select the input field that is the most significant (smallest p-value).

#### Validation

One of the most important parts of any model building is to find the best validation method. IBM SPSS provides 2 generic methods to do the validation. One is to do a N-fold validation and the other is to divide the data into training and testing sets. We use the method in which we divide the data into training and test sample. Based on various experimentations we found that our optimal split was 2:1, where we took around 70-75% of the data for training and 25-30% of the data for testing.

For the cases where the train:test ratio was 2:1 or 3:1, having a smaller training percentage was making the model not learn best on the small training set. Using the above split, we were able to get the accuracy we were targeting along with being in the accuracy threshold range.

## Results

We evaluate all the three candidate models against the model criterion we started with. All three methods sufficed the basic model criterion which we started, namely:

* Estimated risk cap should be less than 0.05
* The combined accuracy to predict should be at least more than 93%
* The individual accuracy should not be less than 90%
* The difference in test and train accuracy should not be more than 10%
  + This will help us understand if the model is overfitting on the training set
* We should find a model to minimize false negatives.

We will explain each model in more detail and then comments on our findings.

**CRT MODEL**

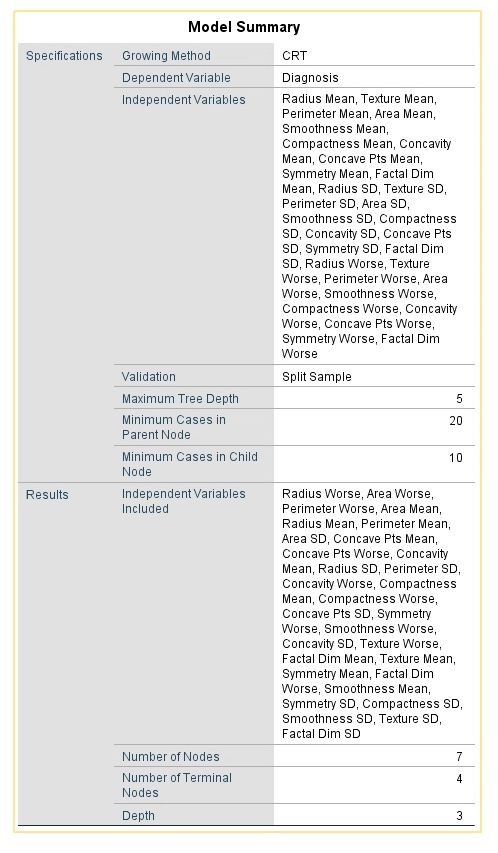
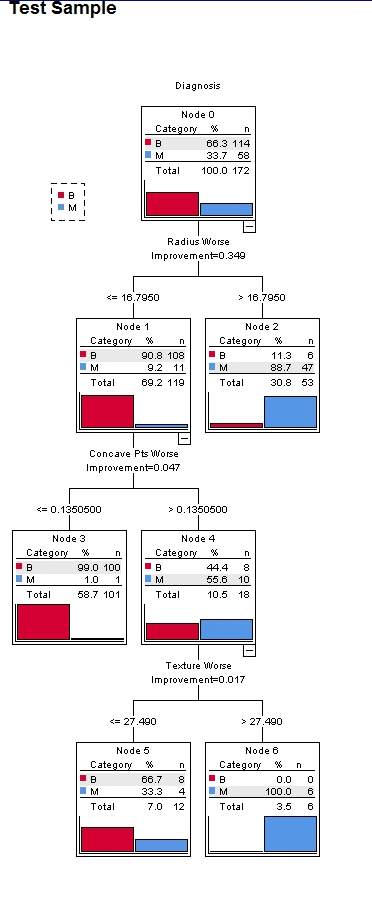
****

Figure 5. CRT Decision tree and Model Summary

#### 

#### Risk and Classification of CRT

#### 

#### 

Figure 6. Risk and Classification of CRT

We can see from the CRT model that the most pertinent variables are Radius Worse, Concave points Worse, Texture worse. These variables are most significant according to our CRT decision tree model.

