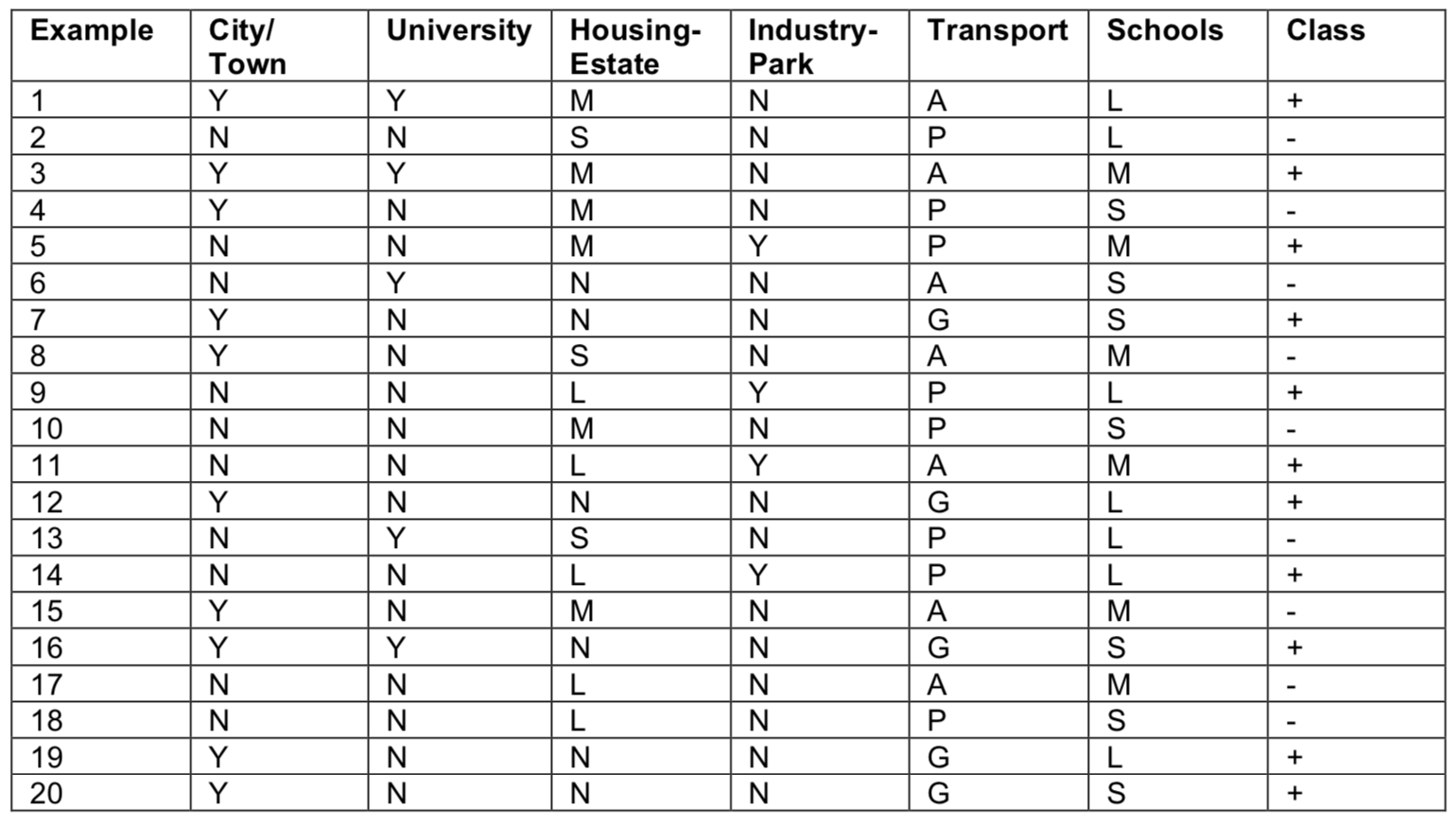
|  |  |
| --- | --- |
| Lab | 7 |
| Student No | 16725829 |
| Registered Programme | DN201 Data Science |
| Date |  |

