STEAMCITY CURRICULUM

TREES VS. CARS
Identifying Polluting Vehicles

thematic: artificial intelligence and new technologies

sub-theme #1: sustainable mobility, transport and regulation

sub-theme #2: environment, well-being and public health

Introduction

The protocol "**Trees vs. Cars**" aims to provide students with an accessible introduction to **supervised learning**, a core concept in machine learning and artificial intelligence. Through a hands-on classroom activity, students will develop their understanding of how computers can learn to make decisions based on patterns in data.

Behind this general objective, students will explore how machines can **classify information** by **learning from examples** rather than following explicitly programmed rules. This approach allows computers to **adapt to new situations and make predictions** based on previous experiences - much like human learning, but through mathematical and statistical processes.

Thanks to the protocol steps, students will create **a model capable of categorizing vehicles into two groups**: those **authorized to drive in the Brussels LEZ and those that are not**, based on three criteria: vehicle type, fuel type, and year of manufacture.

This specific application - **classifying vehicles for Brussels' Low Emission Zone (LEZ)** - represents a **real-world implementation of existing AI in urban management systems**, demonstrating how cities use smart systems to regulate traffic and reduce pollution. The decision tree technology that students will explore is deployed in various smart city applications, from **optimizing public transportation routes to predicting maintenance needs for urban infrastructure**.

Interdisciplinarity

Technology and engineering
Biology
Physics & Chemistry

Sustainable Development Goals

7 AFFORDABLE AND CLEAN ENERGY
11 SUSTAINABLE CITIES AND COMMUNITIES
13 CLIMATE ACTION

Overview

Protocol Structure

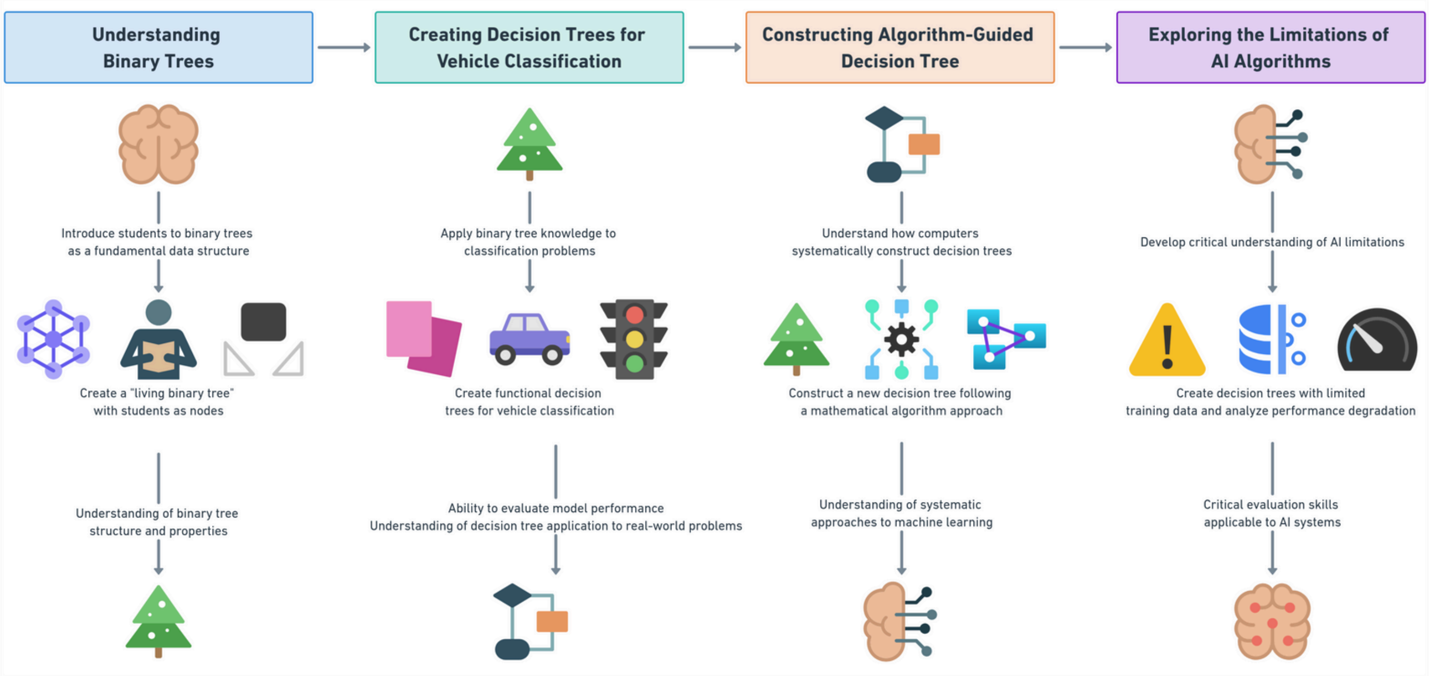
In the "**Trees vs. Cars**" protocol, students will follow a 4-step methodology structured as follow:

Step 1: Understanding Binary Trees
In the first step, students will discover the concept of binary trees, a data structure widely used in the field of supervised learning.

Step 2: Creating Decision Trees for Vehicle Classification
In the second step, they will construct a decision tree, a particular form of binary tree, with the aim of separating a set of vehicles into two categories: the one authorized to drive in the Brussels LEZ, and the others. They will compare the trees created and their performance in categorizing the vehicles.

Step 3: Algorithm-Guided Decision Tree Construction
Based on the knowledge acquired, the students will then construct a new decision tree, this time following an algorithm provided to them. They will then compare the trees created and their performance with each other and with the decision trees created previously.

Step 4: Exploring Limitations of AI Algorithms
In order to open a larger discussion regarding AI systems and conclude on the protocol, the last step will engage students in discovering some of the main limitations of this type of algorithm, showing them that AI is not a magic tool capable of solving any problem, but rather a very effective means of solving certain kind of problems.



