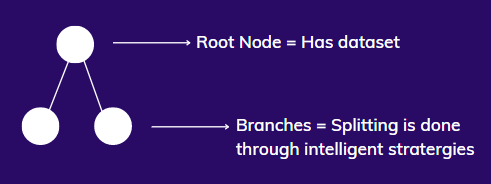
**DECISION TREE USING Chi-square automatic interaction detection (chaid)**

**What is Decision Tree?**

Decision tree is a divide and conquer problem solving strategy.



A decision tree is a non-parametric supervised learning approach that may be used for classification as well as regression applications. It has a tree structure that is hierarchical and consists of a root node, branches, internal nodes, and leaf nodes.

Kass initially introduced this technique in 1980. This method, as the name suggests, is based on the chi-square statistic. A Chi-square test produces a probability value that ranges between 0 and 1.

A chi-square score close to zero implies a considerable difference between the two groups being compared. Similarly, a number around 1 suggests that there is no discernible difference between the two classes.

### Important Terminology related to Decision Trees

1. **Root Node:**It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:**It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:**When a sub-node splits into further sub-nodes, then it is called the decision node.
4. **Leaf / Terminal Node:**Nodes do not split is called Leaf or Terminal node.
5. **Pruning:**When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
6. **Branch / Sub-Tree:**A subsection of the entire tree is called branch or sub-tree.
7. **Parent and Child Node:**A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node

**Why Decision Tree?**

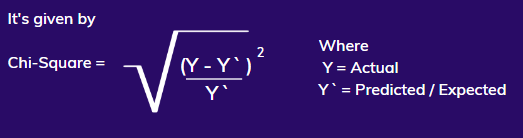
* These are predictive model with higher accuracy.
* Simple to understand.
* It will automate the process so that any distribution can be solved.
* In the decision tree collections of record is called Training Set.
* Each record has set of attributes and one of them is class.
* We find a model for the class attribute as a function of values of other attributes.
* Example: Anomaly Detection can be written as follows ; Anomaly = f(Amount, Time, Location)

**Variables Used in CHAID Algorithm**

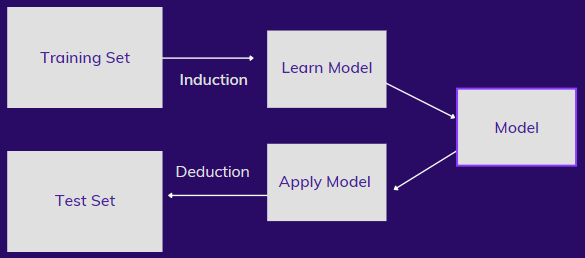
* Variable to be predicted that is, Dependent variable: Continuous OR Categorical
* Independent variables: Categorical ONLY (can be more than 2 categories)
* Thus, if there are continuous predictor variables, then we need to transform them into categorical variables before they can be supplied to the CHAID algorithm.
* Statistical Tests used to determine the next best split:
* Continuous Dependent Variable: F-Test (Regression Problems)
* Categorical dependent Variable: Chi-square (Classification Problems)
* At each phase, the independent variable having the greatest interaction with the dependent variable is chosen.
* If the interaction is small, it is less relevant and will be merged later.

**What is Chi-Square?**

It's a measurement metric to find the important feature used to find distance between the child and the parent nodes.



**Block Diagram of Decision Tree working:**



**Tree Induction**

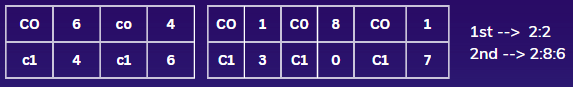
Split the record based on an attribute test that optimizes certain criterion.



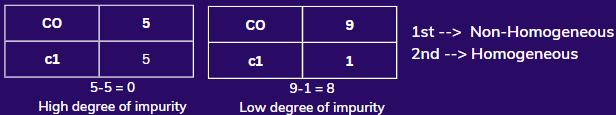
The attribute either falls in maximum or minimum category. The one with maximum is considered. Some of the issues to be addressed:

1. How to specify the attribute test condition?
2. How to determine the best split?
3. When to stop the split?

* Split can be: two-way split and Multi way split.
* Split on continuous attribute are of two types: Discrete(High, medium & Low) and Binary(Yes or No)
* The optimization happens in such a way that the difference in percentage of distribution should be high.



* Nodes with homogeneous class distributions are preferred



* Splitting should be stopped when all records belong to same class or similar attribute value.

