Decision Tree Model

Kuntal Patel

University of North Texas

**Introduction**

Breast Cancer is one of the most common type of the cancers in women which is affecting approximately 12.5% of all women in all around of the world. There are two types of breast cancers: Benign breast lump and Malignant breast lump. If we can recognize the diagnosis of the tumor in early stages of the cancer, we may cure the patient. For the purpose of accurately diagnose tumor either as a benign or malignant we are trying to come up with the model with CHAID (Chi-square Automatic Interaction Detector) algorithm based on decision tree. This algorithm is implemented in IBM SPSS which requires fine tuning of parameters. I have tried ECHAID, CRT and QUEST algorithms also to compare with CHAID.

Analysis is based on Wisconsin breast cancer data with ten variables. These variables are the measured values of affected tissue of the tumor. The idea is to build a model from training data and test how accurately we can predict the results. How variables and interrelations between them help us to decide whether the tumor is malignant or benign is the goal here.

**Why CHAID?**

At each step, CHAID chooses the independent (predictor) variable that has the strongest interaction with the dependent variable. Categories of each predictor are merged if they are not significantly different with respect to the dependent variable. In the given dataset Dependent variable (Diagnosis) is Nominal and independent variables are scalar which is proper setting to use CHAID.

**Descriptive Statistics**

Dataset has total 570 samples. Out of these 569 are labeled as a different feature like,

1) Radius

2) Texture

3) Perimeter

4) Area

5) Smoothness

6) Compactness

7) Concavity

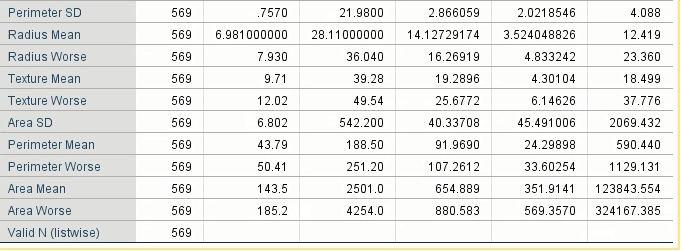
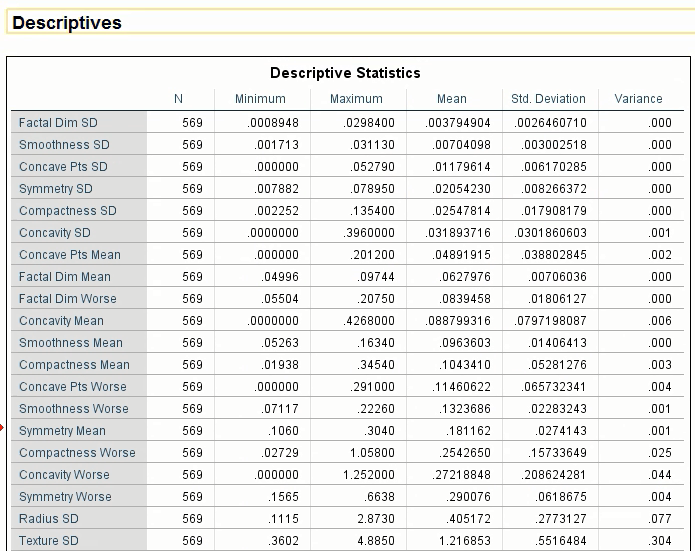
8) Concave points

9) Symmetry

10) Fractal dimension

These each have Mean, Standard deviation and worse values so in total we have 30 columns, and there are no missing data in this dataset other than 570th sample which we have removed before any analysis. Certainly, there are some measurements are zero which is not an issue because the other values are in the similar range. To understand the data distribution, I have used descriptive statistics in SPSS.

From this descriptive statistic we can have some idea of which variables have significant higher value than rest of the variables. For example, Perimeter Worse mean value is 107.2612 whereas Factal Dim Worse is 0. 839458.May be higher valued variables impact more on deciding type of tumor.



**Model Criteria**

The goal of the data analysis in the given dataset is to accurately predict the type of tumor. If the model will fail even for 10 cases out of 100, that will be significant and not allowed in medical diagnosis. To achieve high predictivity and less false negatives model can become more complex with a greater number of nodes, leaves and terminal nodes. The collected data comes from lab which has very low possibility of any kind of discrepancies in the data. Though the total samples are 569 which is very less compared to data generated by other applications, we can able to build a model and can predict the type of tumor with less number of wrong diagnosed cases, the model can be enhanced and become good prototype design for future product.