**Exercise 1**

A brewery that owns a chain of bars/restaurants throughout Ireland needs to assess the suitability of new sites for the establishment of a new premises. When classifying a potential site as suitable/unsuitable it examines a number of attributes about the area where the site is located. These attributes describe whether the site is a city or a large town, whether a university is located in the area, what type of housing estate, if any, is nearby, etc. For each attribute there are a number of attribute values, for example {Y=yes, N=no} for City, {M=medium, S=small, N=none, L=large} for Housing-Estate. A subset of the brewery database is listed below where each existing bar/restaurant is given a class label. Here +(positive) means that the brewery consider the site to be successful and –(negative) means that the site is borderline and they would not wish to make a similar investment in the future. The brewery wish to use the database to find a set of rules that will help them decide on the suitability of a new potential site

****

In the following steps, work out the full decision tree for this problem by filling in the missing values in the white boxes. All numerical answers should be rounded to 3 decimal places.

1. By filling out the following table, calculate which attribute would be selected for expansion under the **root node**:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Id | Attribute | Value | Expanding Root Node (Level 0) | | | |
|  |  |  | H(Y) | H(Y|X=x) | P(X) | I(Y;X) |
| 0 | ROOT NODE |  | 0.993 |  |  |  |
| 1 | City/Town | N |  | 0.971 | 10/20 | 0.067 |
| 2 | City/Town | Y | 0.881 | 10/20 |
| 3 | University | N |  | 0.996 | 15/20 | 0.003 |
| 4 | University | Y | 0.971 | 5/20 |
| 5 | Housing-Estate | L |  | 0.971 | 5/20 | 0.255 |
| 6 | Housing-Estate | M | 1.000 | 6/20 |
| 7 | Housing-Estate | N | 0.650 | 6/20 |
| 8 | Housing-Estate | S | 0.000 | 3/20 |
| 9 | Industry-Park | N |  | 0.988 | 16/20 | 0.203 |
| 10 | Industry-Park | Y | 0.000 | 4/20 |
| 11 | Transport | A |  | 0.985 | 7/20 | 0.266 |
| 12 | Transport | G | 0.000 | 5/20 |
| 13 | Transport | P | 0.954 | 8/20 |
| 14 | Schools | L |  | 0.863 | 7/20 | 0.046 |
| 15 | Schools | M | 1.000 | 6/20 |
| 16 | Schools | S | 0.985 | 7/20 |

1. Starting from the **first node in level 1**, beneath the root node, fill out the following table to determine which attribute would be selected for expansion under that node:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Id | Attribute | Value | Expanding First Node Level 1 | | | |
|  |  |  | H(Y) | H(Y|X=x) | P(X) | I(Y;X) |
| 0 | ROOT NODE |  |  |  |  |  |
| 1 | City/Town | N |  | 0.918 | 3/7 | -0.007 |
| 2 | City/Town | Y | 1.000 | 4/7 |
| 3 | University | N |  | 0.811 | 4/7 | 0.128 |
| 4 | University | Y | 0.918 | 3/7 |
| 5 | Housing-Estate | L |  | 1.000 | 2/7 | 0.278 |
| 6 | Housing-Estate | M | 0.918 | 3/7 |
| 7 | Housing-Estate | N | 0.000 | 1/7 |
| 8 | Housing-Estate | S | 0.000 | 1/7 |
| 9 | Industry-Park | N |  | 0.918 | 6/7 | 0.198 |
| 10 | Industry-Park | Y | 0.000 | 1/7 |
| 11 | Transport | A | 0.985 |  | 1.000 |  |
| 12 | Transport | G |  | 0.000 |
| 13 | Transport | P |  | 0.000 |
| 14 | Schools | L |  | 0.000 | 1/7 | 0.291 |
| 15 | Schools | M | 0.971 | 5/7 |
| 16 | Schools | S | 0.000 | 1/7 |