**Advantages:**

* Inexpensive to construct.
* Less Data Preparation and Versatility.
* Extremely fast at classifying unknown records.
* Easy to interpret for small sized trees.
* Accuracy is comparable to other classification technique for many simple datasets.

**Drawbacks:**

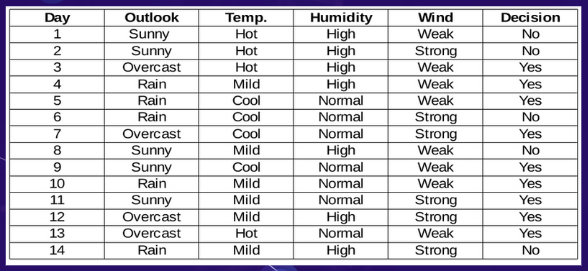
* Overfitting : Pruning

Pruning is a data compression method used in machine learning and search algorithms to minimise the size of decision trees by deleting non-critical and redundant portions of the tree used to categorise instances.

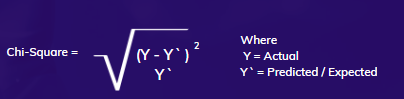
* They are largely unstable compared: A small change in data can lead to major change in the tree.
* Decision tree often involves higher time to train the model.

**Implementation of the Algorithm:**

We have a well-known Golf Dataset. The main objective of this classification activity is to decide if or not to go out to play Golf sport.



* 1. Humidity feature which has a split of High and Normal:



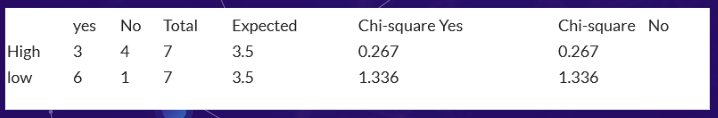
Chi-square yes for high humidity is √(( 3– 3.5)2 / 3.5) = 0.267

whereas actual is 3 and expected is 3.5.

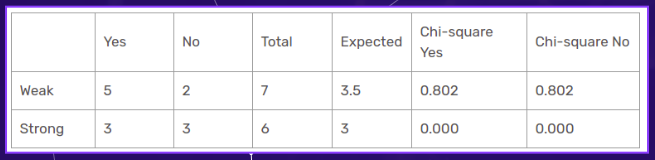
So, the chi-square value of the humidity feature is

= 0.267 + 0.267 + 1.336 + 1.336

= 3.207



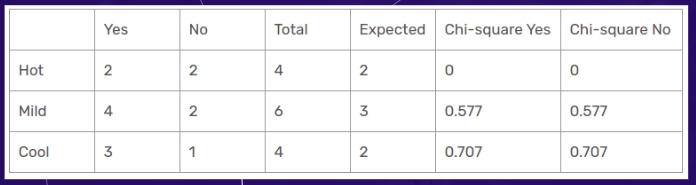
* 1. Wind feature which has a split of Weak and Strong:



Herein, the chi-square test value of the wind feature is

= 0.802 + 0.802 + 0 + 0 = 1.604

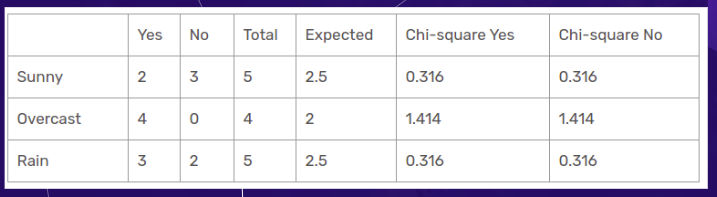
3) Temperature feature which has a split of Hot, Mild and Cool:



Herein, the chi-square test value of the temperature feature is

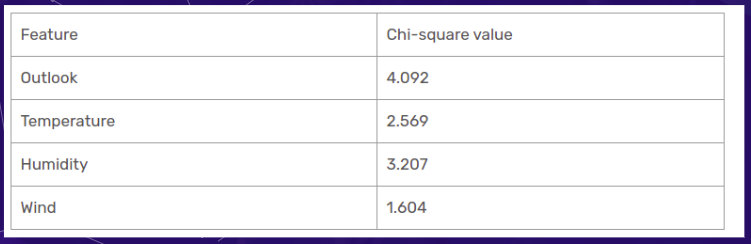
= 0 + 0 + 0.577 + 0.577 + 0.707 + 0.707 = 2.569

4) Outlook feature which has a split of Sunny, Outcast and Rain:

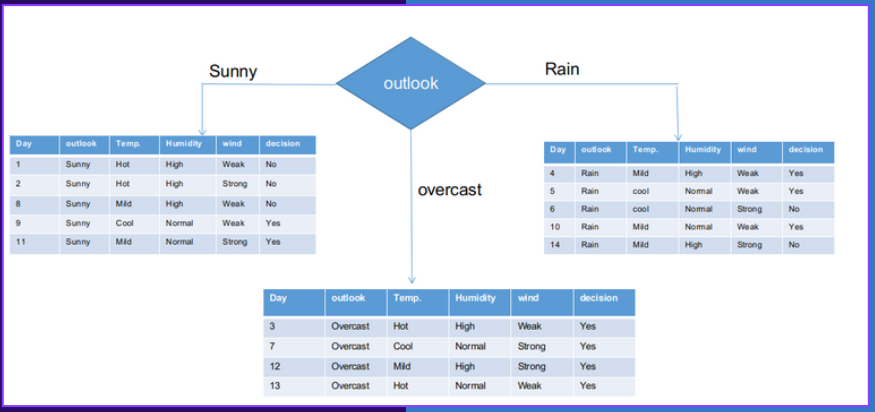


Herein, the chi-square test value of the outlook feature is = 0.316 + 0.316 + 1.414 + 1.414 + 0.316 + 0.316 = 4.092

The outlook column has the most elevated and highest chi-square value.

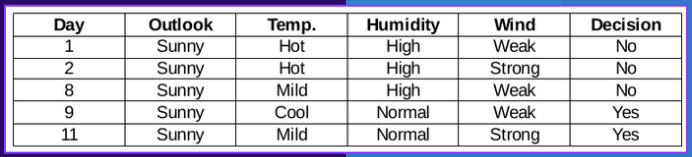


On the image below, we've split the raw data based on the outlook classifications. In the sub informational dataset.

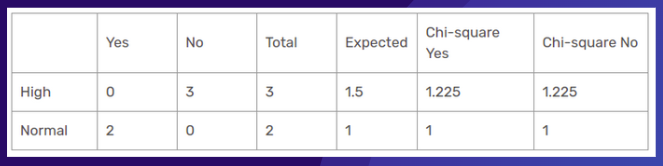


* 1. Outlook is Sunny ---> Humidity, Wind and Temperature:

Both sunny and rain branches have yes and no decisions. And Outcast has all Yes in it. We will apply chi-square tests for these sub informational datasets.

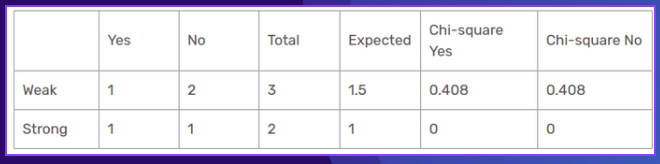


1. Humidity feature for when the outlook is Sunny:



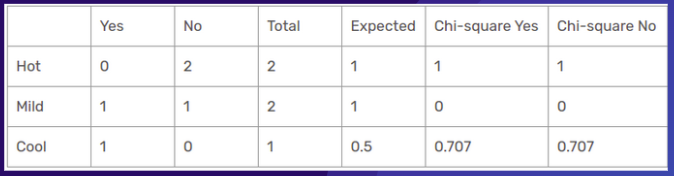
Chi-square value of humidity feature for sunny outlook is = 1.225 + 1.225 + 1 + 1 = 4.449

1. Wind feature for when the outlook is Sunny:



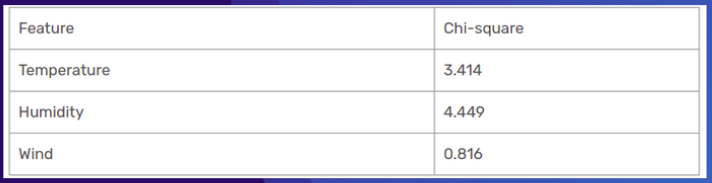
Chi-square value of wind feature for sunny outlook is = 0.408 + 0.408 + 0 + 0 = 0.816