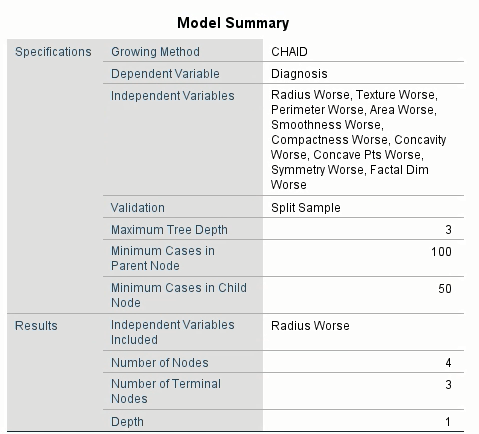
Initially the parameters on which we are focusing are:

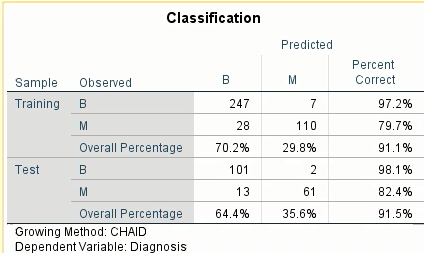
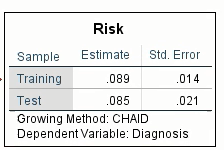
* Risk
* Total accuracy
* Alpha value
* Number of wrong diagnosis cases
* Tree complexity
* Adjusted p value
* Validation
* Minimum case in parent and child nodes
* Variable at the node from which splitting occurs
* Independent variables included
* Number of nodes on which we get 100% deterministic results
* All variables vs few variables from dataset

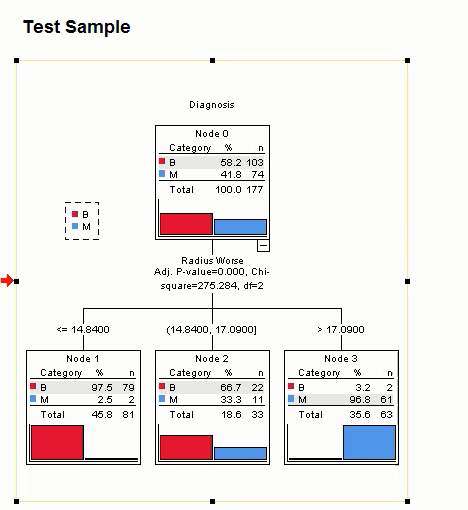
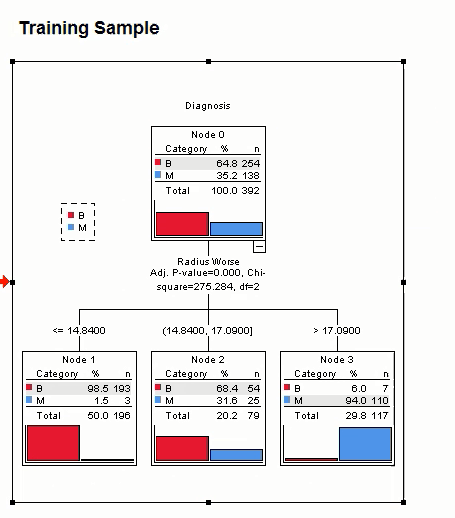
**Model Methodology**

I have performed several trial and error attempts to decide which parameters required tuning and why? Out of those trials I have included four attempts which gradually created better model after each change.

Case 1: Included Worse independent variables only with validation by splitting whole dataset into 67% as training and 33% as test data with all other default parameters and growing method as CHAID.