Getting started

Duration: 4 hours

Level of difficulty:

★★★☆☆

Material needed: All the cards and sheet files are available for printing in the appendixes.

- 1 "Number" cards deck
- 1 "Vehicles dataset" card deck per group
- 1 Algorithm sheet per group

Glossary

Keywords & Concepts	Definitions
Binary Tree	A binary tree is a data structure in which each element (called a node) has a maximum of two child nodes and at most one parent node. The nodes that have no children are called the 'leaves' of the tree.
Machine Learning	Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions.
Supervised learning	Supervised learning is a Machine Learning technique that uses labeled datasets to train artificial intelligence algorithm models to identify the underlying patterns and relationships between input features and outputs.
Model	A model is the result of the training process. It is what enables predictions to be made on new data.
Generalisation	Generalisation is the ability of the model to predict correctly on new data, not used during training.
Model performance	Model performance is the measure of a model's ability to make correct predictions.

Protocol

Step 1 - Understanding Binary Trees

Background and description of the problem to be solved in this step: Binary trees are a ubiquitous data structure in computer science. It has many interesting properties, such as efficiency for searching, inserting and deleting elements inside, the ability to represent mathematical expressions for syntactic analysis in compilers. They are also used in data compression algorithms, such as the Huffman Coding. Understanding the concept of a binary tree is essential in this protocol, as the decision tree of the following steps is a particular form of binary tree.

Learning Objectives: Understanding the concept of binary trees.

Conceptualisation

To understand **how a computer can make a decision**, we need to know **how it organizes input data**. We can start by looking at **how humans have structured and organized large quantities of information**, and the **methods** used to retrieve it. The simplest example is **dictionaries**. A dictionary is a book containing the definitions for words organized alphabetically. This organization makes it easy to find a definition. You do not need to go through the whole dictionary. By identifying the first letter of the word you're looking for, you can directly open a rough page where it might be found. Depending on the words present on this page, it can be determined whether the search should continue before or after it, thus eliminating a large number of pages. This process is then repeated until the desired word is found.

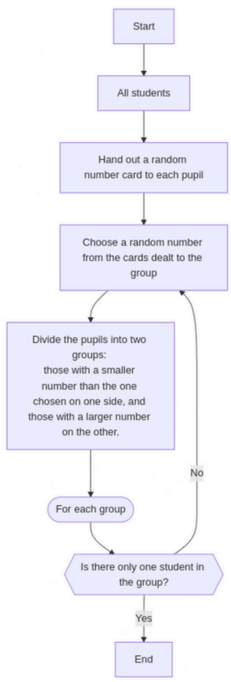
This method is known as **dichotomous searching**. Another example of this approach is the game of guessing a number between 1 and 100, where the best strategy is to divide the field of possibilities in two equal parts, then eliminate one of the two halves. The structure used behind dichotomous searching is a **binary tree**. In this step, students will discover how they already use it.

Students Investigation

The students will discover the concept of a binary tree using binary search trees. The aim is to **create a living binary tree**, with **each student representing a node**.

To begin with, **each pupil is given a card randomly numbered between 1 and the number of pupils** (you can use the number cards in the appendix, or simply use numbered pieces of paper).

The teacher chooses a student randomly to become the **root of the tree**. This student announces the received number, and the rest of the class gets divided in two: those with a **lower number place themselves on the left**, the others **on the right**. This process is repeated until all the students have been placed, forming a complete tree where each 'leaf' is a single student. Here aside is an outline of the process to be followed.



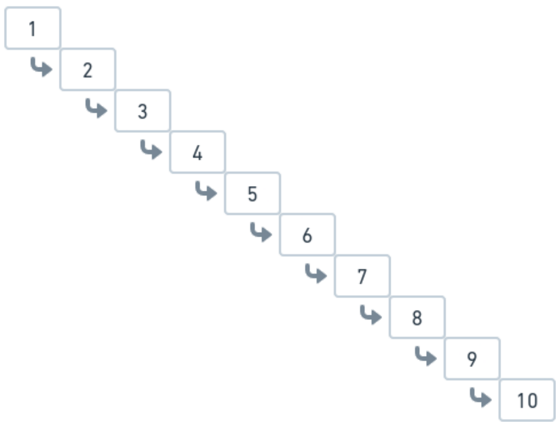
To give a good idea of how the data is organized in a binary tree, the teacher chooses **one student randomly from the leaves of the tree**. The class then has to go through the tree to identify **the number of each student**. The students must understand that they have just created a binary tree, and more specifically a binary search tree, i.e. **a binary tree whose special feature is that it is very efficient for searching inside data.**

The students will then learn **how to balance the tree they have created**. To do this, the teacher will divide the class into groups and **distribute the number cards evenly between these groups** (around ten cards by group). Students will have to build a binary tree, as they did before, but this time they try to create **the most efficient tree**.

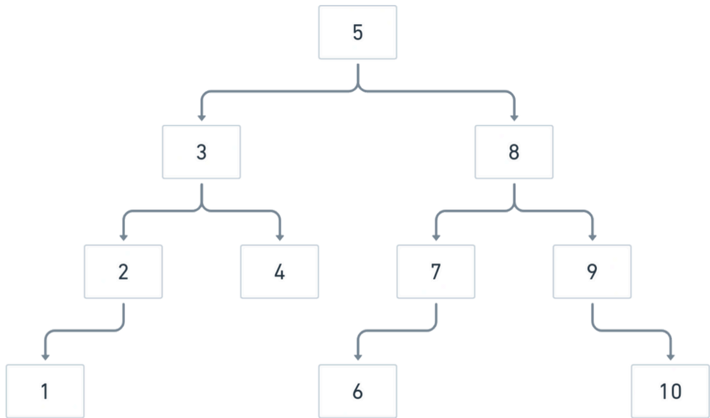
A tree is more efficient if we need **less try to find a specific node**. To create a most efficient tree, we need to try and find the tree with the **shortest branches possible**. If the students have difficulty grasping the concept, the teacher can ask the groups having difficulties to create the **least efficient tree first**, i.e. the tree with the longest branches possible. By understanding how to build the least efficient tree, students should **better understand how to find a better one**.