1. Starting from the **first node in level 2**, beneath the node expanded at step b), fill out the following table to determine which attribute would be selected for expansion under that node:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Id | Attribute | Value | Expanding First Node Level 2 | | | |
|  |  |  | H(Y) | H(Y|X=x) | P(X) | I(Y;X) |
| 0 | ROOT NODE |  |  |  |  |  |
| 1 | City/Town | N |  | 0.971 | 5/5 | 0.029 |
| 2 | City/Town | Y | 0.000 | 0/5 |
| 3 | University | N |  | 0.971 | 5/5 | 0.029 |
| 4 | University | Y | 0.000 | 0/5 |
| 5 | Housing-Estate | L | 1.000 |  | 1.00 |  |
| 6 | Housing-Estate | M |  | 0.00 |
| 7 | Housing-Estate | N |  | 0.00 |
| 8 | Housing-Estate | S |  | 0.00 |
| 9 | Industry-Park | N |  | 0.000 | 2/5 | 0 |
| 10 | Industry-Park | Y | 0.000 | 3/5 |
| 11 | Transport | A |  |  |  |  |
| 12 | Transport | G |  |  |
| 13 | Transport | P |  |  |
| 14 | Schools | L |  | 0.000 | 2/5 | 1 |
| 15 | Schools | M | 1 | 2/5 |
| 16 | Schools | S | 0.000 | 1/5 |

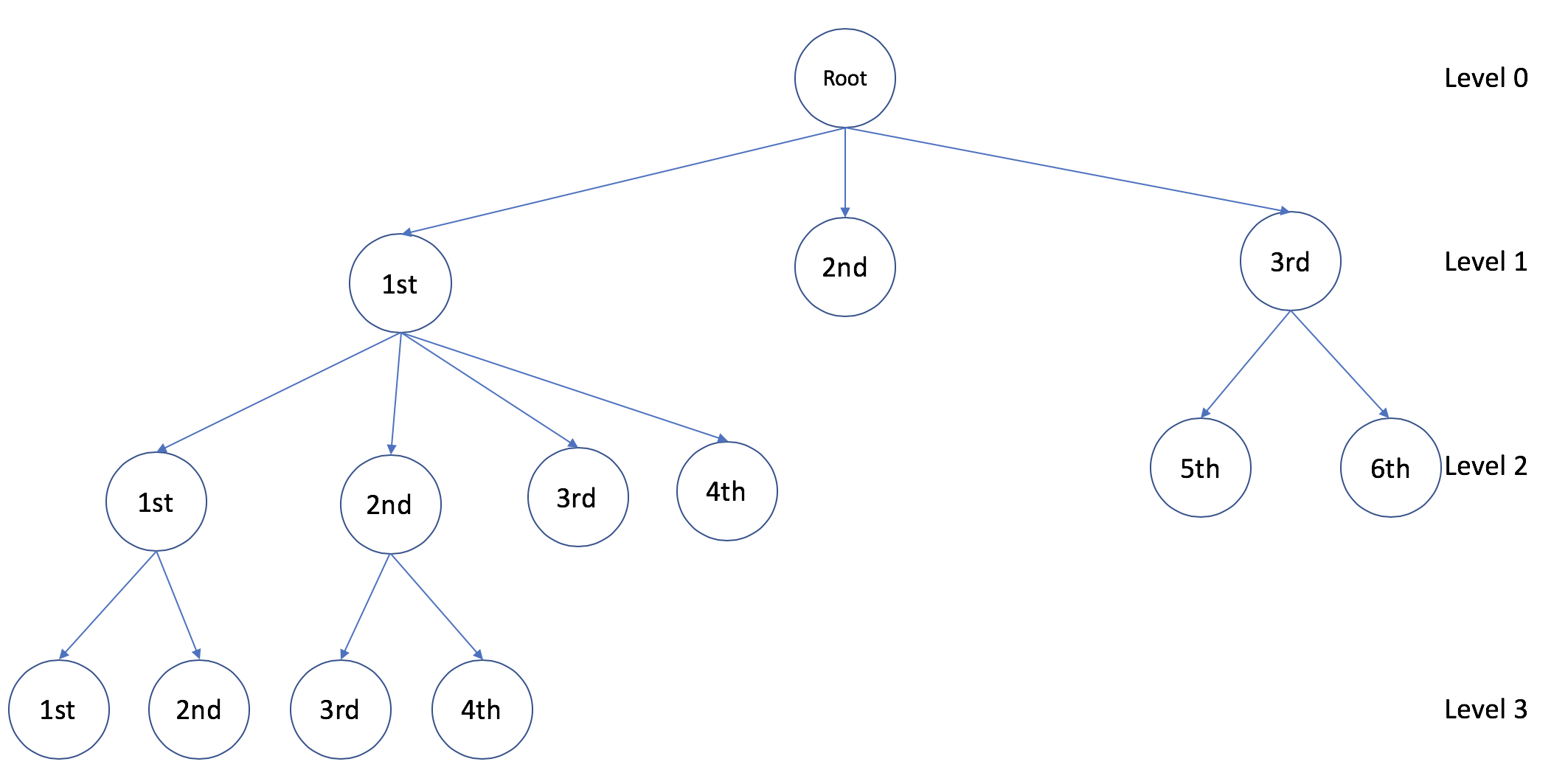
1. Starting from the **second node in level 2**, beneath the node expanded at step b), fill out the following table to determine which attribute would be selected for expansion under that node:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Id | Attribute | Value | Expanding Second Node Level 2 | | | |
|  |  |  | H(Y) | H(Y|X=x) | P(X) | I(Y;X) |
| 0 | ROOT NODE |  |  |  |  |  |
| 1 | City/Town | N |  | 1 | 2/6 | -0.0148 |
| 2 | City/Town | Y | 1 | 4/6 |
| 3 | University | N |  | 0.811 | 4/6 | 0.444 |
| 4 | University | Y | 0 | 2/6 |
| 5 | Housing-Estate | L | 0.9852 |  | 0.00 |  |
| 6 | Housing-Estate | M |  | 1.00 |
| 7 | Housing-Estate | N |  | 0.00 |
| 8 | Housing-Estate | S |  | 0.00 |
| 9 | Industry-Park | N |  | 0.971 | 5/6 | 0.176 |
| 10 | Industry-Park | Y | 0 | 1/6 |
| 11 | Transport | A |  |  |  |  |
| 12 | Transport | G |  |  |
| 13 | Transport | P |  |  |
| 14 | Schools | L |  | 0 | 1/6 | 0.526 |
| 15 | Schools | M | 0.918 | 3/6 |
| 16 | Schools | S | 0 | 2/6 |

