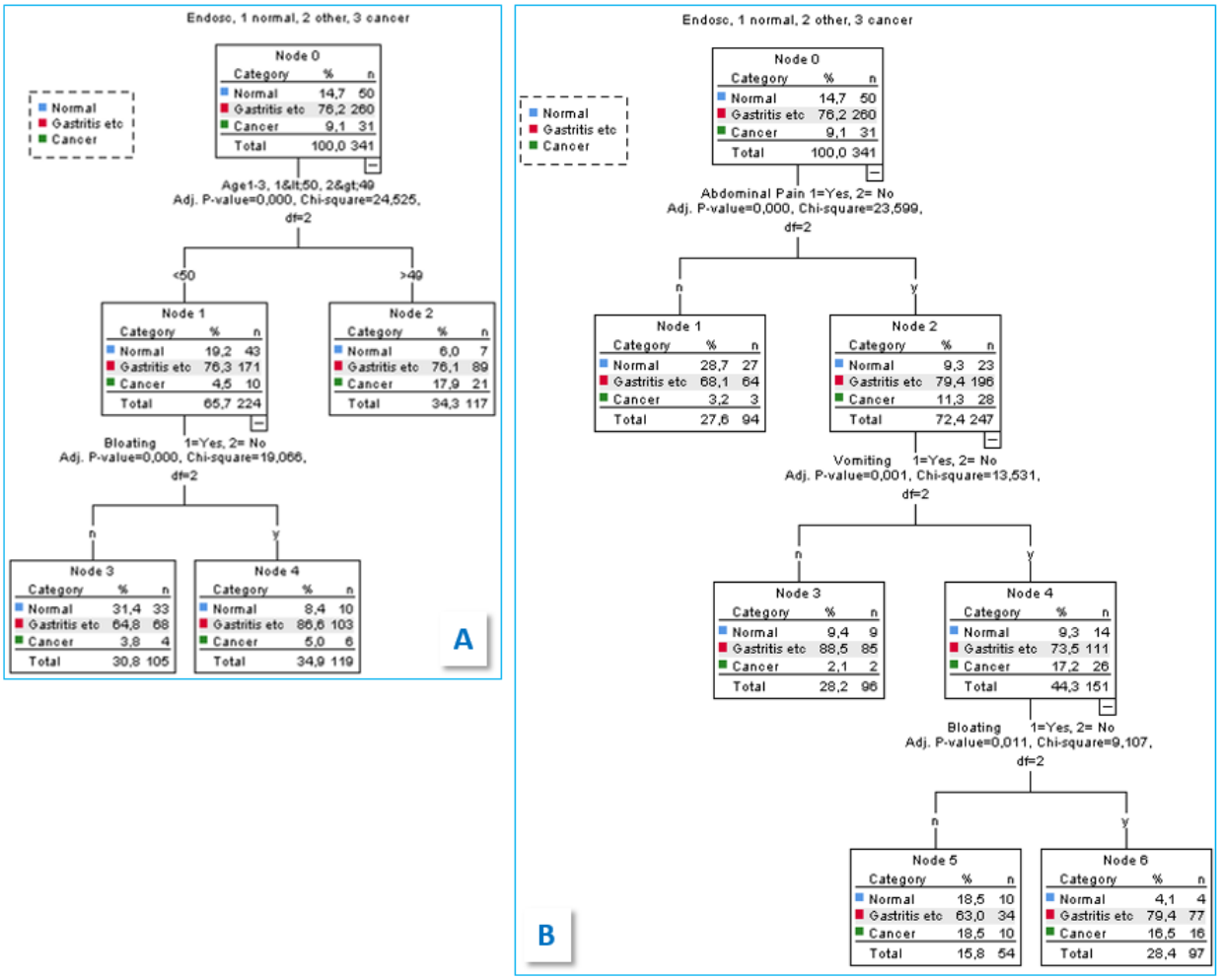
Supplementary File

Supplementary Table 1 Principal components analysis analysis for risk factors, symptoms and *H. pylori* tests which were available for each probands setting the number of factors to three

Parameter	Component 1	Component 2
UBT	0.922	0.166
RUT	0.919	0.147
Infection status summary	0.911	0.197
Heartburn	0.048	0.783
Vomiting	-0.085	0.564
Reflux	-0.119	0.702

Rotated component matrix / revised factor loadings. Rotation method Varimax with Kaiser Normalization. UBT: Urea breath test; RUT: Rapid urease test.



