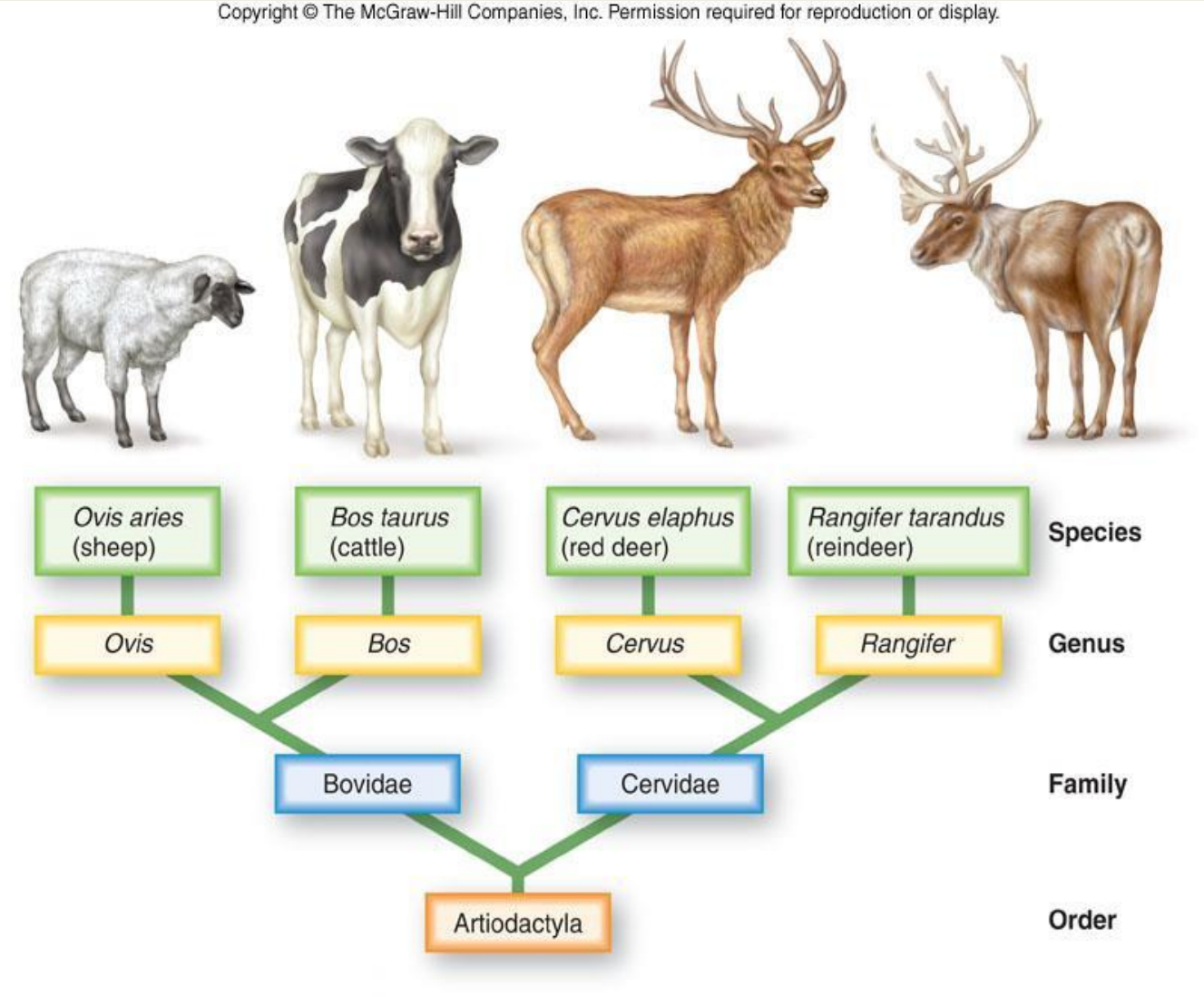
Comparative Study of C5.0 and CART algorithms

Presenter: Alvin Nguyen

Presentation Framework

1. What is Classification?
2. Decision Tree: Binary or Multi- branches
3. CART Overview
4. C5.0 Overview
5. Comparative Study of CART and C5.0 using Iris Flower Data
6. Comparative Study of CART and C5.0 using Titanic Data
7. Comparative Study of CART and C5.0 using Pima Indians Diabetes Data
8. Summary and Conclusion

