Maimoona Khilji
BS-DS
Semester 6

Lab Submission 01

Instructor: Basit Ali

Knowledge report of the Decision Tree of Titanic Training Dataset

Decision tree:

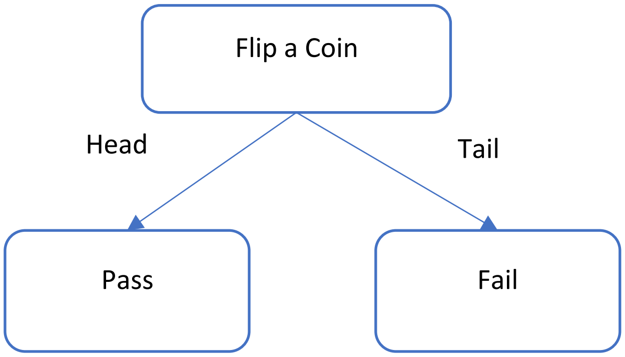
Decision tree is like a flowchart structure that consists of nodes: Root, intermediate and leaf nodes. Root node represents a splitting rule for one specific attribute. It returns further intermediate node on the basis of decision. Leaf node represents a decision or class labels (what we want to find out) and branches represents conjunctions of features that lead to those classes. Each node in the tree is a decision rule.

Initializing from the root node, a feature is evaluated and one of the two node is selected. This procedure is repeated until a final leaf is reached, which normally represents the **Target.**

- **Root Node:** The node that starts the graph. It evaluates the variable that best splits the data.
- **Intermediate Node:** The nodes where variables are evaluated but which are not the final nodes.
- **Leaf Node:** These are the final nodes of the tree, where the predictions of a category or a numerical value are made.

For example

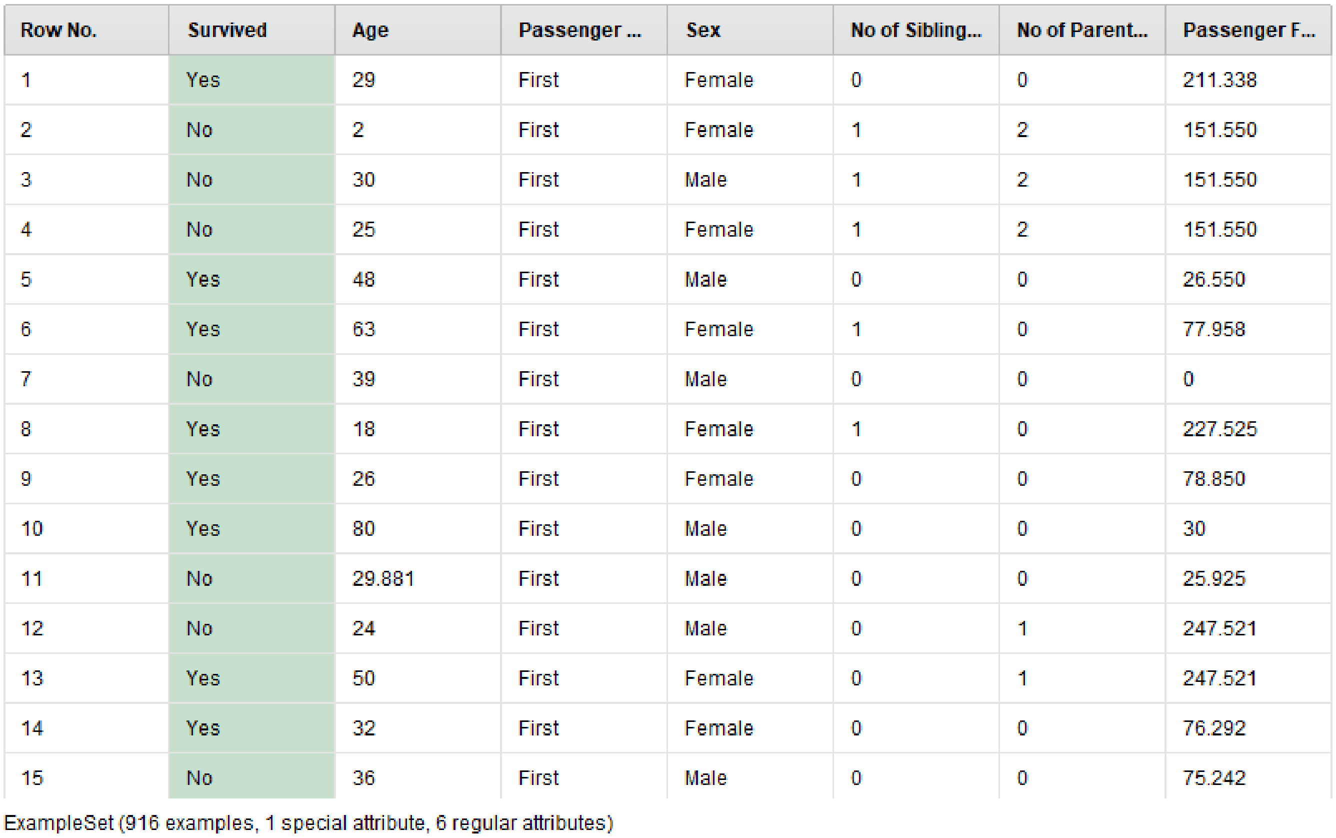
Here we want to find, when a person will win or lose the match on the basis of occurrence of head and tail of a coin. Leaf nodes that are Pass and Fail represents the result (that we want to interpret).