1. Starting from the **third node in** **level 1**, beneath the root node, fill out the following table to determine which attribute would be selected for expansion under that node: (why does the second node on level 1 not need to be expanded?)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Id | Attribute | Value | Expanding Third Node Level 1 | | | |
|  |  |  | H(Y) | H(Y|X=x) | P(X) | I(Y;X) |
| 0 | ROOT NODE |  |  |  |  |  |
| 1 | City/Town | N |  | 0.985 | 7/8 | 0.092 |
| 2 | City/Town | Y | 0 | 1/8 |
| 3 | University | N |  | 0.985 | 7/8 | 0.952 |
| 4 | University | Y | 0 | 1/8 |
| 5 | Housing-Estate | L |  | 0.918 | 3/8 | 0.2655 |
| 6 | Housing-Estate | M | 0.918 | 3/8 |
| 7 | Housing-Estate | N | 0 | 0/8 |
| 8 | Housing-Estate | S | 0 | 2/8 |
| 9 | Industry-Park | N |  | 0 | 5/8 | 0.954 |
| 10 | Industry-Park | Y | 0 | 3/8 |
| 11 | Transport | A | 0.954 |  | 0.00 |  |
| 12 | Transport | G |  | 0.00 |
| 13 | Transport | P |  | 1.00 |
| 14 | Schools | L |  | 1 | 4/8 | 0.454 |
| 15 | Schools | M | 0 | 1/8 |
| 16 | Schools | S | 0 | 3/8 |

1. Finally arrive at the decision tree. Each node in the tree corresponds to a particular attribute and value. Fill in the appropriate attribute and value for each node.



1. Using the decision tree, decide whether the following sites would be considered suitable:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Test Site** | **City/Town** | **University** | **Housing-Estate** | **Industry-Park** | **Transport** | **Schools** | **Decision** |
| i) | N | Y | N | N | P | L | + |
| ii) | Y | N | S | Y | G | S | - |
| iii) | Y | Y | S | N | A | S | - |

**Input your answers into the Moodle Quiz associated with this Lab.**