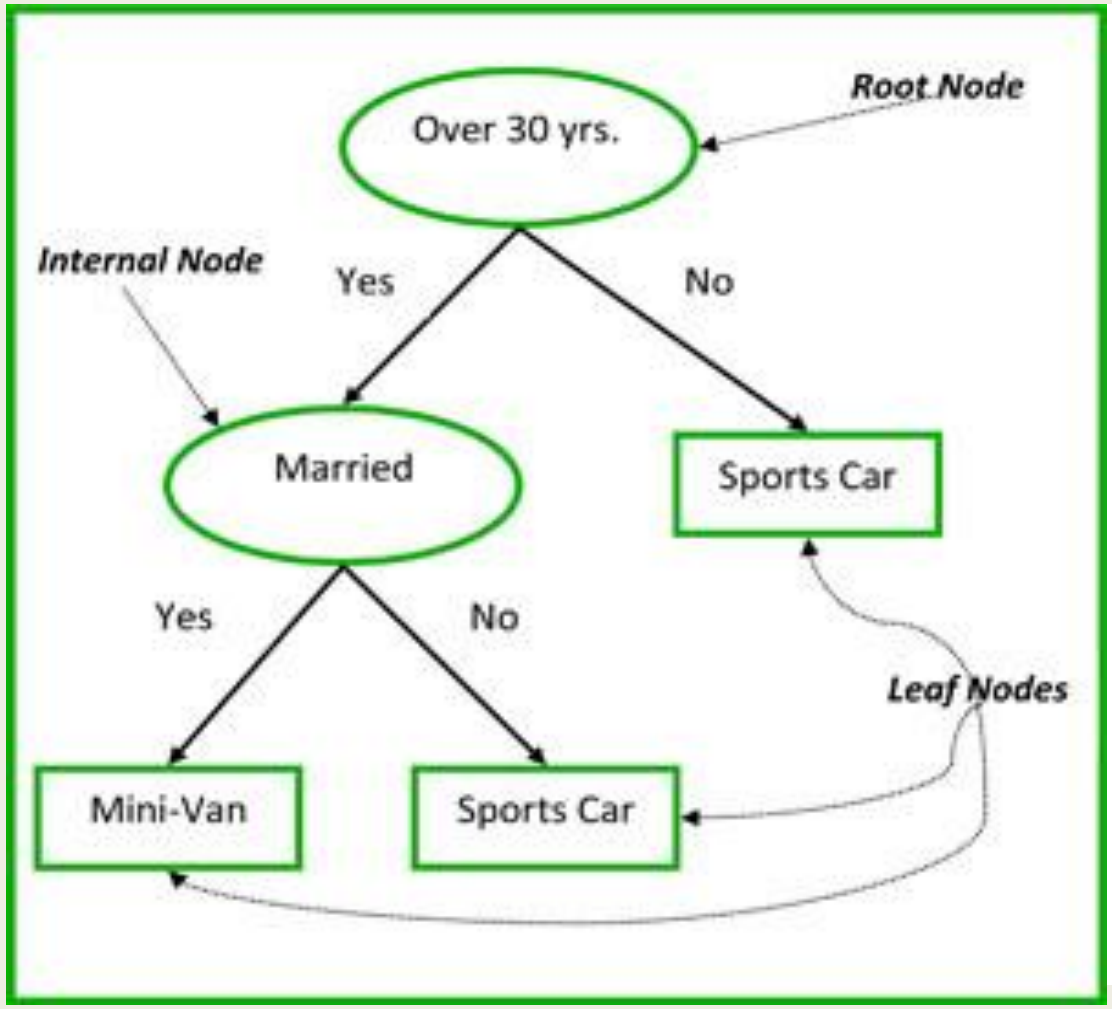
What is Classification in Data Mining?