Maimoona Khilji
BS-DS
Semester 6

Dataset:

Here we plot decision on the sample dataset **Titanic Training** from Rapid Miner. It consists of 7 attributes and 916 records. In attribute, there is one special attribute **Survived** that represents whether passenger survived or not and 6 regular attributes that plays important role in defining or predicting survival rate of passenger.

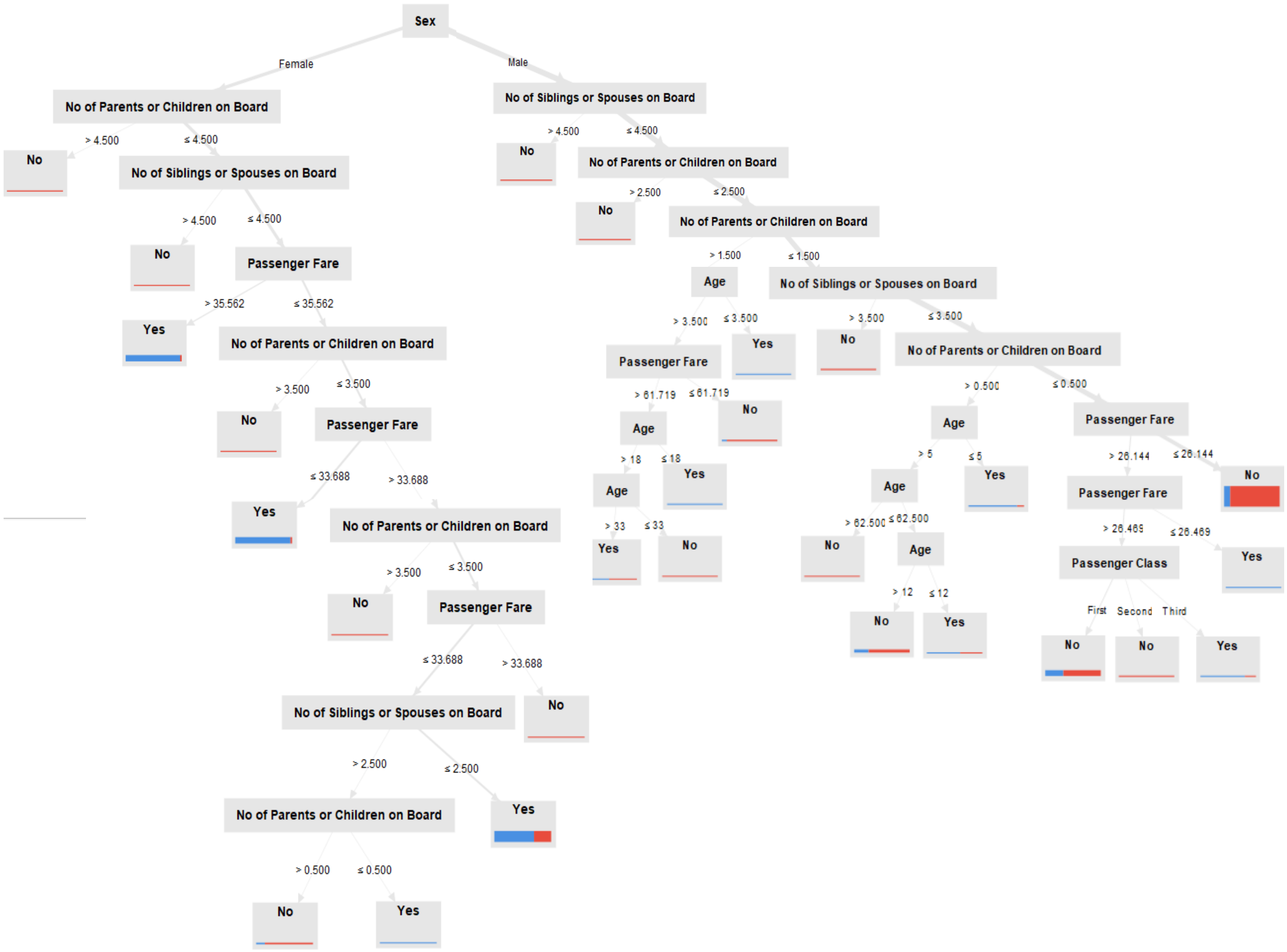


Row No.	Survived	Age	Passenger ...	Sex	No of Sibling...	No of Parent...	Passenger F...
1	Yes	29	First	Female	0	0	211.338
2	No	2	First	Female	1	2	151.550
3	No	30	First	Male	1	2	151.550
4	No	25	First	Female	1	2	151.550
5	Yes	48	First	Male	0	0	26.550
6	Yes	63	First	Female	1	0	77.958
7	No	39	First	Male	0	0	0
8	Yes	18	First	Female	1	0	227.525
9	Yes	26	First	Female	0	0	78.850
10	Yes	80	First	Male	0	0	30
11	No	29.881	First	Male	0	0	25.925
12	No	24	First	Male	0	1	247.521
13	Yes	50	First	Female	0	1	247.521
14	Yes	32	First	Female	0	0	76.292
15	No	36	First	Male	0	0	75.242

ExampleSet (916 examples, 1 special attribute, 6 regular attributes)