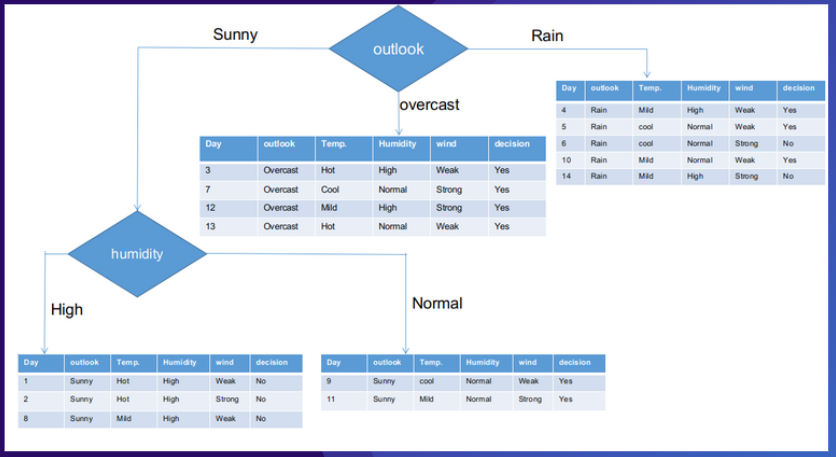
1. Temperature feature for when the outlook is Sunny:



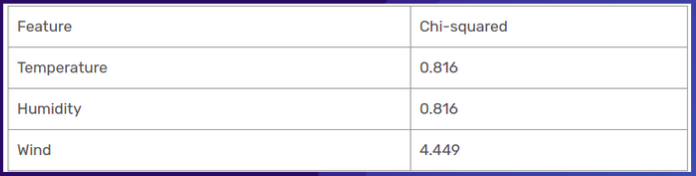
So, the chi-square value of temperature feature for sunny outlook is = 1 + 1 + 0 + 0 + 0.707 + 0.707 = 3.414

Therefore Humidity has the highest Chi-square value so its converted as a child node.



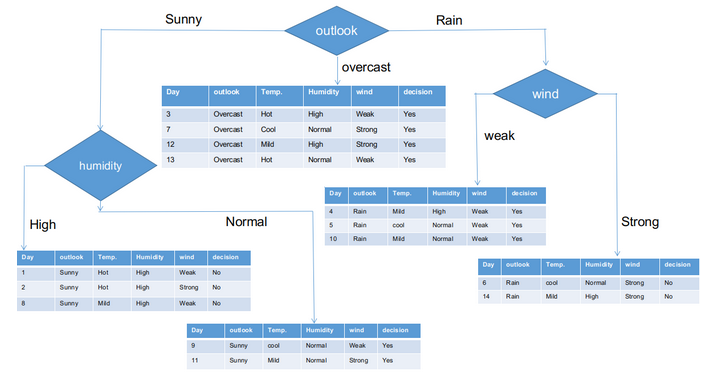


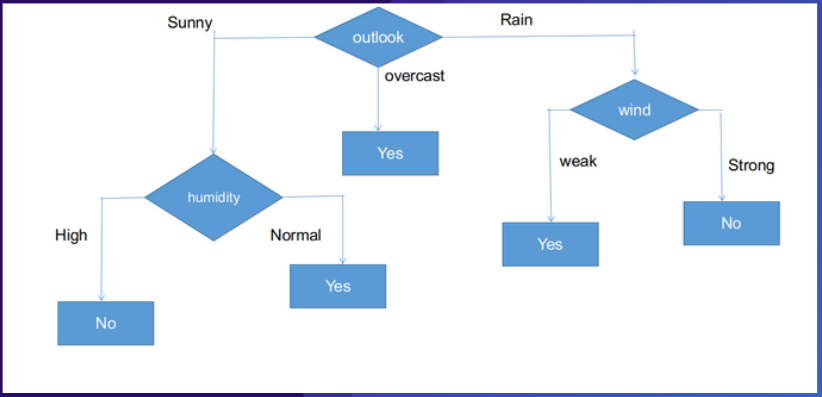
Similarly do it with respect to Rain feature and the chi-square value is given below.



Therefore Wind has the highest Chi-square value so its converted as a child node.

**The final CHAID tree is given below:**





**A general algorithm for a decision tree can be described as follows:**

1. Pick the best attribute/feature. The best attribute is one which best splits or separates the data.
2. Ask the relevant question.
3. Follow the answer path.
4. Go to step 1 until you arrive to the answer.

**References:**

* 1. <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>
  2. <https://www.analyticsvidhya.com/blog/2021/05/implement-of-decision-tree-using-chaid/>
  3. <https://www.analyticsvidhya.com/blog/2021/05/implement-of-decision-tree-using-chaid/>
  4. <https://sefiks.com/2020/03/18/a-step-by-step-chaid-decision-tree-example/>
  5. <https://www.researchgate.net/publication/315348917_CHAID_Decision_Tree_Methodological_Frame_and_Application>
  6. https://www.tandfonline.com/doi/full/10.1080/13467581.2019.1696203