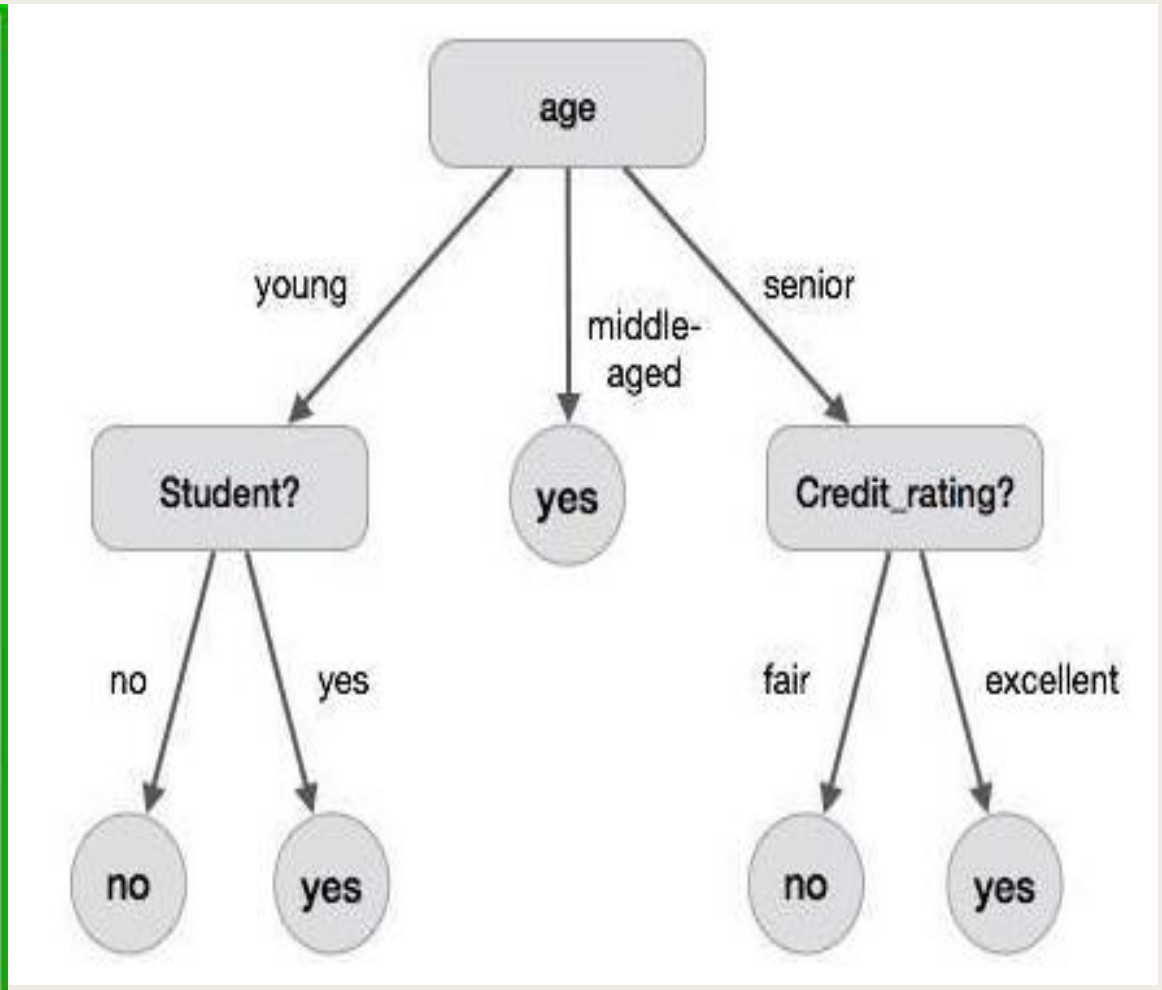


Oxford English Dictionary:

Classification is “the action or process of classifying something according to shared qualities or characteristics”.

Decision Tree: Binary or Multi-branches





CART algorithms (Classification & Regression Trees) by Breiman 1984

- A binary tree using GINI Index as its splitting criteria
- CART can handle both nominal and numeric attributes to construct a decision tree.
- CART uses Cost – Complexity Pruning to remove redundant braches from the decision tree to improve the accuracy.
- CART handles missing values by surrogating tests to approximate outcomes

C5.0 algorithm by Ross Quinlan

- C5.0 algorithm is a successor of C4.5 algorithm also developed by Quinlan (1994)
- Gives a binary tree or multi branches tree
- Uses Information Gain (Entropy) as its splitting criteria.
- C5.0 pruning technique adopts the Binomial Confidence Limit method.
- In a case of handling missing values, C5.0 allows to whether estimate missing values as a function of other attributes or apportions the case statistically among the results.