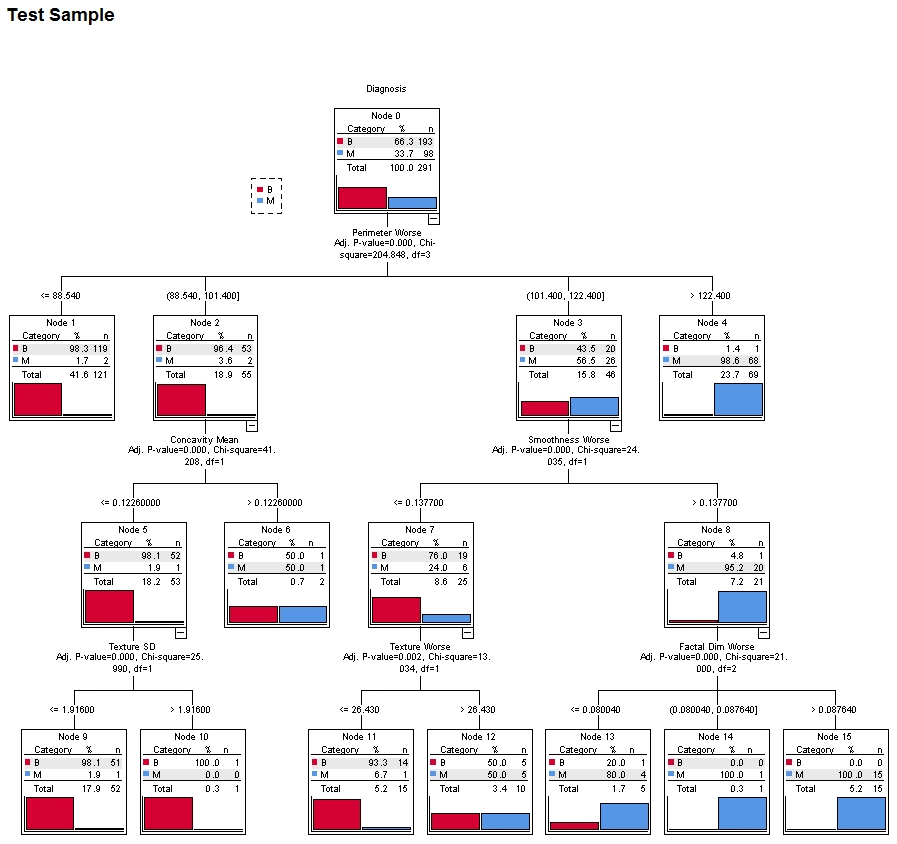
When we look at the tree, we can see that it begins with the splitting of radius worse and has a depth of 3 and has 4 terminal nodes. We select pruning to avoid overfitting which gives us more accuracy [[3](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/)]. As we know CRT decision tree does binary split only our CRT tree is narrower than the CHAID and Quest tree [[4](https://doi.org/10.1155/2014/929768.)].

The parameter that we choose while training the CRT Model are:

* The train vs test dataset was split in the 70/30 ratio.
* Maximum tree depth 3
* Minimum cases in parent node 20
* Minimum cases in child node 10
* Allow pruning
* Remove the patientID from Independent variables.

After analyzing the tree, we can see that out of 4 terminal nodes, 3 of them has fully deterministic capabilities. The CRT tree first starts with the variable Radius Worse and for Radius Worse >16.7950 it has 96.4 % predictability of being malignant. The tree further splits node which has a Radius worse <=16.7950 using Concave pts worse variable and gain 99% confidence for Benign for Concave pts worse <= 0.1350500. And as it cannot determine the value Concave pts worse > 0.1350500, it again splits using Texture worse variable and got an accuracy rating of 100% for the malignant type for Texture worse >27.490. Although out of terminal nodes, 2 of them with the largest data sample (Radius worse>16.7950 and Concave pts worse<=0.1350500) provides with great predictability, it might contain false positive or false negative. So that we might need to look into CHAID and QUEST model for further analysis.