Supplementary Figure 1 SPSS decision trees for GC risk factors. A: Input of all risk factors. Age is the first separator followed by bloating for the patients older than 49; B: Input of all risk factors except age. The first group separation occurs due to the factor abdominal pain followed by vomiting and bloating.

Explanation S1: Terminology related to Decision Trees

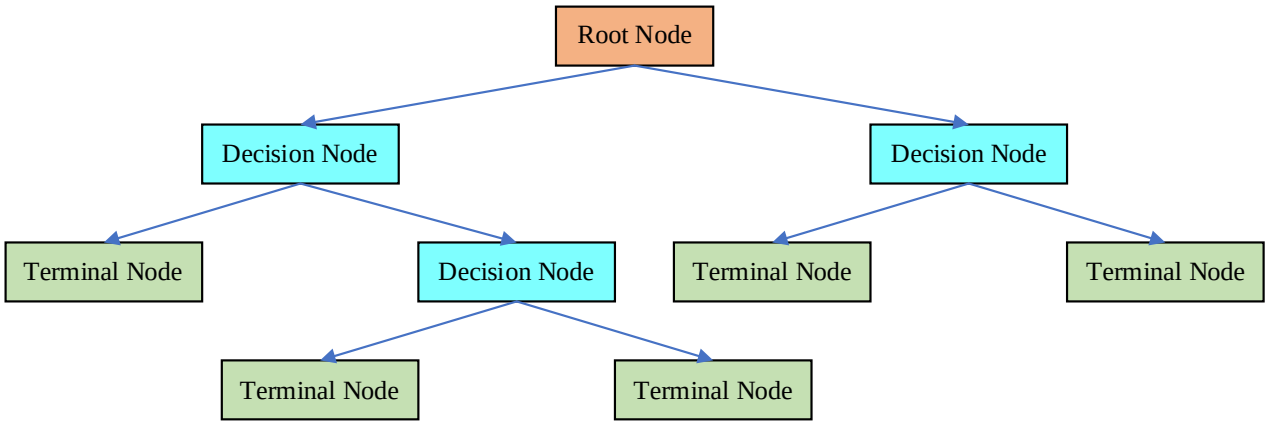
(for more information: https://rpubs.com/anshulkumar/KNN-DecTree-RandForest, Accessed Aug 19, 2021)

Root Node: It represents the entire population of samples and it is further divided into two or more homogeneous sets.

Splitting: It is a process of dividing a node into two or more sub-nodes.

Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node.

Leaf / Terminal Node: Nodes which do not split are called Leaf or Terminal node.



In a decision tree the major challenge is to identify the attribute for the root node in each level. This process is known as attribute selection. We have two popular attribute selection measures:

Information Gain

It is necessary to understand entropy.

Entropy: Entropy is the measures of impurity, disorder, or uncertainty in a bunch of examples. Entropy controls how a decision tree decides to split the data. It affects how a decision tree draws its boundaries. Entropy values range from 0 to 1, low values of entropy are needed.

$$Entropy = -\sum_{i}^{C} p_i \log_2 p_i$$

Where

- C - number of classes and
- p_i - probability of randomly picking an element of class i (i.e. the proportion of the dataset made up of class i)

Information gain is a statistical property that measures how well a given attribute separates the training examples according to their target classification. Constructing a decision tree is all about finding an attribute that returns the highest information gain and the smallest entropy. Information gain is a decrease in entropy. It computes the difference between entropy before split and average entropy after split of the dataset based on given attribute values.

$$Information\ Gain = Entropy_{before} - \sum_{k=1}^{n} w_k * Entropy_k$$

Where

$Entropy_{before}$ - Entropy before split,

n - number of child nodes,

$Entropy_k$ - Entropy of k^{th} child node and

Wk - Weight of k^{th} schild node (samples in kth child node / total number of samples in parent node)

Gini Gain/Index

Gini Impurity is the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly labeled according to the class distribution in the dataset. It is calculated as

$$Gini\ Impurity = \sum_{i=1}^{C} p_i * (1 - p_i)$$

Where

- C is the number of classes and
- p_i is the probability of randomly picking an element of class i.

When training a decision tree, the best split is chosen by maximizing Gini Gain, which is calculated by subtracting the weighted impurities of the branches from the original impurity.

Entropy vs. Gini Impurity

The maximum value for entropy is 1 whereas the maximum value for Gini Impurity is 0.5. As the Gini Impurity does not contain any logarithmic function to calculate, it takes less computational time as compared to entropy.