Question 1

Correct

Mark 1.40 out of 2.00

Categorize each of the following by the steps involved in knowledge discovery from databases. Choose "none" if none of the steps apply.

Determine if the precision and recall trade-off in detecting credit card fraud is acceptable.

interpretation

Correct answer, well done.

Marks for this submission: 0.40/0.40.

Study the bias-variance trade-off.

none

Correct answer, well done.

Marks for this submission: 0.40/0.40. Accounting for previous tries, this gives **0.00/0.40**.

Encode categorical variables using one-hot encoding.

preprocessing

Correct answer, well done.

Marks for this submission: 0.40/0.40.

Obtain a decision tree that determines whether a credit card transaction is fraud.

data mining

Correct answer, well done.

Marks for this submission: 0.40/0.40. Accounting for previous tries, this gives **0.20/0.40**.

Determine what to mine about viewers' preferences on movies.

preparation

Correct answer, well done.

Marks for this submission: 0.40/0.40.

A correct answer is: `"interpretation"`

A correct answer is: `"none"`

A correct answer is: `"preprocessing"`

A correct answer is: `"data mining"`

A correct answer is: `"preparation"`

Question 2

Correct

Mark 2.00 out of 2.00

Identify the main knowledge discovery process for each of the following tasks. Choose "none" if none of the processes apply.

Measure the performance of different marketing strategies.

none

Correct answer, well done.

Marks for this submission: 0.40/0.40.

Learn from users' listening history how to suggest music tracks.

association rule mining

Correct answer, well done.

Marks for this submission: 0.40/0.40.

Predict the lifetime of a machine part.

regression

Correct answer, well done.

Marks for this submission: 0.40/0.40.

Categorize customer feedback into positive or negative.

classification

Correct answer, well done.

Marks for this submission: 0.40/0.40.

Segment a mammogram into candidate objects (in pixels) to detect microcalcifications.

cluster analysis

Correct answer, well done.

Marks for this submission: 0.40/0.40.

A correct answer is: `"none"`

A correct answer is: `"association rule mining"`

A correct answer is: `"regression"`

A correct answer is: `"classification"`

A correct answer is: `"cluster analysis"`

Question 3

Correct

Mark 0.00 out of 4.00

Reorder the following blocks to give the pseudocode of the RIPPER (Repeated Incremental Pruning to Produce Error Reduction) algorithm. You need NOT specify the indentation/grouping of different blocks.

- Input: Training data D
- Output: Decision list L

- ✓ Split D into D' and D''.
- ✓ Initialize R to be the rule $Y = 1$ without antecedent.
- ✓ Search for the conjunct C that maximizes the FOIL gain on D' of adding C to R.
- ✓ If the FOIL gain is positive, add the conjunct C to R and repeat the search for another conjunct to add.
- ✓ Eliminates the last added conjunct C from R if doing so reduces the error rate on D''.
- ✓ Repeat the step that eliminates a conjunct until there is no more reduction in the error rate.
- ✓ Add R to L and remove the covered instances from D'.
- ✓ If D' is not empty, repeat from the step that initializes R.

Your answer is correct.

Correct

Marks for this submission: 4.00/4.00. Accounting for previous tries, this gives **0.00/4.00**.

Question 4

Correct

Mark 4.00 out of 4.00

Consider the data D on customer satisfaction below with two nominal input attributes $X_1, X_2 \in \{0, 1\}$, where the target $Y \in \{0, 1, 2\}$ is the level of satisfaction:

$$\begin{bmatrix} i & X_1 & X_2 & Y \\ 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 \\ 3 & 0 & 0 & 2 \\ 4 & 0 & 0 & 1 \\ 5 & 1 & 0 & 0 \\ 6 & 0 & 0 & 2 \end{bmatrix}$$

Calculate the following quantities related to the Gini impurity measure on the data.

$\text{Gini}(D) =$ 2/3 $\frac{2}{3}$

Correct answer, well done.

Marks for this submission: 0.80/0.80.