The least efficient tree is one that contains **only one branch**, where the numbers are arranged in order, as in the example. This tree is said to be **unbalanced.**

On the other hand, the most efficient tree is **completely balanced**, i.e. all the branches of the tree **have the same length** (plus or minus 1).



unbalanced least efficient tree



balanced most efficient tree

Conclusion & Further Reflexion

- **Knowledge Mobilized**: Students have explored the concept of binary trees through a hands-on activity. They have understood the tree's structure and its efficiency in searching for data by experiencing the creation of both efficient and inefficient trees.
- **Classroom Implementation Reflection**: Students have actively participated in building a binary search tree by positioning themselves in the classroom based on numerical order. Through guided reflection, they have understood the concept of more efficient trees and the importance of balancing branches to minimize search depth.
- **General Learning Outcomes**: Students have gained a foundational understanding of binary trees and binary search trees. They have learned to appreciate the significance of balanced trees for efficient data searching and develop problem-solving skills for optimizing tree structures.

Step 2 - Creating Decision Trees for Vehicle Classification

Background and description of the problem to be solved in this step: After discovering the concept of a binary tree, the students will see to what extent binary trees can be used for machine learning, in particular data classification.

Learning Objectives: Understanding what a decision tree is. Learn how to use a decision tree to classify data and make predictions. Learn how to measure the performance of a model

Conceptualisation

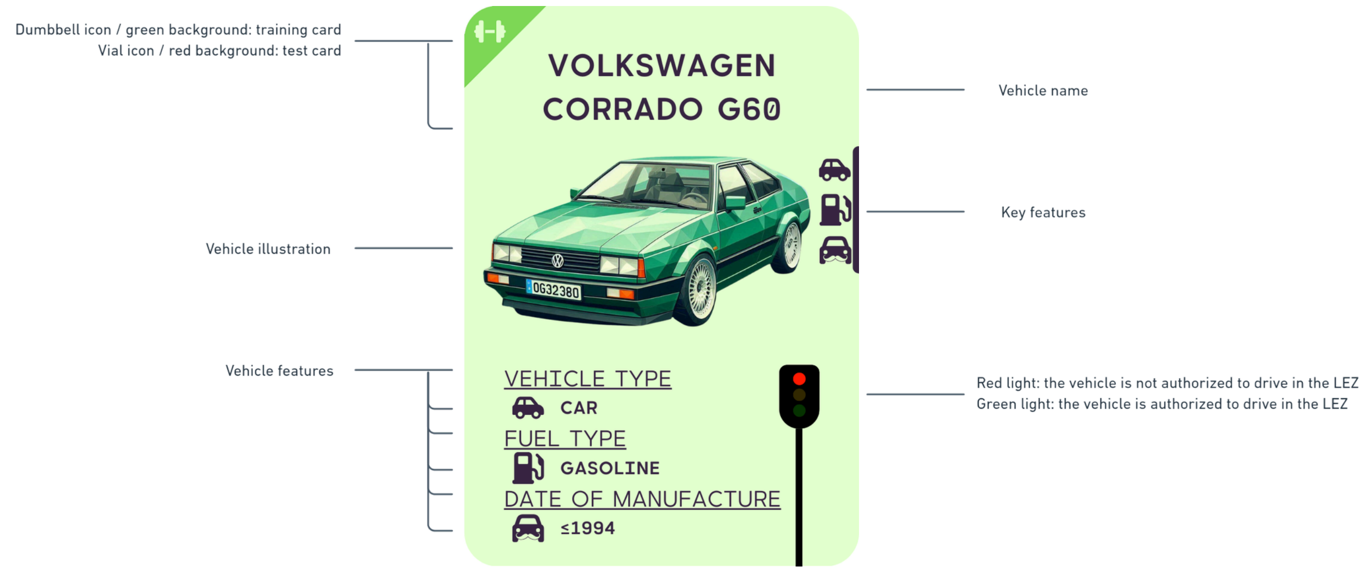
Once the binary tree concept has been mastered, the data structure created and studied in step 1 can be **enriched** and made **more complex** to be used to **support decision maker**. This structure is called a **decision tree** and instead of storing numbers, each node is a **close-ended question**. Each student of a node represents a possible answer (generally “yes” or “no”). Leafs symbolizes a result of the decision process.

Like binary trees, which can be used to quickly eliminate many elements from a search, a decision tree avoids **unnecessary questions and speeds up the decision-making process**.

In this step of the protocol, students will apply their knowledge in buiding a more complex decision tree to determine which vehicle is allowed or not in a LEZ. In urban areas around the world, Low Emission Zones (LEZs) are becoming increasingly common as cities strive to reduce air pollution and improve public health. These zones restrict access to vehicles based on their emissions levels, ensuring that only those that meet certain environmental standards can enter. Understanding the criteria for vehicle entry is crucial for both city planners and vehicle owners. **How can we use a decision tree to determine whether a vehicle is allowed or not allowed to enter a LEZ?**

Students Investigation

During this step, students will have to familiarise themselves with a dataset, represented by the deck of vehicle cards available for printing in the appendix. Below is a description to understand each visual element:



For each vehicle, there are 3 characteristics, each with 3 possible values:

Vehicle Type

- Car
- Utility
- Motorcycle

Fuel Type

- Gasoline
- diesel
- Electric

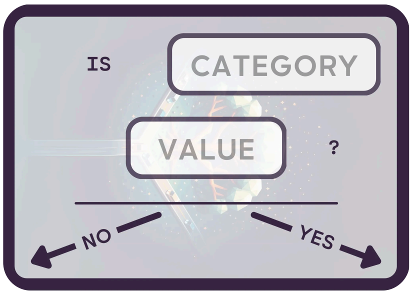
Date Of Manufacture

- ≤ 1994
- 1995 - 2006
- 2006

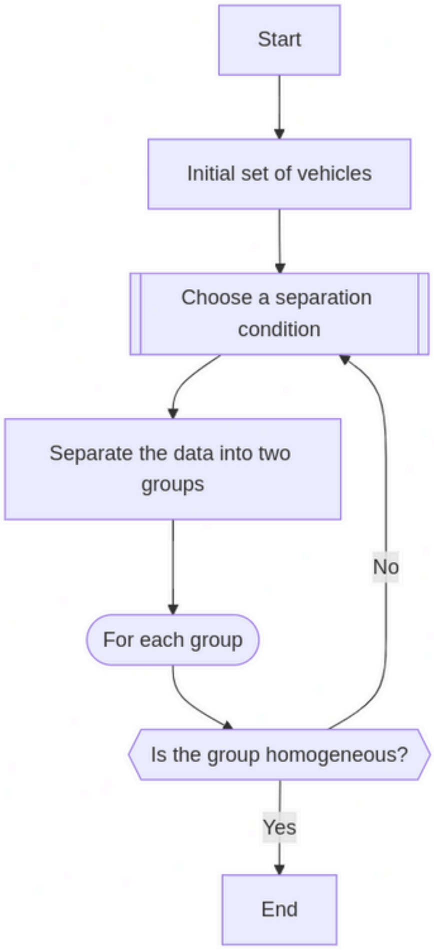
Each vehicle also has a **label**, represented by the **traffic light** at the bottom right of the card.

Once they have familiarised themselves with the cards, the students will work in groups to build a decision tree using **only the training cards** (the test cards will be used later). To do so, the students will have to carry out the following steps for each node in the decision tree:

- **Choose a separation condition** in the form: 'Is [category] [value]? For example: 'Is the fuel type diesel?' and record it on **one of the question cards** present with the dataset cards:



- **Divide the vehicle cards into two groups** according to whether each vehicle **answers this condition true or false**, and divide them up on either side of the question card.
- **Continue this process** for each new group formed, **choosing new conditions**
- **Stop** when a group contains **only vehicles authorised in the LEZ** (green light) or **only unauthorised vehicles** (red light)



Here aside is a flow chart to help you understand the steps involved.

To evaluate the performance of their model, the students will use **test cards**. They will 'move' each test card **down their decision tree**, following the conditions at each node until they reach a leaf. **If the label on the card matches the label on the leaf reached**, the prediction is considered **correct**. The students count the number of correct predictions, then divide this number by the total number of cards tested. The result is a percentage and indicate the **accuracy of the model**. The closer the score is to **100%,** the better is the model. This method allows students to objectively assess their tree's ability to decide correctly on new data.

Finally, the groups will compare their trees and their performance. It may be interesting to ask the groups what method or strategy they used to choose the data separation condition at each node.

Conclusion & Further Reflexion

- **Knowledge Mobilized**: Students have explored the concept of binary trees and their application in machine learning, particularly in data classification. They have learned the fundamentals of decision trees, how to structure them, and methods to evaluate their predictive performance.
- **Classroom Implementation Reflection**: Through hands-on activities, students have built decision trees using a dataset of vehicle cards, defining separation conditions and iteratively organizing data. Reflection has involved testing their models with unseen data, calculating accuracy, and comparing strategies and results with peers.
- **General Learning Outcomes**: Participants have gained an understanding of decision trees as decision support tools, developing critical thinking to structure data effectively. They have enhanced their data science skills by measuring model accuracy and learning to assess the impact of their design choices on performance.

Step 3 - Constructing Algorithm-Guided Decision Tree

Background and description of the problem to be solved in this step: In the previous step, the students created a decision tree. To do this, they chose which separation condition to use for each branch. However, a computer is not capable of making arbitrary decisions, and must instead rely on a statistical and mathematical tools to analyze data. The students will therefore create a new decision tree, but this time using a method to determine the best separation condition for each branch of the tree.

Learning Objectives: Understanding how a computer build-up a good decision tree. See which method is the most effective for choosing the separation condition

Conceptualisation

In the previous step, participants explored how to define separation criteria to determine, at each stage, whether or not a vehicle is authorized to enter the LEZ. Depending on the group, different questions were asked and different approaches adopted. Although the end result may be the same, it is essential to ask whether there is a solution (or algorithm) capable of identifying the best criteria for reaching a decision reliably and quickly.

Students Investigation

During this phase, the students will construct a new decision tree as in the previous phase, but following a **very precise method for determining the condition to be chosen at each node**.

To guide them, students will have the following table at their disposal, which can either be laminated so that it can be used several times, or printed out in several copies (6-8). It can be used for determining the best separation condition for vehicles. The table is divided into categories (Type of vehicle, Type of fuel, Date of manufacture) with their respective values. To find the best separation condition, students can follow these steps:

1. Fill in the columns **A** (authorised vehicles) and **NA** (unauthorised vehicles) for each value in each category.
2. Identify the row(s) with the smallest value in either column **A** or column **NA**. Tick the corresponding box in the **ABSOLUTE MIN?** column for these rows.
3. From the row(s) where **ABSOLUTE MIN?** is ticked, identify the row(s) with the highest value of **A** or **NA**, and tick the **ABSOLUTE MAX?** box accordingly.
4. The optimal separation condition is the one corresponding to the row that has the boxes ticked in both **ABSOLUTE MIN?** and **ABSOLUTE MAX?**.