Comparative Study of C5.0 and CART using Iris Flower Data

Data Description:

150 samples in total

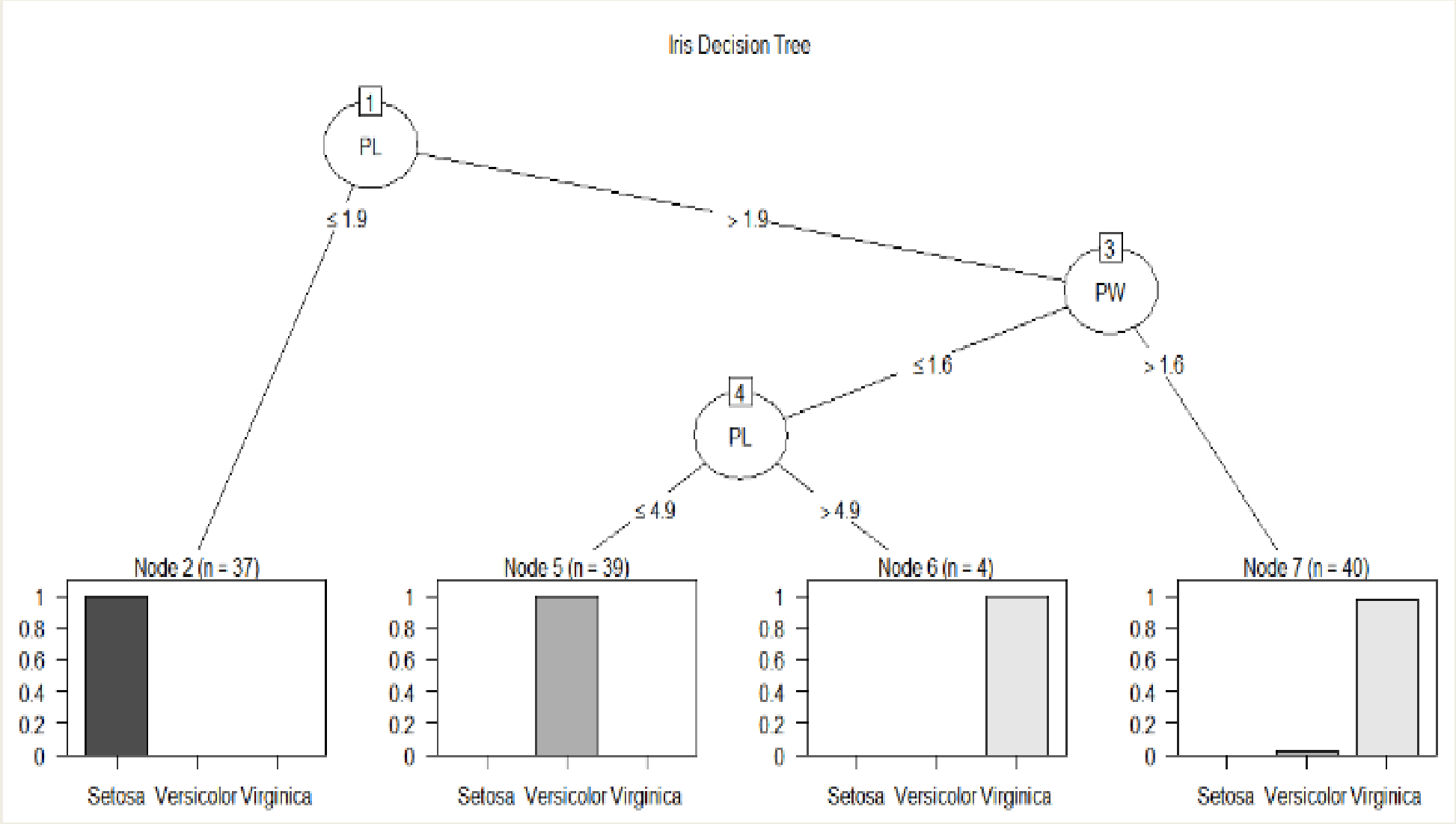
50 samples from each of 3 species (Setosa, Virginica, and Versicolor).

And each sample is explained by 4 numerical attributes: Sepal Length, Sepal Width, Petal Length and Petal Width.

