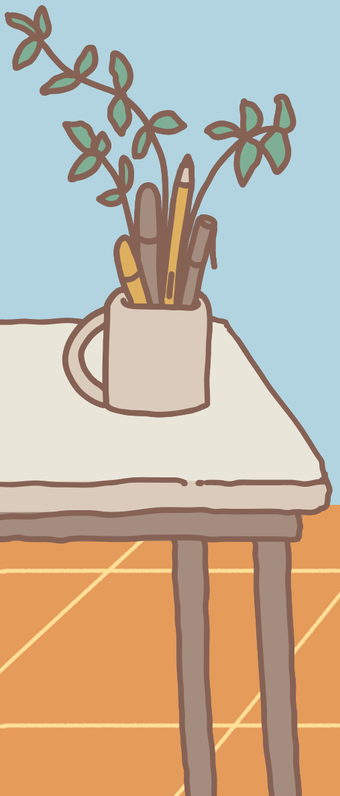
Attribute Selection(CHAID)

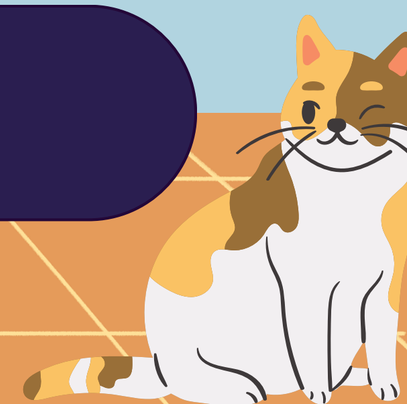
Lecture: Mr. Chan Sophal

Members:

SEAN VENGNGY e20201133
Thornthea Gechhai e20201321
Aov keatmeng e20201812
SOK SREYSEEY e20201226
Men chanchhorporn e20201146
HIN BUNRA e20201287
SANG Rithpork e20200232
TENG CHANSOPANHA e20201711
RUN SAVIN e20200897
HONG KIMMENG e20200559
Pean Chhinger e200201339