Maimoona Khilji
BS-DS
Semester 6

Decision Tree of Survival rate in Titanic



Maimoona Khilji
BS-DS
Semester 6

Criterion for root node

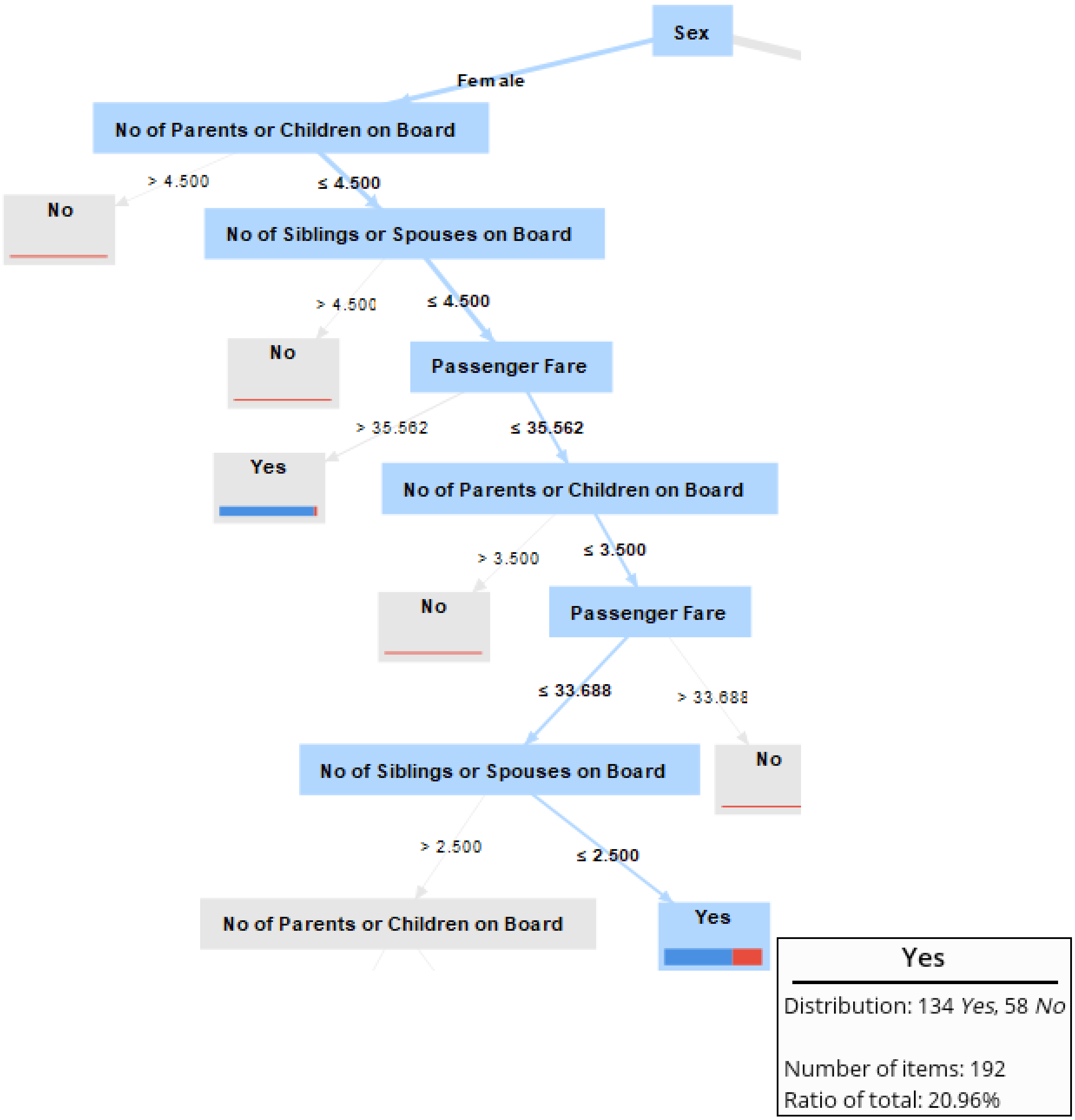
Selects the criterion on which Attributes will be selected for splitting. It can have one of the following values:

- **Information gain:** The entropies of all the Attributes are calculated and the one with least entropy is selected for split. This method has a bias towards selecting Attributes with a large number of values.
- **Gain ratio:** A variant of information gain that adjusts the information gain for each Attribute to allow the breadth and uniformity of the Attribute values.
- **Gini index:** A measure of inequality between the distributions of label characteristics. Splitting on a chosen Attribute results in a reduction in the average gini index of the resulting subsets.
- **Accuracy:** An Attribute is selected for splitting, which maximizes the accuracy of the whole tree.
- **Least square**: An Attribute is selected for splitting, that minimizes the squared distance between the averages of values in the node with regards to the true value.

Important Points:

- There are two Binomial attributes:
    - Sex
    - Survived
- As our objective is to find the **survival rate**, so this special attribute will be taken as leaf node. In other words, branches will made or nodes will be divide till the final leaf nodes reached that is **survived**.
- The other binomial attribute is Sex, so this will be taken as root node. As it splits the whole data set in a best way. It can be chosen using criterion parameter of **Accuracy**.

- We will use this path because this gives the details of 20% females of total population.
- We can predict that in **192 Females,** 134 females survived Because
    - They were with entropy value of **No of Parents or Children on Board (attribute)** less than or equal to **3.5**.
    - Having entropy of **No of siblings or spouses on board (attribute)** less than or equal to **2.5.**
    - Paid less fare, with entropy of **Passenger fare (attribute)** less than or equal to **33.68**