Results: 



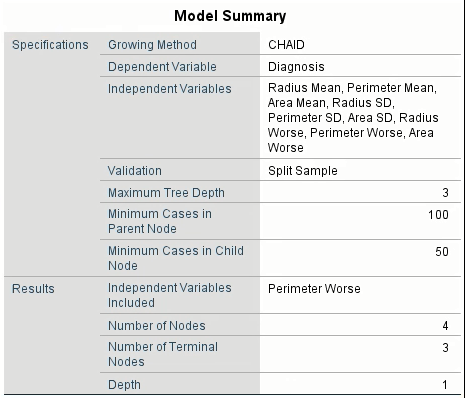


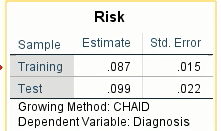
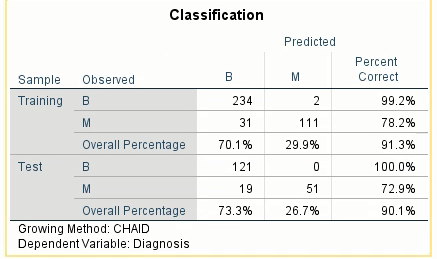
Analysis:

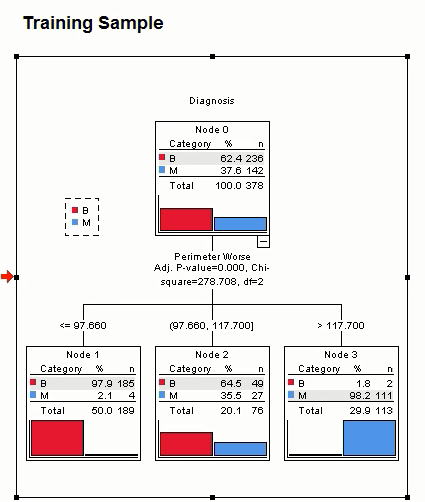
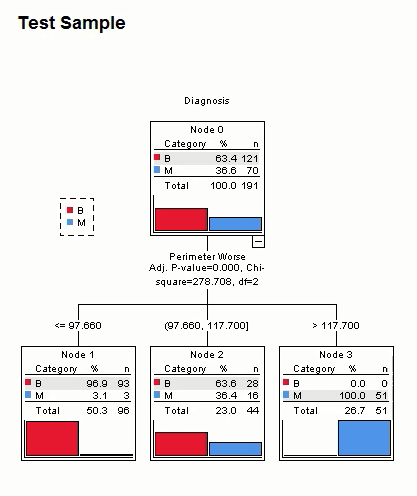
* Out of 4 nodes 3 nodes are terminal nodes.
* Tree depth – 1 level
* Less complex
* None of the node is deterministic to predict type of tumor 100%.
* High risk value:0.085

Case 2: Included Radius, Perimeter and Area variables only with validation by splitting whole dataset into 67% as training and 33% as test data with all other default parameters and growing method as CHAID.

Results:



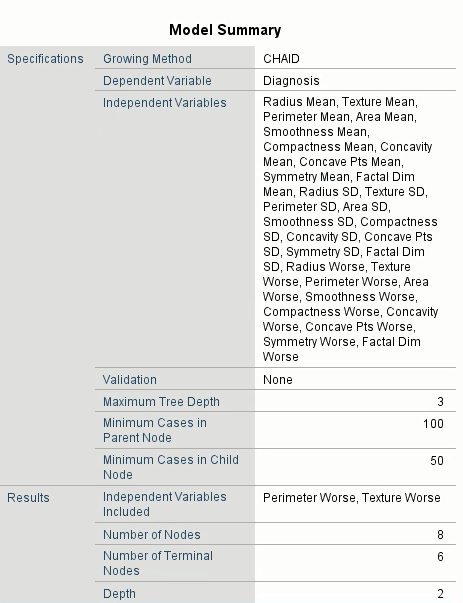
Analysis:

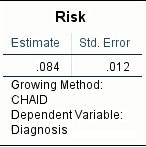
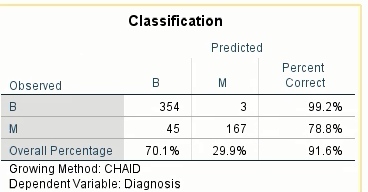
* Out of 4 nodes 3 nodes are terminal nodes.
* Tree depth – 1 level
* Less complex
* None of the node is deterministic to predict type of tumor 100%.
* High risk value:0.099