Category	Value	A	NA	Absolute min.?	Absolute max.?
Vehicle Type	Motorcycle			☐	☐
	Car			☐	☐
	Utility			☐	☐
Fuel Type	Diesel			☐	☐
	Gasoline			☐	☐
	Electric			☐	☐
Date of Manufacture	≤1994			☐	☐
	1995-2006			☐	☐
	≥2006			☐	☐

The rest of the process is the same as in the previous step. The result would be the following flowchart available here: https://bit.ly/binarytreeai. If the students carry out the algorithm correctly, they should all obtain the same decision tree. In the tests, they should also obtain the best performance.

Conclusion & Further Reflexion

- **Knowledge Mobilized:** Students have explored the process of decision-making by computers. They have learned to apply a structured algorithm to create consistent and efficient decision trees.
- **Classroom Implementation Reflection:** Through hands-on activities, students have followed a detailed algorithm to construct decision trees based on statistical analysis of dataset characteristics.
- **General Learning Outcomes:** Participants have gained an understanding of computational decision-making processes, developing skills to apply algorithms in data classification. They have enhanced their ability to analyze datasets methodically and recognize the importance of precision in machine learning workflows.

Step 4 - Exploring the Limitations of AI Algorithms

Background and description of the problem to be solved in this step: Students have learned how to build and use a decision tree. This step aims to show some of the inherent limitations of machine learning.

Learning Objectives: Discover some of the limitations of such algorithms.

Conceptualisation

In the context of this protocol, decision trees appear to be a particularly suitable tool, offering the possibility of quickly obtaining an answer. However, like any representation method, **they have limits and a scope of use beyond which their effectiveness diminishes**.

In this step, students analyze and explore the limits inherent in this structure. They also consider the quality of the data used to construct the tree, and the impact this data may have on the accuracy and reliability of the predictions produced by this structure.

Students Investigation

The students will first discover the main limitation applicable to all machine learning algorithms: the size of the training dataset. To do this, the students will create a new decision tree, but this time by removing half of the training cards, i.e. 6 cards. With the exception of a few potentially lucky groups, the students should obtain a tree that performs worse than the one created in step 3. The teacher will then ask the students why the tree obtained is less efficient, to make them realise that by removing data, the algorithm lacks the examples and cases needed to make relevant choices.

Next, the students will construct a final decision tree, this time by swapping the training cards and the test data. The students should then all end up with the same tree that makes prediction errors. As with the previous tree, the teacher will ask them why the tree performs less well than the one in step three. Here, unlike in the previous tree, there are even more training data than test cards! The students need to understand that the quantity of data is not everything. Indeed, if the training data is not representative of all the possible data, the training algorithm could give more or less importance to certain criteria than is actually the case in all the possible data. For this activity, the training data has been carefully selected to maximise the performance of the model, but it is quite rare for a model to make correct predictions in 100% of cases. It is possible to approach this value, but very difficult, if not impossible, to reach it.

Conclusion & Further Reflexion

- **Knowledge Mobilized**: Students have engaged with the fundamental limitations of machine learning algorithms, focusing on two key aspects: insufficient training data and the representativeness of datasets. They have developed critical thinking about why decision trees may fail under certain conditions.
- **Classroom Implementation Reflection**: Through experiments, students have created and evaluated decision trees with reduced training data and unrepresentative datasets.
- **General Learning Outcomes**: Participants have learned to identify and explain the limitations of decision trees and machine learning algorithms in general. They have gained insight into the importance of balanced and representative datasets, enhancing their ability to critically assess the reliability of predictive models.

Bibliography

Books and Guides

- **"Machine Learning: An Algorithmic Perspective" by Marsland, S.**
  Presents the algorithms behind decision trees, with a practical focus.

- **"Data Mining: Practical Machine Learning Tools and Techniques" by Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J.**
  A practical book that explains how to implement decision trees using modern tools.

- **"Data Mining with Decision Trees: Theory and Applications" by Rokach, L., & Maimon, O.**
  A specialized book on decision trees, covering theoretical foundations and practical applications.

On-line resources

- **Mermaid** (https://mermaid.live)
  Mermaid is an online tool that allows users to create diagrams and flowcharts using a simple text-based syntax. It's particularly useful for developers and project managers who want to visually present complex processes in a clear and efficient way. With the ability to integrate into various platforms, Mermaid is a versatile resource for enhancing presentations and documentation.

- **Chapter 3.2 BINARY SEARCH TREES, Algorithme by ROBERT SEDGEWICK, KEVIN WAYNE.**
  https://algs4.cs.princeton.edu/lectures/keynote/32BinarySearchTrees.pdf