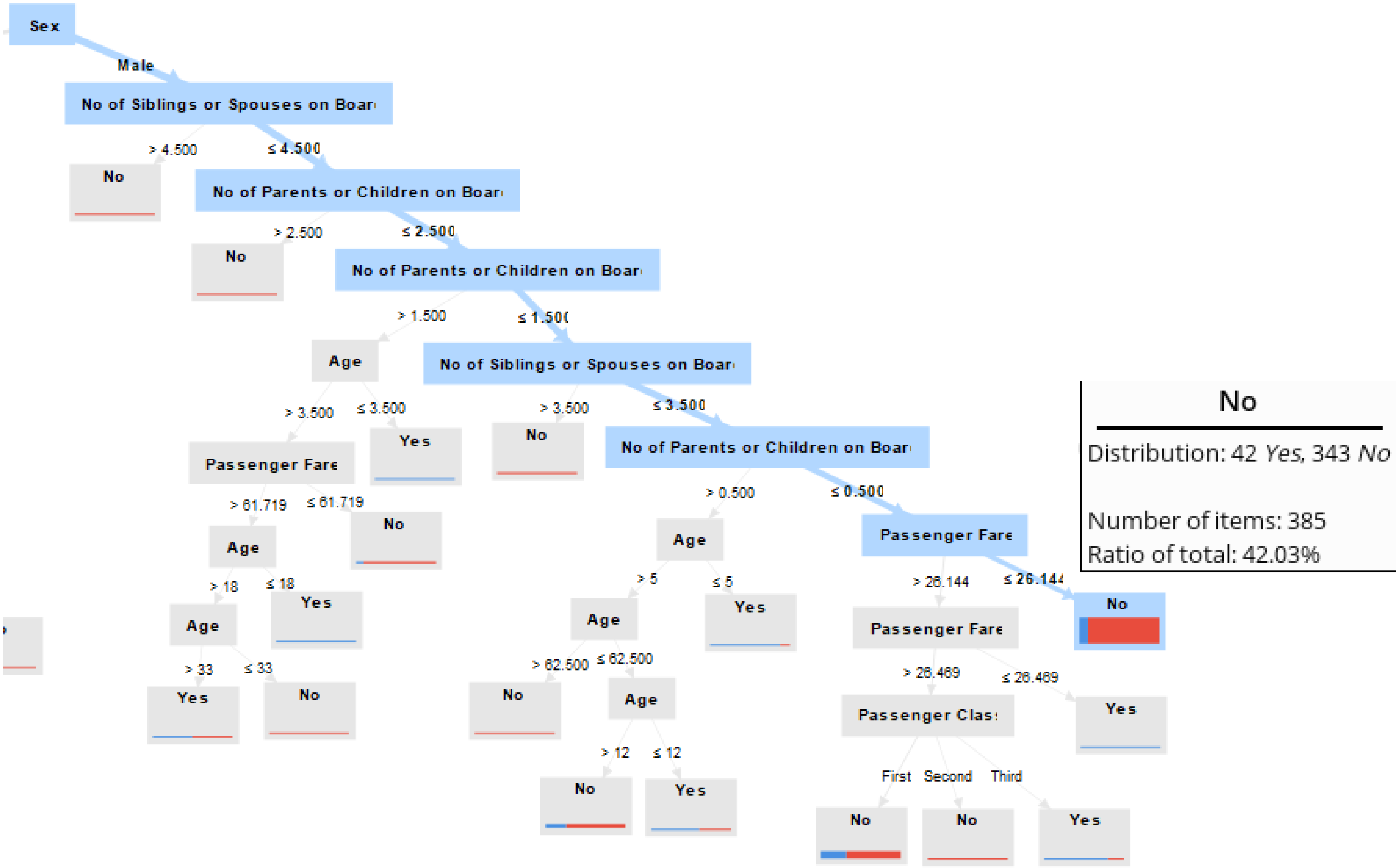

Sex
Female
No of Parents or Children on Board
> 4.500
≤ 4.500
No
No of Siblings or Spouses on Board
> 4.500
≤ 4.500
No
Passenger Fare
> 35.562
≤ 35.562
Yes
No of Parents or Children on Board
> 3.500
≤ 3.500
No
Passenger Fare
≤ 33.688
> 33.688
No of Siblings or Spouses on Board
No
> 2.500
≤ 2.500
No of Parents or Children on Board
Yes
Yes
Distribution: 134 Yes, 58 No
Number of items: 192
Ratio of total: 20.96%

- We will use this path because this gives the details of 42.03% males of total population.

- We can predict that in **385 males,** 343 males did not survived Because
  - They were with entropy value of **No of Parents or Children on Board (attribute)** less than or equal to **0.5**

Maimoona Khilji
BS-DS
Semester 6

- Having entropy of **No of siblings or spouses on board (attribute)** less than or equal to **3.5.**
- Paid less fare, with entropy of **Passenger fare (attribute)** less than or equal to **26.44.**



No

Distribution: 42 *Yes*, 343 *No*

Number of items: 385
Ratio of total: 42.03%