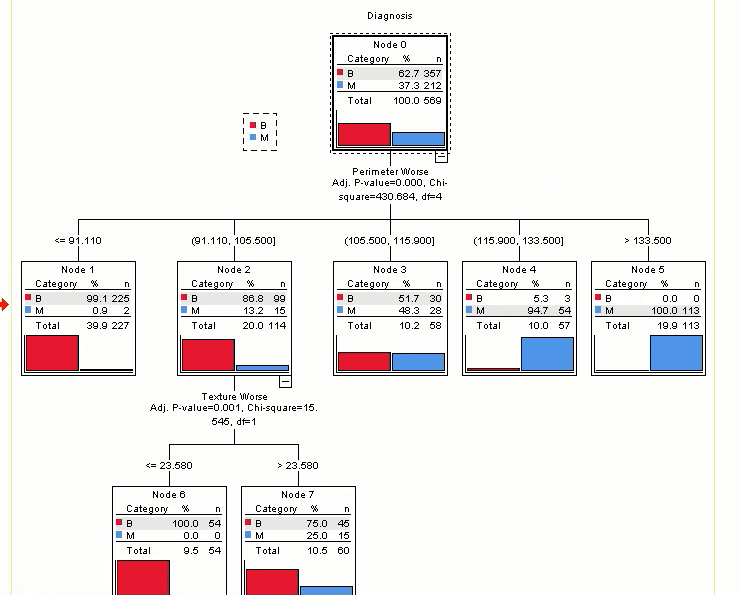
Since I am not including all variables in this case CHAID is not automatically detect pattern and can not automatically predict type of tumor. That is why risk value is even worse that case 1.

Case 3: Included all variables without any validation with all other default parameters and growing method as CHAID.

Results:



Tree diagram

Analysis:

* Out of 8 nodes 6 nodes are terminal nodes.
* Tree depth – 2 level
* Complex
* 2 nodes are deterministic to predict type of tumor 100%.
  + 1)Perimeter Worse >133.500
  + 2) Perimeter Worse is in the range of (91.110,105.500); Texture Worse <= 23.580.
* High risk value:0.084

Reason to not choosing Validation:

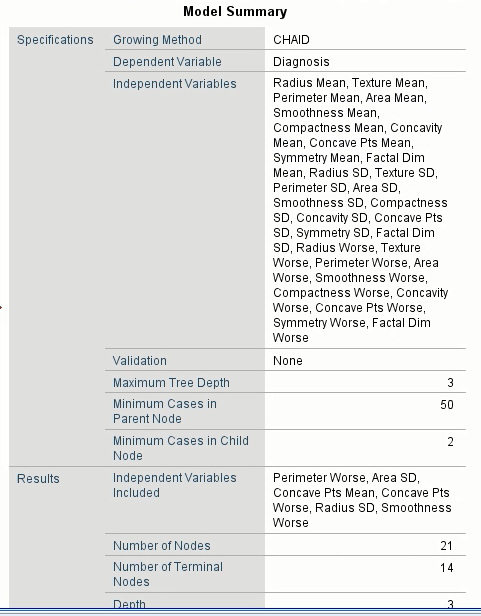
Split-sample validation should be used with caution on small data files (data files with a small number of cases). Small training sample sizes may yield poor models, since there may not be enough cases in some categories to adequately grow the tree. Since we have selected NONE option in validation, we are getting more complex tree compare to previous 2 cases.

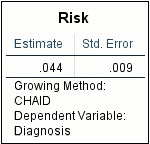
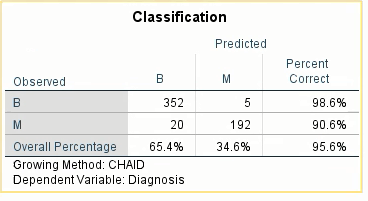
Case 4: Included all variables without any validation with all other default parameters but changed minimum parent nodes as 50 instead of default 100 and minimum child nodes as 2 instead of 50 and growing method as CHAID.