$\text{Gini}_{X_1}(D) =$ 7/12 $\frac{7}{12}$

Correct answer, well done.

Marks for this submission: 0.80/0.80.

$\text{Gini}_{X_2}(D) =$ 2/3 $\frac{2}{3}$

Correct answer, well done.

Marks for this submission: 0.80/0.80.

$\Delta\text{Gini}_{X_1}(D) =$ 1/12 $\frac{1}{12}$

Correct answer, well done.

Marks for this submission: 0.80/0.80.

$\Delta\text{Gini}_{X_2}(D) =$ 0 0

Correct answer, well done.

Marks for this submission: 0.80/0.80.

The answer $\frac{2}{3}$, which can be typed as `2/3`, would be correct.

The answer $\frac{7}{12}$, which can be typed as `7/12`, would be correct.

The answer $\frac{2}{3}$, which can be typed as `2/3`, would be correct.

The answer $\frac{1}{12}$, which can be typed as `1/12`, would be correct.

The answer 0, which can be typed as `0`, would be correct.

Question 5

Correct

Mark 7.64 out of 8.00

Given the F_β-score is $\frac{5}{7}$ and β is 2 for a spam email classifier (used to classify emails as spam or not spam), what are the minimum and maximum possible precision and recall.

- F_β is an increasing function of precision (recall), which is a fraction between 0 and 1.
- $F_\beta = (1+\beta^2)\frac{1}{\beta^2 \mathrm{TPR}^{-1} + \mathrm{PPV}^{-1}}$, where TPR is the recall, and PPV is the precision.

	minimum		maximum	
precision	1/3	$\frac{1}{3}$	1	1
	Correct answer, well done. Marks for this submission: 0.73/0.73.		**Correct answer, well done.** Marks for this submission: 0.73/0.73.	
recall	2/3	$\frac{2}{3}$	1	1
	Correct answer, well done. Marks for this submission: 0.73/0.73.		**Correct answer, well done.** Marks for this submission: 0.73/0.73.	

Assume further that the classifier was tested with 10 emails that are spam (positive instances) and 10 emails that are not spam (negative instances).

Select the questions answered by the corresponding evaluation metrics:

- Question: Given an email is classified as not spam, what is the chance it is not spam?

  Correct answer, well done.

  Marks for this submission: 0.73/0.73. Accounting for previous tries, this gives **0.36/0.73**.

  Answer: The negative predictive value is $\frac{8}{11}$.

- Question: Given an email is not spam, what is the chance it is classified as spam?

  Correct answer, well done.

  Marks for this submission: 0.73/0.73.

  Answer: The false positive rate is $\frac{1}{5}$.

Using the additional information above, complete the confusion matrix below with the corresponding integer counts:

	classified as spam	classified as not spam
spam	7	3
	7 **Correct answer, well done.** Marks for this submission: 0.73/0.73.	3 **Correct answer, well done.** Marks for this submission: 0.73/0.73.
not spam	2	8
	2 **Correct answer, well done.** Marks for this submission: 0.73/0.73.	8 **Correct answer, well done.** Marks for this submission: 0.73/0.73.

Finally, calculate the error rate:

error rate: 1/4 $\frac{1}{4}$

Correct answer, well done.

Marks for this submission: 0.73/0.73.

The answer $\frac{1}{3}$, which can be typed as `1/3`, would be correct.

The answer 1, which can be typed as `1`, would be correct.

The answer $\frac{2}{3}$, which can be typed as `2/3`, would be correct.

The answer 1, which can be typed as `1`, would be correct.

A correct answer is: `"Given an email is classified as not spam, what is the chance it is not spam? "`

A correct answer is: `"Given an email is not spam, what is the chance it is classified as spam? "`

The answer 7, which can be typed as `7`, would be correct.

The answer 3, which can be typed as `3`, would be correct.

The answer 2, which can be typed as `2`, would be correct.

The answer 8, which can be typed as `8`, would be correct.

The answer $\frac{1}{4}$, which can be typed as `1/4`, would be correct.