### CHAID Model



**Model Summary**

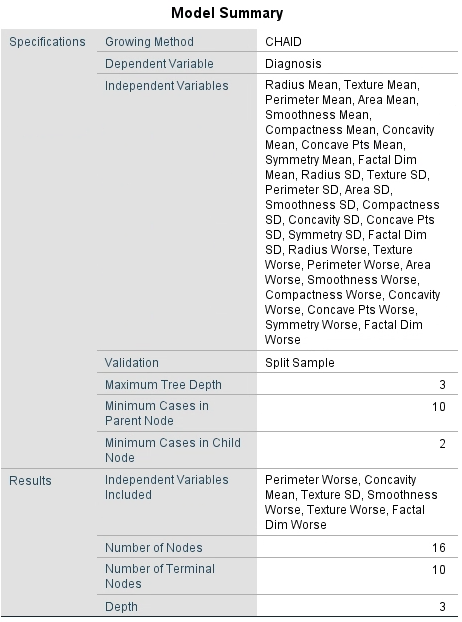
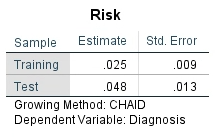


Figure 7. CHAID Decision Tree and Model Summary

#### Risk and Classification of CHAID



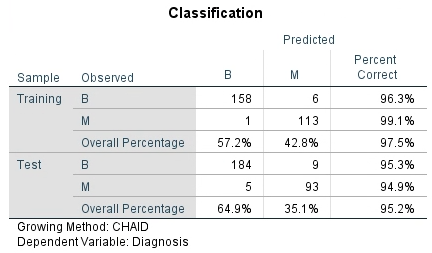


Figure 8. Risk and Classifications of CHAID decision tree

In the CHAID Analysis, the most relevant variables are the following: Perimeter Worse, Concavity Mean, Texture Standard Deviation, Smoothness Worse, Texture Worse, and Fractal Dimension Worse. These variables were the most significant thus, included in the decision tree module.

Moving onto the tree itself, we can see that it has a depth of 3, 10 terminal nodes, and 16 total nodes. This decision tree is comparatively wider and goes deeper than decision trees created using the CRT and QUEST method due to CHAID’s non-binary nature, meaning that more than 2 nodes/terminal nodes can spawn from a terminal node; we can have more information as to how and why it categorizes the tissue as malign or benign based on the significant variables. Now that we have established an understanding regarding the structure of the tree and its nodes, we can begin discussing more detailed features that it has and its effectiveness compared to QUEST and CRT methodologies.

Beginning with the first terminal node, SPSS utilized Perimeter worse as its 1st determining variable. In 1st split, we can see that tissue with a Perimeter Worse of less than 88.540 has a 98.3% chance of being benign. On the opposite end of that spectrum, we can see that tissue with a Perimeter Worse greater than 122.400, it is almost guaranteed to be malignant (98.6% chance). However, there is a little grey area in this split with the terminal nodes in Node 2 and Node 3.

In node 2, if the Perimeter Worse is between 88.540 and 101.4, then there is a 95.4% chance that it will be benign and a 3.6% chance that it will be malignant. SPSS decided to explore this node further with the Concavity Mean Variable. In the second split, we can see that tissue with a concavity mean of less than 0.1225, it has a 98.1% chance of being benign. While on the other hand, tissue with a concavity mean greater than the aforementioned value, has a .5 chance of being either. This is one of the limitations of this model and could be addressed by setting up stricter rules and developing a significantly more detailed parent and child node split.

In node 3, we can see that there is almost an even split between benign and malignant tissue (43.5% and 55.5%, respectively). SPSS chose to use the Smoothness Worse variable to analyze this further. In tissue with a smoothness worse of less than 0.137, there is a significant chance that it will be benign, while if it is greater, it has a much higher chance of being malignant (98.6%). To address the close to even split nodes, SPSS decided to use Texture worse to investigate it further. It was discovered that tissue with a texture worse value of less than 26.43, is almost guaranteed to be benign, while those with greater value, is at a perfectly even split. This is another limitation of this model and can be addressed by the same methods as in the 2nd split.