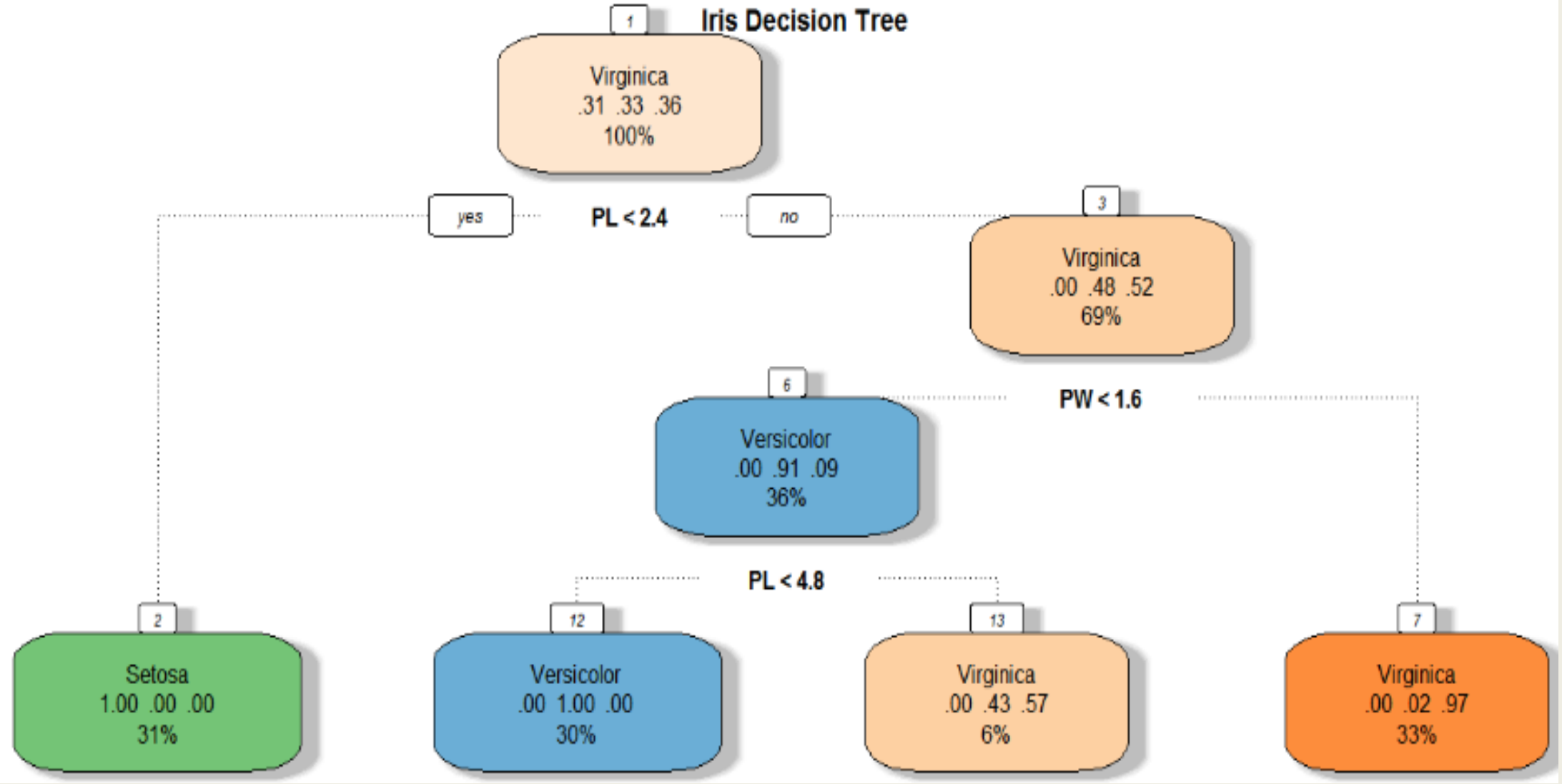
80% of the data using for training set and the remaining 20% for testing the tree model.

```
> head(Iris1)
   SL  SW  PL  PW Classification
1 5.0 3.5 1.3 0.3         Setosa
2 7.9 3.8 6.4 2.0      Virginica
3 5.8 2.8 5.1 2.4      Virginica
4 5.7 2.6 3.5 1.0     Versicolor
5 5.4 3.9 1.7 0.4         Setosa
6 5.2 2.7 3.9 1.4     Versicolor
```

C5.0 Algorithm Classification Decision Trees For Iris Dataset



CART Algorithm's Decision Tree



Generalization Capacity of the Trees

Iris	Predicted		
	Setosa	Versicolor	Virginica
Setosa	13	0	0
Versicolor	0	8	2
Virginica	0	0	7

C5.0 percentage of accuracy: 93.33%

CART_Predicted			
	Setosa	Versicolor	Virginica
Setosa	13	0	0
Versicolor	0	8	2
Virginica	0	0	7

CART percentage of accuracy: 93.33%

Comparative Study of CART and C5.0 using Titanic Dataset

- **Data Description:**
- The Titanic dataset describes the survival status of individual passengers on the Titanic. The dataset frame contains 1309 instances on the following 14 variables:

VARIABLE DESCRIPTIONS

Pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survival	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare (British pound)
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