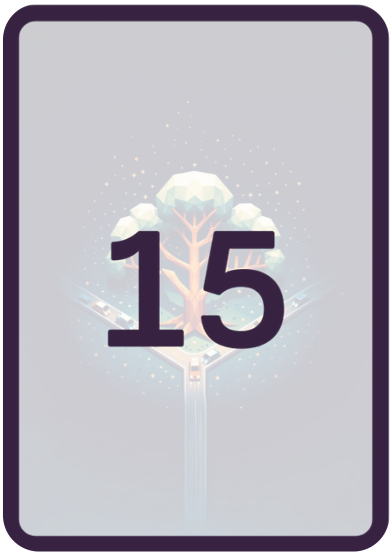
“Number” cards deck


10

11

12

13

14

15

16

17

18

19

20

21

22

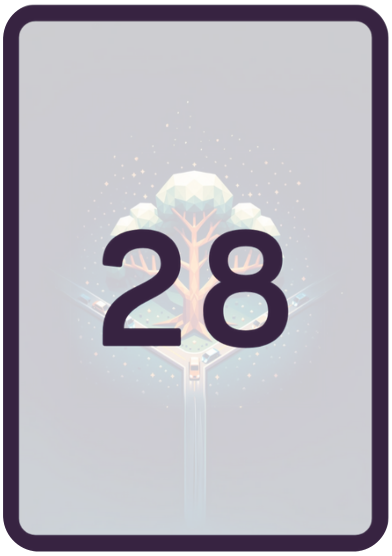
23

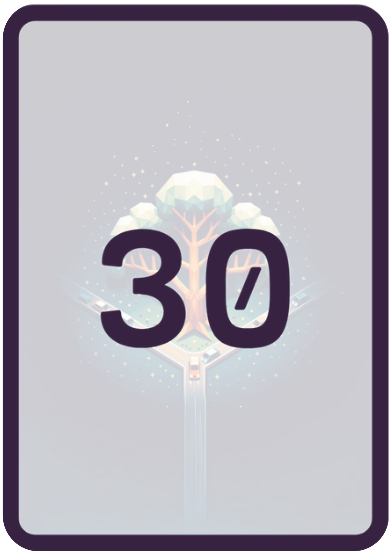
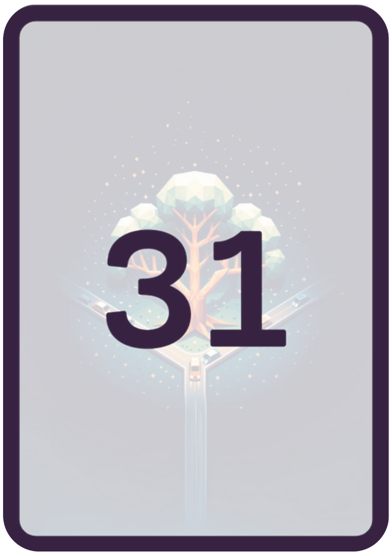
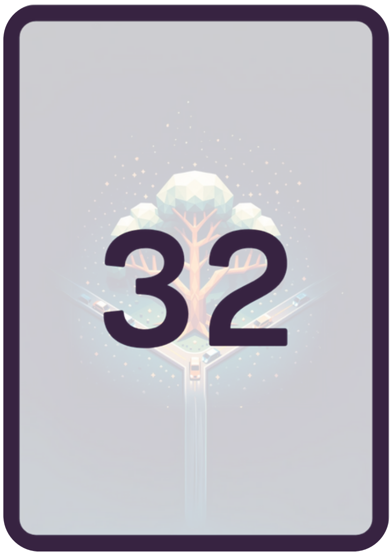
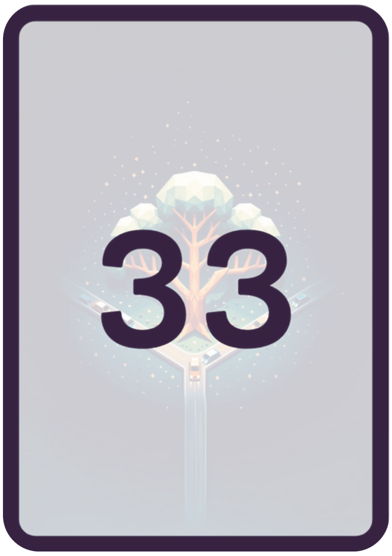
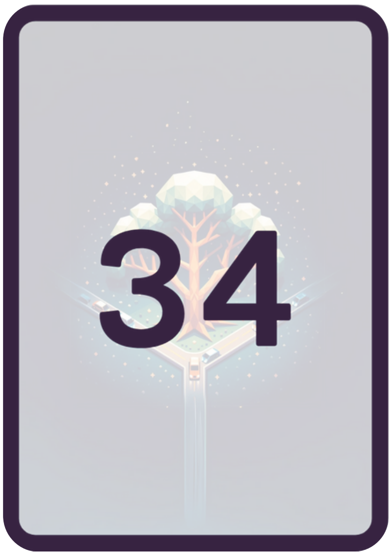
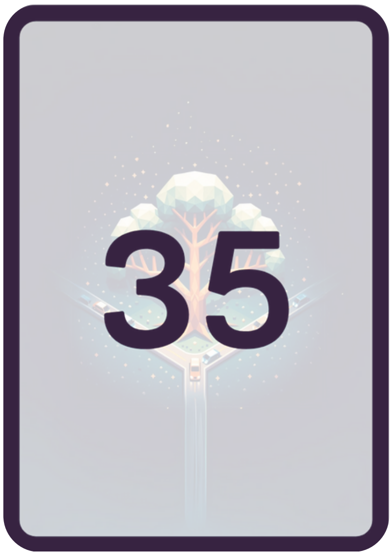
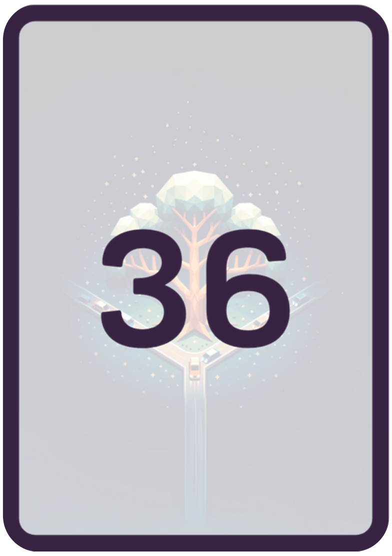
24