### QUEST Model

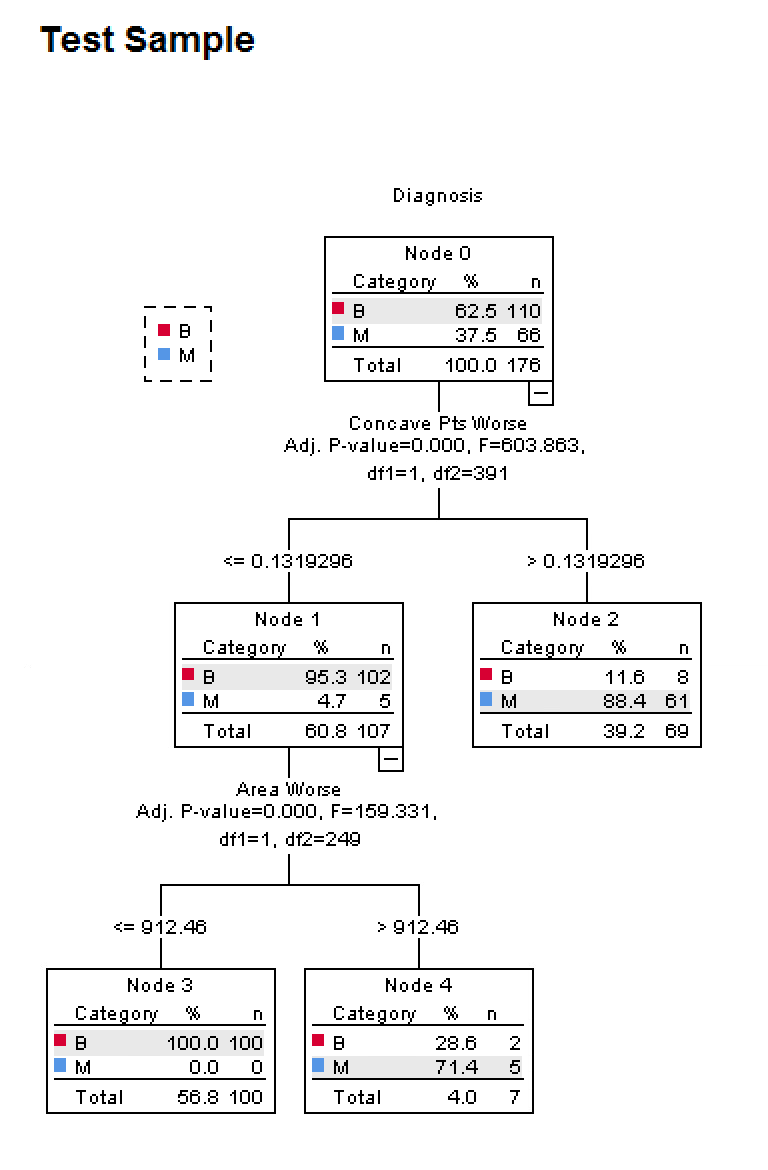


Figure9: Quest decision Tree Model

The QUEST model as shown in figure above is a very simple model to understand. It uses only 2 variables to build the whole model. The Root node is Concave Points Worse and the second layer node is Area Worse. Both of these attributes have a non-integer variable. The tree has a depth of 2 with only 3 terminal nodes. Based on the model above we can see that 2 of the 3 nodes are very deterministic in the predictive capabilities. The node where Concave Points Worse is >0.1319296, can predict M for 61 test sample with an accuracy of 88.4%. In contrast the 2nd level node with Concave Pts Wors <=0.1319296 AND Area Worse <=912.46 has the predictive capability of 100% for 100 Benign test set data.

