11a, 2. Angekündigter Kleiner Leistungsnachweis Informatik, 06.06.2024
Beispiellösung

Aufgabe 1:
Ein solcher Datensatz muss einen Wert für die Form des Plättchens und einen für dessen (vorhandenes oder nicht vorhandenes) Muster umfassen sowie eine Angabe, ob das Plättchen aus Gold ist oder nicht – Letzteres ist das sogenannte Label.
Die 14 Datensätze müssen in Trainingsdaten und Testdaten aufgeteilt werden. Üblicherweise wird der größere Teil der Daten als Trainingsdaten verwendet: Aus ihnen wird abgeleitet, wie der Entscheidungsbaum aufgebaut sein soll. Mit den Testdaten wird ein so entstandener Entscheidungsbaum dann getestet, um einschätzen zu können, wie gut er die Plättchen klassifiziert.

Aufgabe 2a:
- Zunächst wird der Algorithmus feststellen, dass die vorhandene Menge von Plättchen hinsichtlich ihres Labels nicht homogen ist und dass es noch Attribute gibt, nach denen die Plättchenmenge „aufgesplittet“ werden kann (s. u.).
- Deshalb wird er (die Verarbeitung nicht abbrechen, sondern) folgendermaßen das „beste/wichtigste Attribut“ zum „Aufsplitten“ bestimmen:
  - Wenn wir alle neun Plättchen mit demselben Label versehen, machen wir mindestens 4 Fehler (bei Klassifizierung als Gold).
  - Wenn wir die Plättchen nach ihrer Form in zwei Teilmengen aufteilen, sind von den quadratischen 2 aus Gold und 2 nicht, von den dreieckigen sind 3 aus Gold und 2 nicht.

	Gold	kein Gold
quadratisch	**2**	2
dreieckig	3	**2**

    Wir machen dann bei der Klassifizierung also mindestens 4 Fehler (quadratisch kein Gold, dreieckig Gold). Der Informationsgewinn (gegenüber der Ausgangsfehlerzahl von 4) beträgt in diesem Fall 4 – 4 = 0.
  - Wenn wir die Plättchen nach ihrem Muster in zwei Teilmengen aufteilen, sind von den blanken 4 aus Gold und 1 ist es nicht, von den gestreiften ist 1 aus Gold und 3 sind es nicht.

	Gold	kein Gold
blank	4	**1**
gestreift	**1**	3

    Wir machen dann bei der Klassifizierung also mindestens 2 Fehler (blank Gold, gestreift kein Gold). Der Informationsgewinn (gegenüber der Ausgangsfehlerzahl von 4) beträgt in diesem Fall 4 – 2 = 2.
  - Der Informationsgewinn ist bei Aufteilung nach dem Muster (positiv und) am höchsten, das Muster ist also das „beste/wichtigste Attribut“.
- Also wird die Menge der Plättchen gemäß vorhandenem oder nicht vorhandenem Muster in zwei Teilmengen aufgeteilt. (Dann wird für jede Teilmenge der Algorithmus wieder durchgeführt.)

Aufgabe 2b:
Teilt man die beiden nach dem Vorgehen in 2a entstandenen Teilmengen nach dem einzigen verbleibenden Attribut, der Form, jeweils wieder in zwei Teilmengen auf, so ist der Informationsgewinn für die blanken Plättchen negativ, für die gestreiften Plättchen ist er bestenfalls 0. Der Entscheidungsbaum würde also nicht besser bzw. in einem Fall sogar schlechter klassifizieren, wenn diese Aufteilung vorgenommen würde.

Aufgabe 3:
Wenn Hennie beim Aussuchen von Plättchen einen Entscheidungsbaum verwendet, kann sie bestenfalls mehr Gold gewinnen, als sie es sonst täte – was erfreulich wäre –, schlimmstenfalls entgeht ihr ein gewisser Gewinn – was verkraftbar wäre.
Beim Einsatz zur Diagnose schwerer Krankheiten besteht die Chance, dass eine Krankheit erkannt wird, die sonst unerkannt bliebe. Wenn dann eine entsprechende Behandlung möglich wird, rettet das möglicherweise Leben. Andererseits kann eine irrtümliche Diagnose einer schweren Krankheit einen Patienten und sein Umfeld (auch unabhängig von möglicherweise darauf folgenden gesundheitsschädlichen Fehlbehandlungen) schwer belasten, was, wenn es sich eigentlich vermeiden lässt, kaum zu verantworten ist.
Deshalb erscheint es sinnvoll, ein zur Verfügung stehendes Diagnosesystem auch zu nutzen (möglicherweise sogar routinemäßig); die mithilfe eines Entscheidungsbaums ermittelte Diagnose muss aber unbedingt von einer erfahrenen Ärztin oder einem erfahrenen Arzt kritisch überprüft werden.