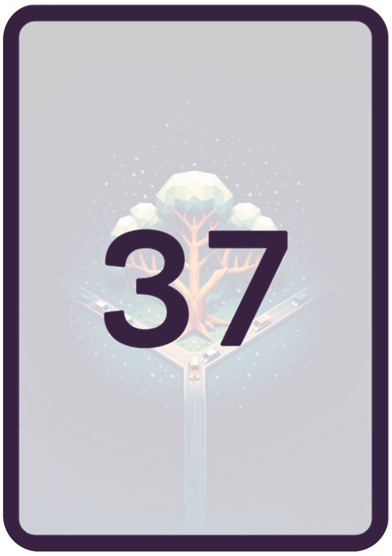
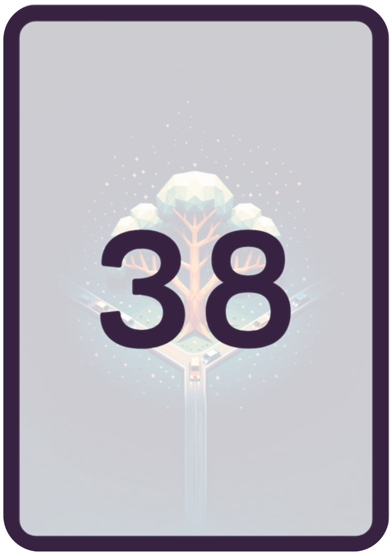
25

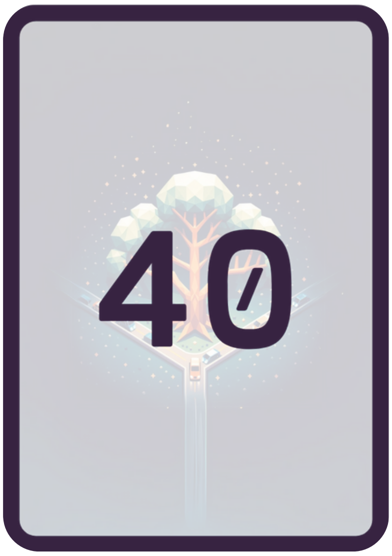
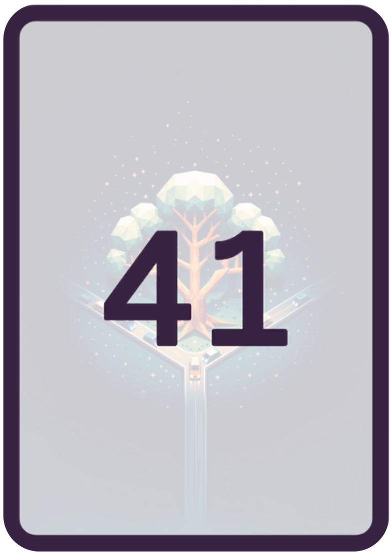
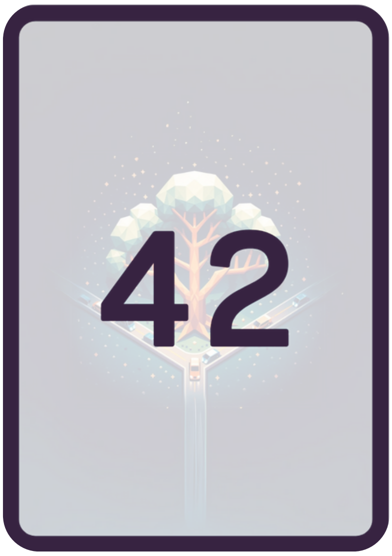
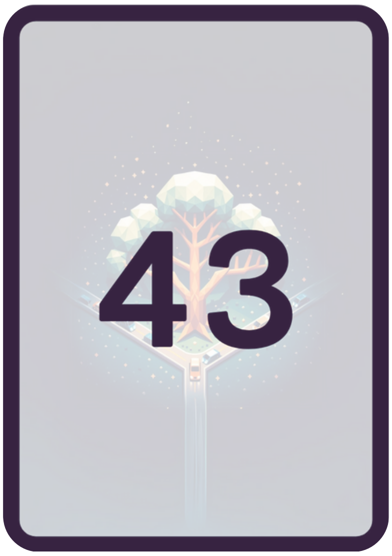
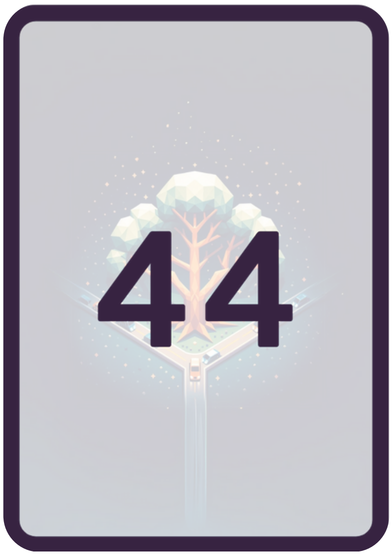
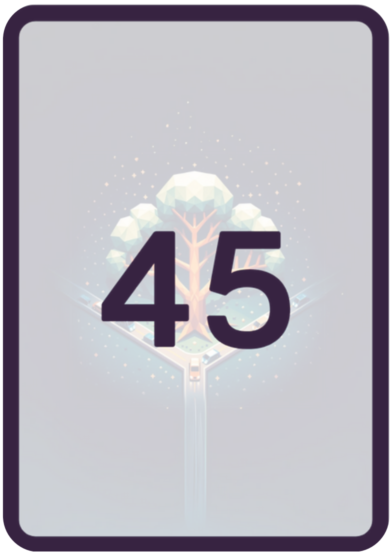
26