Add Some Conversions and Modifications to the Dataset

A	B	C	D	E
Field	**Modification**		**Field**	**Modification**
Name	Ignored		Pclass	Normalized into 1st, 2nd and 3rd
Passenger ID	Ignored		Sibsp	Normalized into None, Lessthan3, Morethan3
ticket	Ignored		Parch	Normalized into None, One, and Both
fare	Ignored		Age	Normalized into Child, Adolescent, Adult and Old
cabin	Ignored		Embark	Normalized into S.amp, Cher, and Queen
home.dest	Ignored		Survived	Normalized into YES and NO
boat	Ignored			
body	Ignored			

A glimpse of New Titanic Dataset

	A	B	C	D	E	F	G
1	**sex**	**Age**	**embarked**	**pclass**	**sibsp**	**parch**	**Survived**
2	female	Adult	Southampthon	2nd	None	One	YES
3	male	Old	Cherbourg	1st	None	None	NO
4	male	Adult	Southampthon	3rd	Lessthan3	Both	NO
5	male	Adult	Southampthon	1st	None	None	YES
6	female	Adult	Southampthon	1st	None	Both	YES
7	female	Adult	Cherbourg	1st	Lessthan3	None	YES
8	male	Adult	Queenstown	3rd	None	None	YES

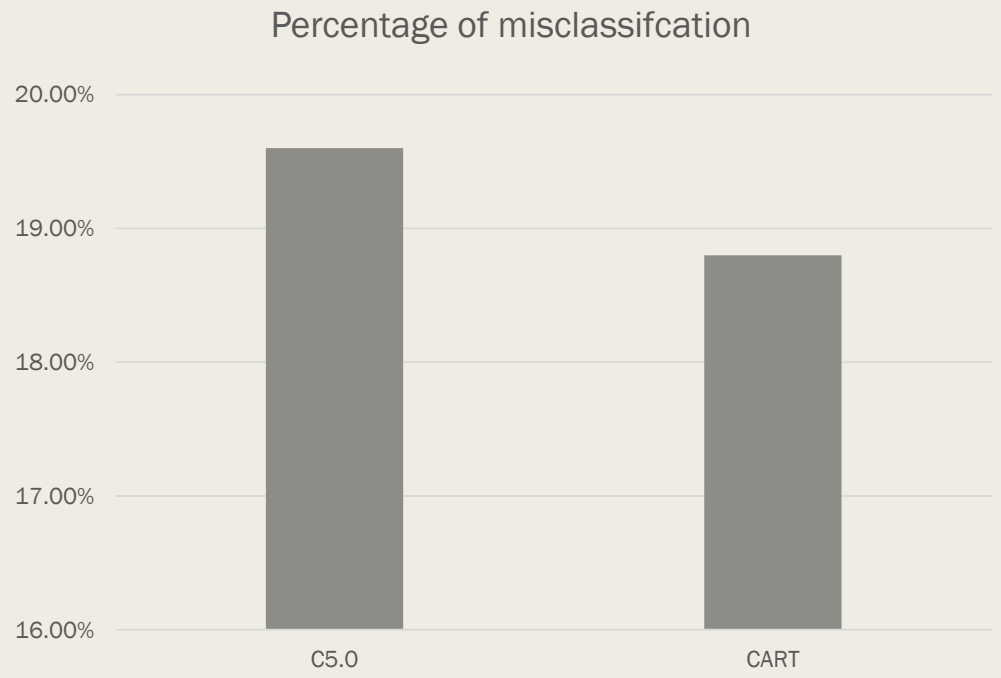
Rulesets & Findings

```
Decision tree:

sex = male:
:...Age in {Adolescent,Adult,Old}: NO (370/70)
:   Age = Child: YES (21/6)
sex = female:
:...pclass in {1st,2nd}: YES (143/12)
    pclass = 3rd:
    :...embarked = Cherbourg: YES (13/1)
        embarked in {Queenstown,Southampthon}:
        :...sibsp in {Lessthan3,Morethan3}: NO (32/11)
            sibsp = None: YES (49/23)
```

CART has a lower probability of misclassification than C5.0

Comparsion of Classifcation capacity		
C5.0 / CART	NO	YES
NO	321 / 330	42 / 33
YES	81 / 85	184 /180



Same predictive accuracy percentage

The Predictive Accuracy of Decision Trees Using the Test Set.