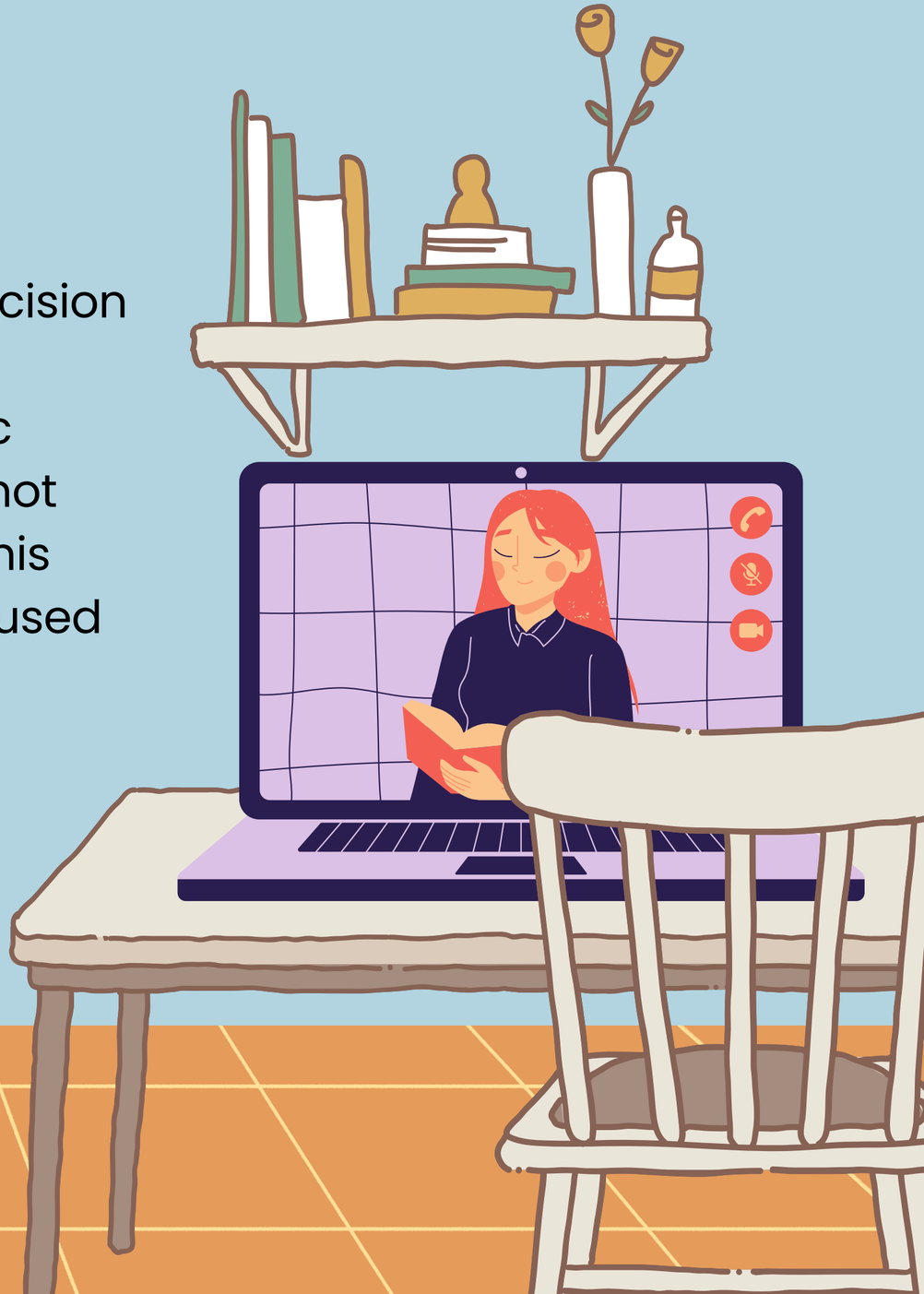


Overview



1. **Introduction**

CHAID, or Chi-Squared Automatic Interaction Detector, is a decision tree algorithm that is commonly used in Machine Learning for classification and regression tasks. CHAID is a non-parametric method for building decision trees, which means that it does not make any assumptions about the distribution of the data. In this slide, we will explore what CHAID is, how it works, and how it is used in Machine Learning.



2. What is CHAID

CHAID is a decision tree algorithm that is based on the chi-squared test of independence.

The CHAID algorithm works by recursively splitting the data into subsets based on the categorical predictor variables that have the strongest association with the response variable. At each step, CHAID calculates the chi-squared test of independence between the response variable and each of the categorical predictor variables. The variable with the strongest association is chosen as the split variable, and the data is divided into subsets based on the categories of that variable. This process is repeated for each subset until the stopping criteria are measured

3. How does CHAID work

CHAID works by recursively partitioning the data into subsets based on the predictor variables that have the strongest association with the response variable. The algorithm starts with the entire data set and then splits it into subsets based on the predictor variable that has the strongest association with the response variable. This process is repeated for each subset until the stopping criteria are met.

Step 1: Start with complete data

Step 2: Statistical significance of each variable¶

Step 3: Selecting the best variable to split based on least p-value

Step 4: Repeting steps 1,2 and 3 until the stopping criterion¶

4. Program for CHAID

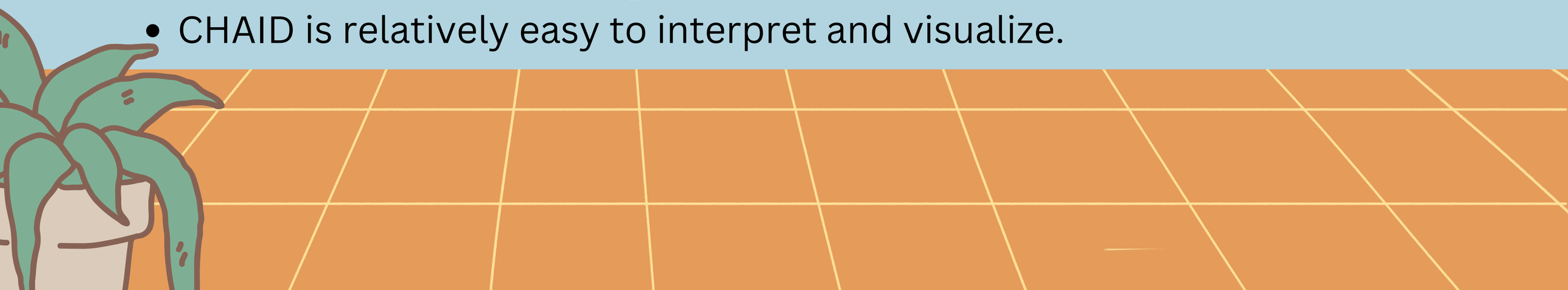
```
import pandas as pd
from CHAID import Tree

#Load dataset
df = pd.read_csv('dataset.csv')

#Split dataset into predictor and target variables
X = df.iloc[:, :-1]
y = df.iloc[:, -1]

#Create a decision tree using CHAID algorithm
tree = Tree.from_df(df, 'target_variable', max_depth=3)

#Print the decision tree
tree.print_tree()
```

- In this code, we first load a dataset as a Pandas DataFrame. We then split the dataset into predictor variables X and the target variable y.
- Next, we create a decision tree using the Tree.from_df method from the CHAID package. We pass in the DataFrame, the name of the target variable, and a maximum depth of the tree.
- Finally, we print the decision tree using the print_tree method.
- Note that you'll need to install the CHAID package first using pip or conda.

5.Advantages of CHAID

There are several advantages to using CHAID as a decision tree algorithm:

- CHAID is a non-parametric method, which means that it does not make any assumptions about the distribution of the data.
- CHAID is a powerful tool for exploring the relationships between categorical variables.
- CHAID can handle both categorical and continuous variables.
- CHAID can handle missing data and is robust to outliers.
- CHAID is relatively easy to interpret and visualize.



6.Application of CHAID

1. Marketing: CHAID can be used to identify the characteristics of customers who are most likely to purchase a particular product or service.
2. Medical Research: CHAID can be used to identify risk factors for certain diseases or conditions.
3. Social Science: Chaid is used to identify the factors that predict a particular behaviour or outcome.

Favorite Ideas

7. Limitations of CHAID

There are also several limitations to using CHAID as a decision tree algorithm:

1. CHAID is a greedy algorithm, which means that it may not always find the optimal tree.
2. CHAID can be sensitive to small changes in the data.
3. CHAID is not suitable for large data sets, as it can become computationally intensive.
4. CHAID can produce complex trees that are difficult to interpret.


Thank You
for your participation
love