27


28

29

30

31

32

33

34

35

36


37

38

39

40

41

42

43

44

45

46

47

48

49

50

“Vehicles dataset” card deck

AURANTHETIC CHARGER

VEHICLE TYPE
MOTORCYCLE
FUEL TYPE
ELECTRIC
DATE OF MANUFACTURE
≤1994

CITROËN C-ZERO

VEHICLE TYPE
CAR
FUEL TYPE
ELECTRIC
DATE OF MANUFACTURE
>2006

DUCATI 996 2000

VEHICLE TYPE
MOTORCYCLE
FUEL TYPE
GASOLINE
DATE OF MANUFACTURE
1995-2006

FIAT DOBLÒ

VEHICLE TYPE
UTILITY
FUEL TYPE
GASOLINE
DATE OF MANUFACTURE
>2006

FORD RANGER EV

VEHICLE TYPE
UTILITY
FUEL TYPE
ELECTRIC
DATE OF MANUFACTURE
1995-2006

FORD TRANSIT

VEHICLE TYPE
UTILITY
FUEL TYPE
DIESEL
DATE OF MANUFACTURE
≤1994

HARLEY DAVIDSON SPORTSTER

VEHICLE TYPE
MOTORCYCLE
FUEL TYPE
GASOLINE
DATE OF MANUFACTURE
≤1994

HONDA CBF 125 2010

VEHICLE TYPE
MOTORCYCLE
FUEL TYPE
GASOLINE
DATE OF MANUFACTURE
>2006

HONDA CIVIC VII

VEHICLE TYPE
CAR
FUEL TYPE
GASOLINE
DATE OF MANUFACTURE
1995-2006

MERCEDES-BENZ SPRINTER I

VEHICLE TYPE
UTILITY
FUEL TYPE
DIESEL
DATE OF MANUFACTURE
1995-2006

NISSAN E-NV200

VEHICLE TYPE
UTILITY
FUEL TYPE
ELECTRIC
DATE OF MANUFACTURE
>2006

PEUGEOT 106 ELECTRIC

VEHICLE TYPE
CAR
FUEL TYPE
ELECTRIC
DATE OF MANUFACTURE
1995-2006

PEUGEOT 604

VEHICLE TYPE
CAR
FUEL TYPE
DIESEL
DATE OF MANUFACTURE
≤1994

PEUGEOT 2008 II

VEHICLE TYPE
CAR
FUEL TYPE
GASOLINE
DATE OF MANUFACTURE
>2006

PEUGEOT J5 ELECTRIC

VEHICLE TYPE
UTILITY
FUEL TYPE
ELECTRIC
DATE OF MANUFACTURE
≤1994

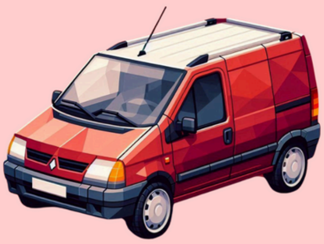
PORSCHE CAYENNE

VEHICLE TYPE
CAR
FUEL TYPE
DIESEL
DATE OF MANUFACTURE
>2006

RENAULT EXPRESS

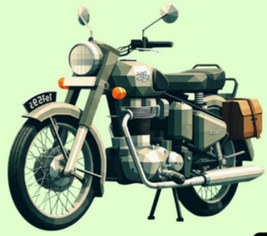
VEHICLE TYPE
UTILITY
FUEL TYPE
GASOLINE
DATE OF MANUFACTURE
≤1994

RENAULT KANGOO



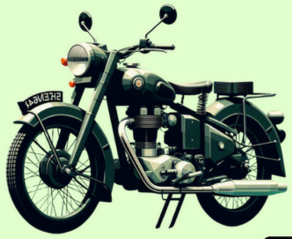
VEHICLE TYPE
UTILITY
FUEL TYPE
GASOLINE
DATE OF MANUFACTURE
1995-2006

ROYAL ENFIELD 325 TAURUS



VEHICLE TYPE
MOTORCYCLE
FUEL TYPE
DIESEL
DATE OF MANUFACTURE
≤1994

ROYAL ENFIELD SOMMER 462



VEHICLE TYPE
MOTORCYCLE
FUEL TYPE
DIESEL
DATE OF MANUFACTURE
1995–2006

TRACK 800 CDI

VEHICLE TYPE
MOTORCYCLE
FUEL TYPE
DIESEL
DATE OF MANUFACTURE
>2006

VECTRIX VX–1

VEHICLE TYPE
MOTORCYCLE
FUEL TYPE
ELECTRIC
DATE OF MANUFACTURE
1995–2006

VOLKSWAGEN CADDY CARGO

VEHICLE TYPE
UTILITY
FUEL TYPE
DIESEL
DATE OF MANUFACTURE
>2006

VOLKSWAGEN CORRADO G60

VEHICLE TYPE
CAR
FUEL TYPE
GASOLINE
DATE OF MANUFACTURE
≤1994

VOLKSWAGEN GOLF IV

VEHICLE TYPE
CAR
FUEL TYPE
DIESEL
DATE OF MANUFACTURE
1995–2006

ZERO S

VECHICLE TYPE
MOTORCYCLE
FUEL TYPE
ELECTRIC
DATE OF MANUFACTURE
>2006

IS CATEGORY

VALUE ?

NO YES

IS CATEGORY

VALUE ?

NO YES

IS CATEGORY

VALUE ?

NO YES

IS CATEGORY

VALUE ?

NO YES

IS CATEGORY

VALUE ?

NO YES

IS CATEGORY

VALUE ?

NO YES

IS CATEGORY

VALUE ?

NO YES

IS CATEGORY

VALUE ?

NO YES

IS CATEGORY

VALUE ?

NO YES

Algorithm sheets

Category	Value	A	NA	Absolute min.?	Absolute max.?
Vehicle Type	Motorcycle			☐	☐
	Car			☐	☐
	Utility			☐	☐
Fuel Type	Diesel			☐	☐
	Gasoline			☐	☐
	Electric			☐	☐
Date of Manufacture	≤1994			☐	☐
	1995-2006			☐	☐
	≥2006			☐	☐

Category	Value	A	NA	Absolute min.?	Absolute max.?
Vehicle Type	Motorcycle			☐	☐
	Car			☐	☐
	Utility			☐	☐
Fuel Type	Diesel			☐	☐
	Gasoline			☐	☐
	Electric			☐	☐
Date of Manufacture	≤1994			☐	☐
	1995-2006			☐	☐
	≥2006			☐	☐