#### Model Summary

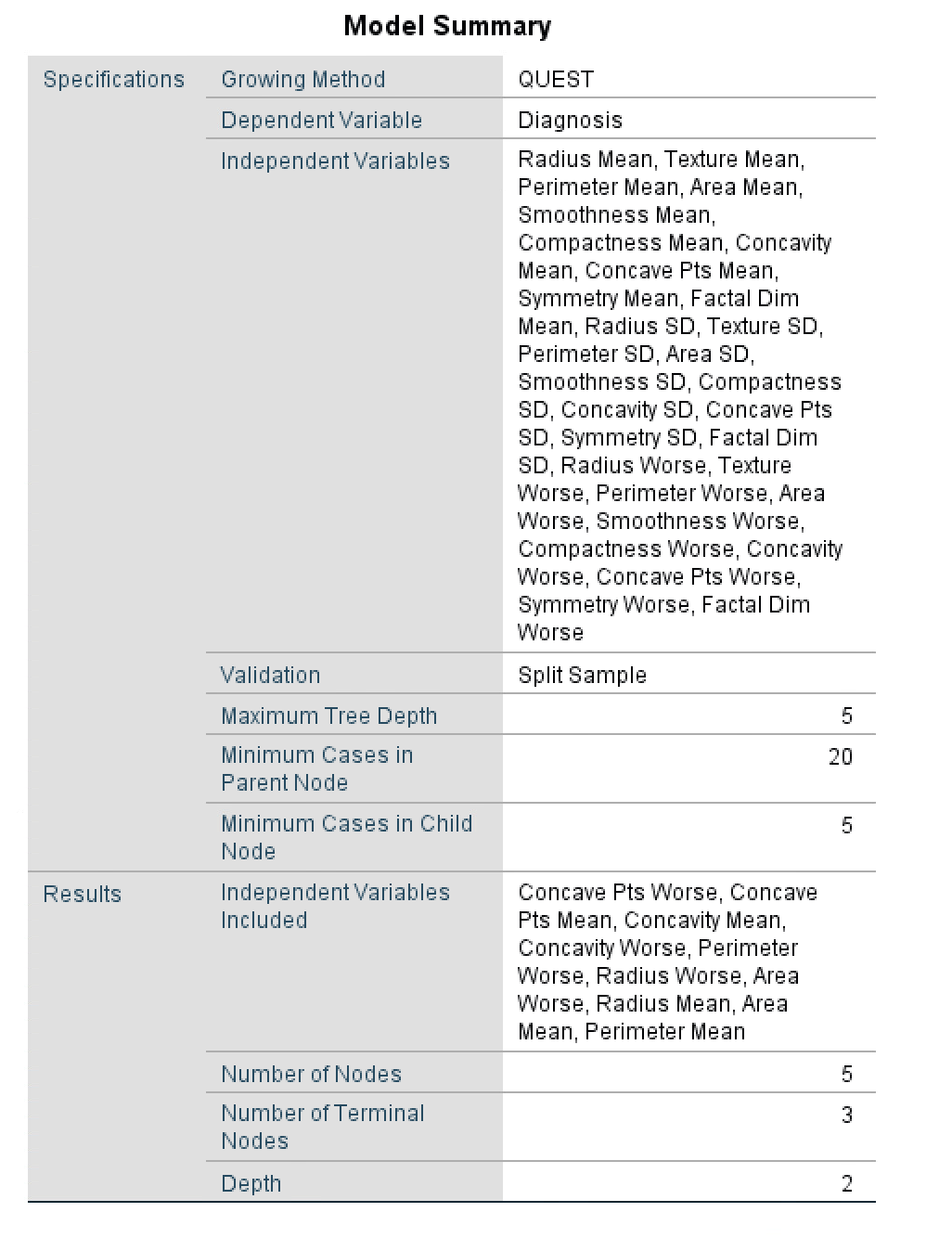
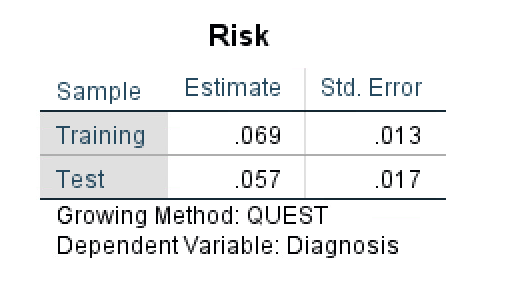