C5Opredicted

	NO	YES
NO	221	35
YES	48	114

C5.0 predictive accuracy percentage: 80.01%

CARTpredicted

	NO	YES
NO	224	32
YES	52	110

CART's predictive accuracy percentage: 79.9%

Comparative Study C5.0 and CART using Diabetes Data

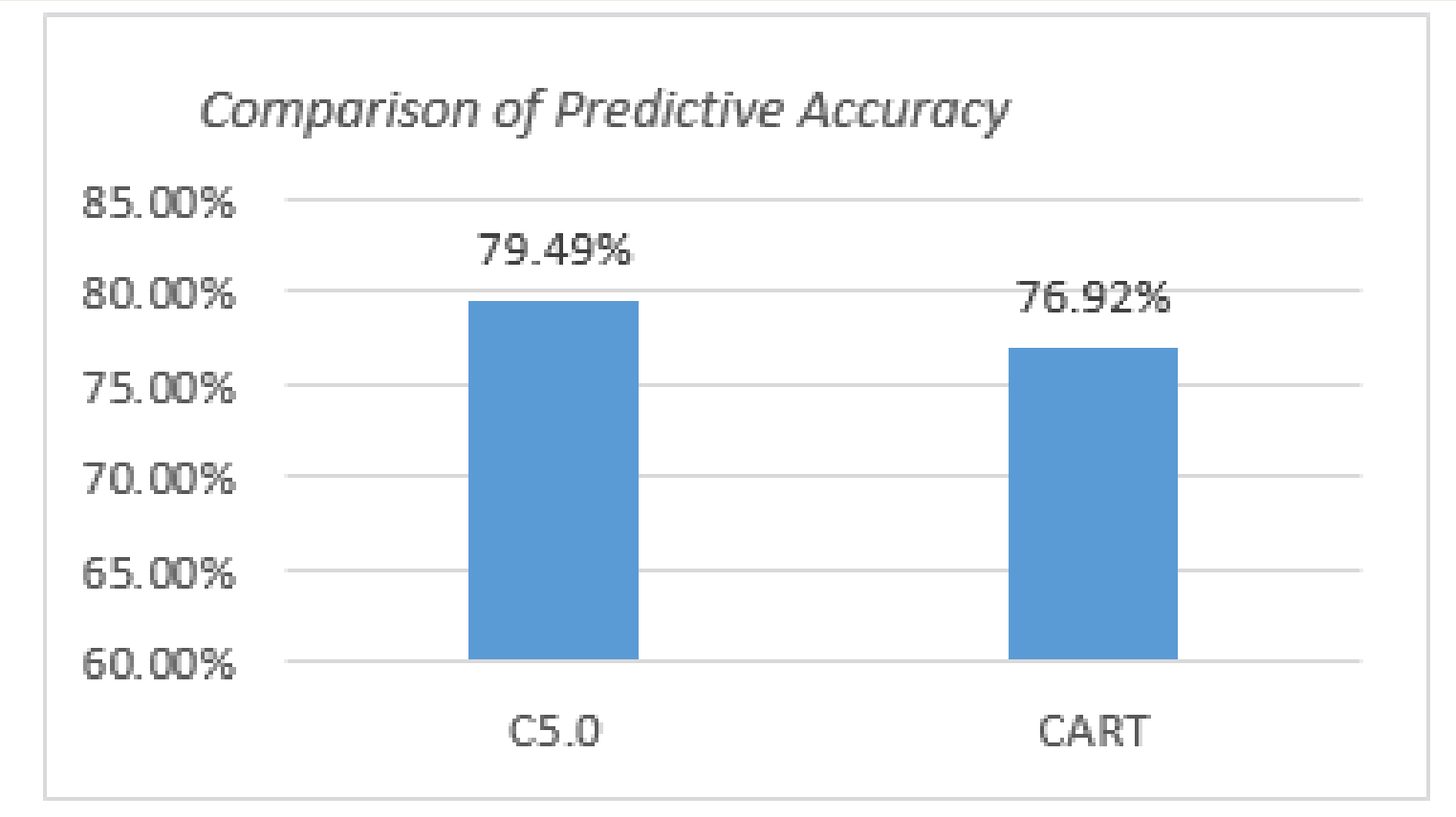
- **Data Description:**

A total of 768 instances in Prima Indians Diabetes Database described by the 9 following attributes: number of times pregnant, Plasma glucose concentration, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), Serum insulin (mu U/ml), BMI, Diabetes pedigree function, Age (years), Class variable (Sick or Healthy).