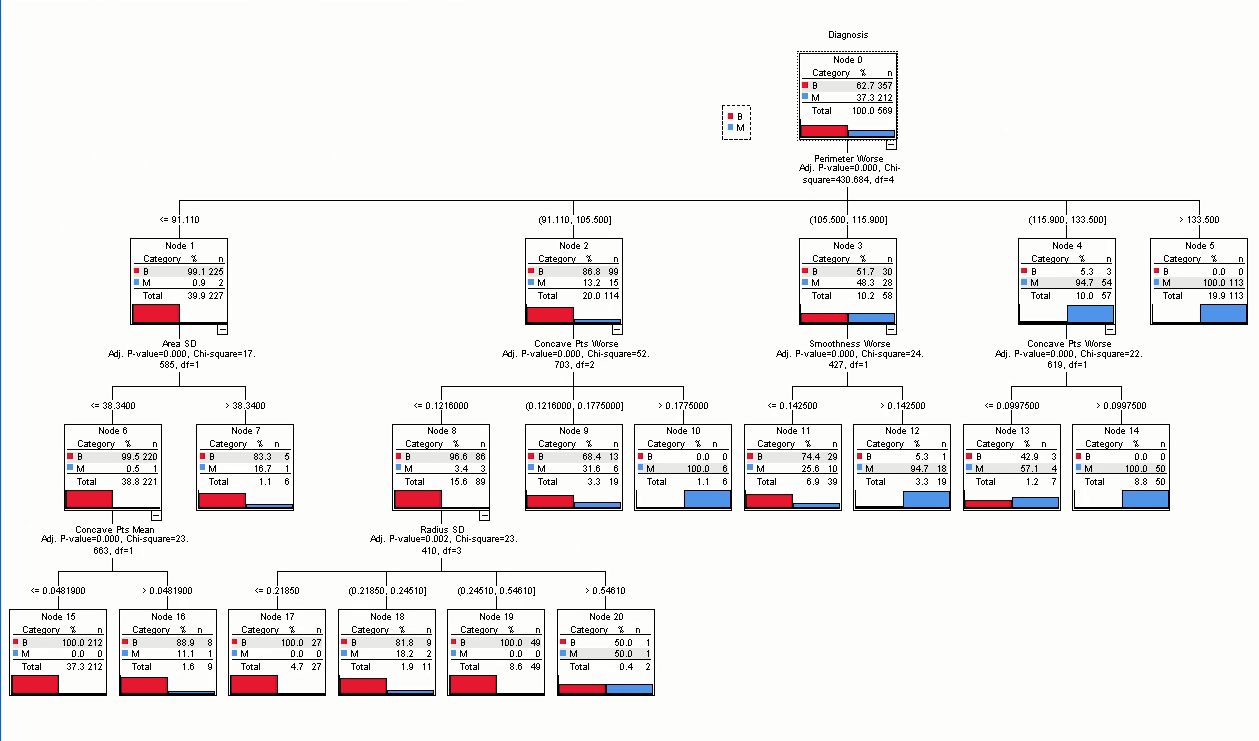
Reason to tune the parameter Minimum Number of Cases:

For data files with a small number of cases, the default values of 100 cases for parent nodes and 50 cases for child nodes may sometimes result in trees with no nodes below the root node; in this case, lowering the minimum values may produce more useful results.

Results:





Analysis:

* Out of 21 nodes 14 nodes are terminal nodes.
* Tree depth – 3 level
* Complex
* 6 nodes are deterministic to predict type of tumor 100%.
  + 1)Perimeter Worse >133.500
  + 2) Perimeter Worse is in the range of (91.110,105.500); Concave Pts Worse >0.1775000.
  + 3) Perimeter Worse is in the range of (115.900,133.500); Concave Pts Worse >0.0997500.
  + 4) Perimeter Worse <= 91.110; Area SD <= 38.3400; Concave Pts Mean <= 0.0481900.
  + 5) Perimeter Worse is in the range of (91.110,105.500); Concave Pts Worse <= 0.1216000; Radius SD <= 0.21850
  + 6) Perimeter Worse is in the range of (91.110,105.500); Concave Pts Worse <= 0.1216000; Radius SD is in the range of (0.24510,0.54610).
* Less risk value:0.044

Case 5: CRT, ECHAID and QUEST compared to CHAID

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CHAID | ECHAID | CRT | QUEST |
| Risk | 0.044 | 0.044 | 0.058 | 0.053 |
| Predictive | Out of 14 terminal nodes 6 (42.85%) | Out of 11 terminal nodes 4  (36.36%) | Out of 4 terminal nodes 0  (0%) | Out of 10 terminal nodes 0  (0%) |
| Overall Accuracy | 95.6% | 95.6% | 94.2% | 94.7% |

**Conclusion**

After tuning the parameter ‘Minimum number of parent and child nodes values’, Case 4 can predict the type of tumor 42.85% times better than 33.33% of case 3 using CHAID. Even after tuning parameters and using ECHAID, CRT or QUEST the model was not improving overall accuracy and not lowering Risk value further.

Reference:

<ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/client/Manuals/IBM_SPSS_Decision_Trees.pdf>