Figure10: Model Summary of QUEST decision Tree

#### 

#### Risk And Classification of QUEST



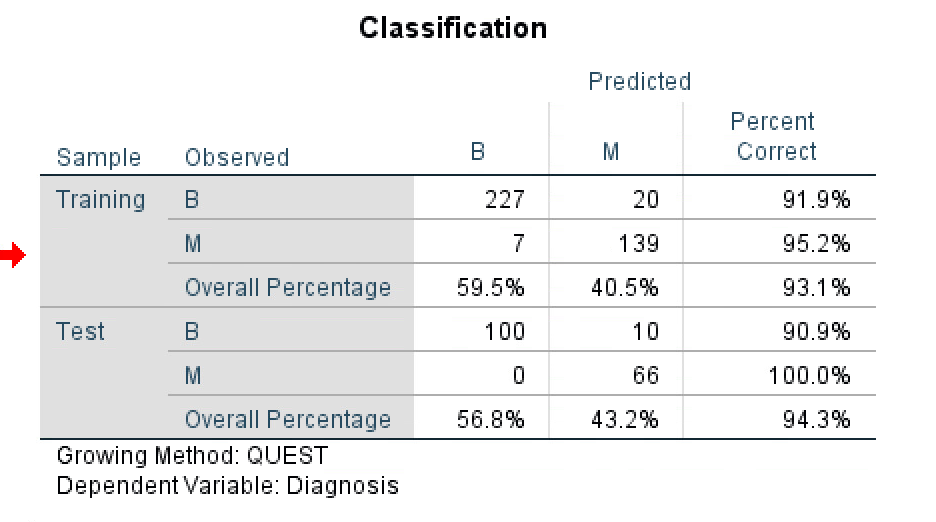


Figure11: Risk and Classifications of QUEST Decision Tree

The parameter that we choose while training the QUEST Model are:-

* The train vs test dataset was split in the 70/30 ratio.
* Maximum tree depth 5
* Minimum cases in parent node 20
* Minimum cases in child node 5
* Allow pruning
* Remove the patientID from Independent variables.

If we look at the result the following observation is very important. The overall training and test accuracies are 93.1% and 94.3% respectively. This gives us the confidence that the model is not overfitting on the training data. Another important observation is that the accuracy to diagnose benign and malignant is very close in both the test and training set. The risk estimate is as low as 0.057 in the test set with std. Error being only 0.017. This also gives us confidence in the robustness of the model.

## Variable Importance

Referring to our best model which is the CHAID model, the most relevant variables are the following: Perimeter Worse, Concavity Mean, Texture Standard Deviation, Smoothness Worse, Texture Worse, and Fractal Dimension Worse. The root node is the Perimeter Worse, This is not integer value. Hence the CHAID model divides this variable into 4 buckets, <=88.54, (88.54 to 101.4), (101.4 to 122.4) and >122.400. Based on the model we can observe that if the value of Perimeter Worse is less than 101.4 the model generally predicts benign. On the lower level, the other variable like the Concavity Mean feature is divided into 2 parts, <=0.122600 and >0.122600. When the value is <0.122600, the model tends to predict benign.

## Our Decision

Based on our findings using each methodology, we concluded that using the CHAID decision tree will be the best predictor of benign and malignant tissue. Compared to the QUEST and CRT methods, all three models had relatively low risk, but CHAID had to highest percentage correct. However, it is not to say that the QUEST and CRT models are unviable. They both had very high percentage correct scores and all had approximately the same low-risk number. For instance, in the QUEST model, we had 100% accuracy in detecting malignant cells and the CRT model would be a viable substitute for the CHAID model because the overall percentage correct is extremely close. We believe that CHAID had the highest rating because since there were many variables, CHAID was able to discern the most significant variables and gives us a slightly more accurate decision tree.