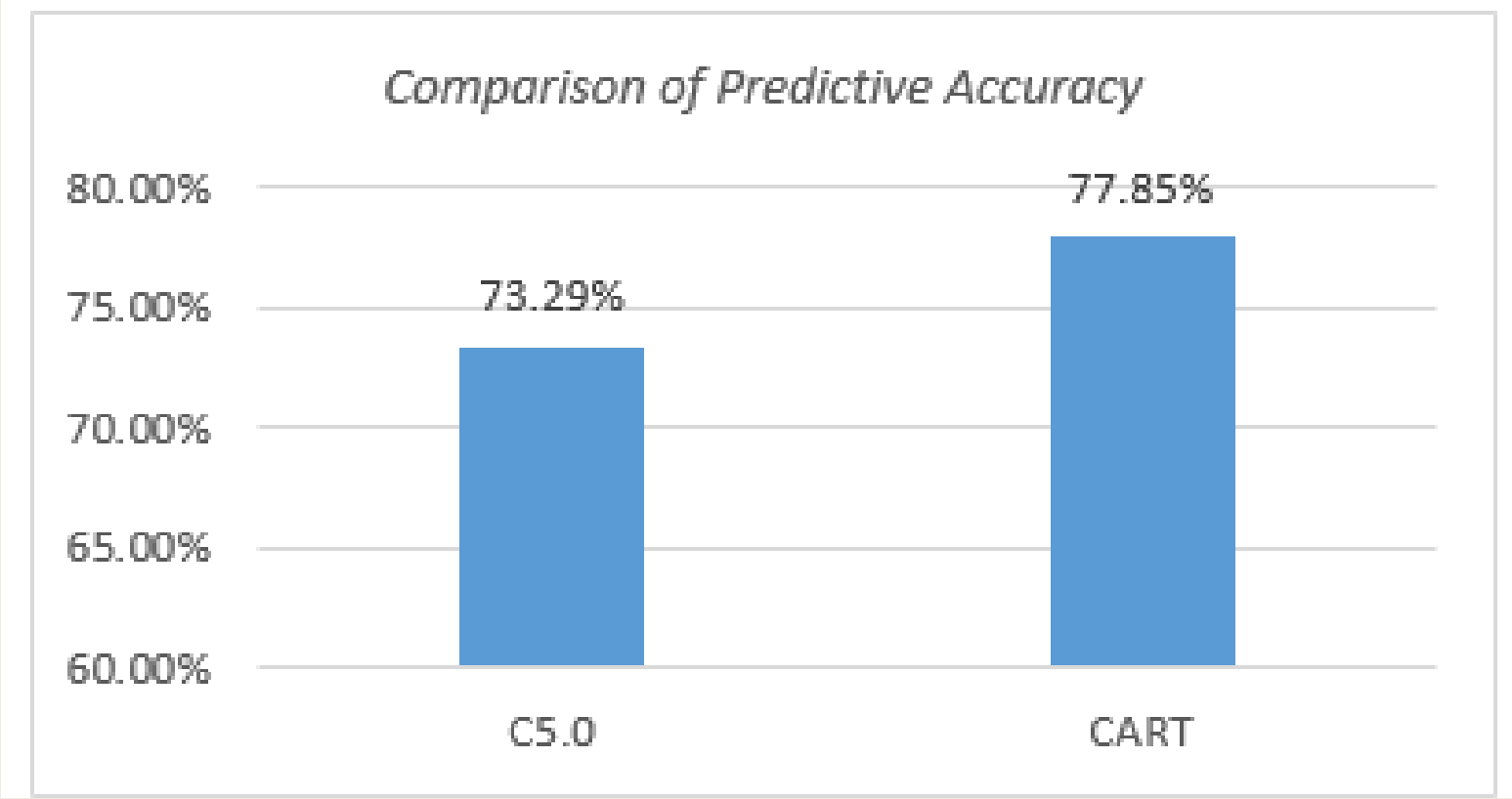
Roughly 49% of the dataset contains missing values.

Two options: Discard the missing values or Include them.

Scenario 1: Discard the Missing Values



Scenario 2: Missing Values Included



Summary and Conclusions

1) C5.0 can have a multiway splitting or binary decision tree, whereas CART only gives a binary tree.
2) C5.0 use Information Gain or Entropy as an attribute selection measure to build a decision tree while CART use Gini index.
3) For the pruning process, CART uses pre-pruning technique called Cost – Complexity pruning to remove redundant braches from the decision tree to improve the accuracy, whereas C5.0 pruning technique adopts the Binomial Confidence Limit method to reduce the size of the tree without any loss of its predictive accuracy.
4) Finally, in a problem of handling missing values CART surrogates test to approximate outcomes while C5.0 apportions values probability among outcomes.

Q&A section

- Thank you