## Predictive Capability

In order to discuss predictive capability, we must take the risk number of our models into consideration. Although high percentage correct scores can be a good indicator, the risk number tells us the stability of our scores. This plays a significant role in the predictive capability of our model and also affects the reliability of it.

Taking our models into consideration we will be primarily looking at the values from the test tree, the risk number per model goes as follows:

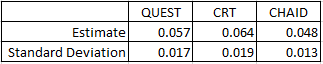


Figure12: The risk number per Model

As previously mentioned, our goal was to have not only the highest percentage correct score but to have the lowest as possible risk numbers. With that being said, we decided to set up parameters around the risk number as we were building our models. We decided that a risk number of 0.1 and above would be too high, as that would impact the reliability of outcomes.

As stated in the Validation section, we used the method in which we divided the data into training and test sample. Based on various experimentations, we found that our optimal split was 2:1, where we took around 70-75% of the data for training and 25-30% of the data for testing. This allows us to see how replicable our results are and to make sure that the high percentages were not a fluke.

Based on the graph we can create some rules to automate the diagnosis based on the profile data:

* If the Parameter Worse is <=88.540 then with 98.3% confidence we can diagnose the tumor to be benign
* If the Perimeter Worse measurement is in the range 88.54 to 101.40 AND the Concavity Mean is <= 0.1226 AND Texture SD is <= 1.916 then we can conclude the tumor is benign with 98.1% confidence
* But, predictability when Perimeter Worse is in the range (88.54 to 101.40) AND Concavity Mean is >0.12260 The the confidence of predicting B or M is 50%. For models like this, we cannot be sure, and hence we should not use it to predict.
* The model predictability when Perimeter Worse is >122.4 is 98.6% for being malignant.

## Practical Application & Recommendations

Decision trees were used in many of the applications earlier. The practical aspect of a decision tree can be divided into two parts: a personal decision tree and business decision tree. Some of the personal use cases where we use the decision tree algorithm and do not realize are booking a flight, renting or buying a home, planning a vacation, etc.

Some of the areas where we use the decision tree in business or research are: Song et.al. uses a decision tree for the analysis of risk factors related to major depressive disorder (MDD) in a four-year cohort study. Some of the hidden use cases of the decision tree in research are also well explained by Song et,al namely Variable selection: assessing the relative importance of variables, handling of missing values, prediction, and data manipulation. Fair Isaac commonly known for FICO is the pioneer in using decision trees. Earlier they used to use a decision tree for fraud detection. Now as in [7] they have a product called FICO® Analytic Modeler Decision Tree Professional which integrates data-driven rules with policies to target segments that support a company’s objectives. Analysts can quickly access and begin using a robust segmentation tool that evaluates multiple business perspectives simultaneously.

## 

## Reference

1. Breast cancer. (2019, November 22). Retrieved from https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475.
2. Decision Trees. (n.d.). Retrieved from https://www.sciencedirect.com/topics/computer-science/decision-trees.
3. Song YY, Lu Y: Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015; 27(2): 130–5.Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/
4. Mutasem Sh. Alkhasawneh, Umi Kalthum Ngah, Lea Tien Tay, Nor Ashidi Mat Isa, and Mohammad Subhi Al-Batah, “Modeling and Testing Landslide Hazard Using Decision Tree,” Journal of Applied Mathematics, vol. 2014, Article ID 929768, 9 pages, 2014.
5. Quest Node. (n.d.). Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3RA7\_15.0.0/com.ibm.spss.modeler.help/questnode\_general.htm.
6. Decision Tree Applications for Data Modelling (Artificial Intelligence). (n.d.). Retrieved from http://what-when-how.com/artificial-intelligence/decision-tree-applications-for-data-modelling-artificial-intelligence/.
7. FICO® Analytic Modeler Decision Tree Professional. (n.d.). Retrieved from https://www.fico.com/en/products/fico-analytic-modeler-decision-tree-professional.