FACULTY of SCIENCE and ENGINEERING
Department of Computer Science and Information Systems

End-of-Semester Exam

Academic Year:	2021/2022	**Semester:** Spring
Module Title:	Data Mining	**Module Code:** CS4168
Exam Duration:	2 Hours	**Total Marks:** 50
Lecturer:	Dr. N. S. Nikolov	(Equal to 50% of the final grade)

Instructions to Candidates:

This exam consists of 2 sections.

Section A:

- Section A consists of 30 multiple-choice questions, each having exactly one correct answer.
- You **must use the bubble sheets** to answer the questions in this section.
- Answer ALL questions. Each question will be marked as follows:
  - Correctly answered: 1 mark
  - Incorrectly answered: -0.25 marks
  - Unanswered: 0 marks

Section B:

- Choose to answer either Q31 or Q32.
- Write your answers in the **answer booklet**.
- If both questions are answered, the one with higher marks will be considered.

SECTION A: Multiple Choice Questions – answer ALL questions **30 marks**

Q1. According to Witten and Frank, data mining is the extraction of implicit, _____ and potentially useful information from data.

Q2. The input of a machine learning algorithm typically consists of a large number of examples which are:

Q3. When treating missing values in a dataset, it is recommended to drop a column if more than _____ of the values in it are missing.

Q4. In scikit-learn, _____ results in a distribution with a standard deviation equal to 1 and a mean approximately 0.

Q5. Not removing outliers from a dataset can increase the probability for training a/an _____ predictive model.

Q6. The confidence in the accuracy of a predictive model is higher when the training, validation and test datasets are:

Q7. Cross-validation and repeated holdout are methods used in the evaluation of:

Q8. In *k*-fold cross validation, the _____ in the training dataset are split into *k* subsets of equal size.

Q9. In classification, the harmonic mean of the precision and recall is known as:

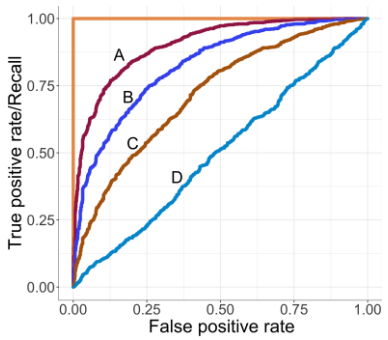
Q10. The prime use of a validation dataset is for:

Q11. Assume a classification model does not perform well on a well-designed test dataset. Which one of the following actions is inappropriate?

Q12. A regression model can be evaluated by calculating the _____ on a test dataset.

Q13. The plot below contains the ROC curves of four alternative classification models for the same classification problem. Which one of them is the best model?

Q14. You have trained four classifiers that determine if a certain mushroom is edible (positive class) or poisonous (negative class). Which one of the classifiers, represented by their confusion matrices, would use to select the mushrooms for your risotto?



Q15. Scikit-learn pipelines help avoid _____ in cross-validation, by ensuring that all steps in the pipeline are applied separately to the training and test datasets.

Q16. All steps in a scikit-learn pipeline, except the last one, must be _____.

Q17. If the examples in a training dataset are not linearly separable, it may be possible to make them linearly separable by:

Q18. The _____ algorithm trains a single-layer neural network.

Q19. ID3 employs _____ to build a decision tree.

Q20. One of the parameters of kNN is the _____ .

Q21. Logistic regression and Perceptron work best if the examples are _____ separable.

Q22. The parameter k in the clustering algorithm k-means is the:

Q23. The optimal value of the parameter k in the k-means clustering algorithm can be determined with the _____ method.

Q24. _____ methods learn which features best contribute to the accuracy of the model while the model is being trained.

Q25. _____ feature selection methods apply a statistical measure to assign a scoring to each feature.

Q26. _____ methods consider the selection of a set of features as a search problem, where different combinations of features are prepared, evaluated and compared to other combinations.

Q27. The Hughes effect refers to the observation that with a fixed number of training examples the accuracy of a classifier first:

Q28. In collaborative filtering, the decomposition of the user-item interaction _____ into the product of two lower dimensionality _____ is known as _____.

Q29. Content-based filtering and collaborative filtering are _____ approaches.

Q30. In text mining, the tf-idf weighting scheme is used for reducing the weight of words occurring:

END OF SECTION A

SECTION B: Problem Solving – choose to answer either Q31 or Q32 — 20 marks

Q31. (20 marks) Consider the **training** and **test datasets** in Tables 1 and 2, respectively.

Table 1. Training dataset.

Education	Income (in thousands)	Property	Loan_Status
Primary	41	Rural	Yes
Primary	20	Rural	Yes
Primary	25	Urban	No
Secondary	50	Rural	No
Secondary	31	Urban	Yes
Secondary	150	Urban	Yes
Tertiary	45	Rural	Yes
Tertiary	30	Urban	Yes

Table 2. Test dataset.

Education	Income (in thousands)	Property	Loan_Status
Secondary	58	Urban	Yes
Secondary	35	Rural	Yes
Primary	39	Rural	Yes
Tertiary	80	Urban	Yes
Tertiary	20	Rural	Yes
Primary	60	Urban	No

a. **(5 marks)** Draw a boxplot for all values in column **Income** in both Table 1 and Table 2, i.e., the array of values [20, 20, 25, 30, 31, 35, 39, 41, 45, 50, 58, 60, 80, 150]. Calculate *Q1*, *Q2*, *Q3* and IQR. Are there any outliers in the array and why?

b. **(12 marks)** Replace the values in column **Income** by two categories:
High (for Income >= 40) and **Normal** (for Income < 40). Then apply the ID3 algorithm to the **training dataset** to build a decision tree for predicting the value of **Loan_Status**.

c. **(3 marks)** Calculate the accuracy of the decision tree you have built on the **test dataset** in Table 2. Also calculate the precision and the recall for class **Yes** on the **test dataset**.

Hints:

Boxplot: The box in a boxplot extends from $Q1$ to $Q3$, known as interquartile range (IQR). Any values smaller than $Q1 - 1.5 * (Q3-Q1)$ or larger than $Q3 + 1.5 * (Q3-Q1)$ are considered outliers and are drawn as separate dots. The two whiskers extend from $Q1$ and $Q3$ to the smallest and the largest values within $1.5 * (Q3-Q1)$, respectively.

Info function: Let $(a + b + c) = p$ and $(d + e) = q$. Then:

$$\text{info}[a,b,c] = -\left(\frac{a}{p}\right) \times \log_2\left(\frac{a}{p}\right) - \left(\frac{b}{p}\right) \times \log_2\left(\frac{b}{p}\right) - \left(\frac{c}{p}\right) \times \log_2\left(\frac{c}{p}\right)$$

$$\text{info}([a,b,c],[d,e]) = \left(\frac{p}{p+q}\right)\text{info}[a,b,c] + \left(\frac{q}{p+q}\right)\text{info}[d,e]$$

accuracy = (TP + TN)/(TP + FP + TN + FN), **precision** = TP/(TP+FP), **recall** = TP/(TP+FN)

Q32. (20 marks) Consider the user-movie rating matrix in Table 3, as well as the user-user similarity matrix in Table 4 and the movie-movie similarity matrix in Table 5.

Table 3. User-movie rating matrix.

	Movie1	**Movie2**	**Movie3**	**Movie4**	**Movie5**	**Movie6**
User1	4		3	5		
User2			3		5	
User3	5	4		4		
User4		2			5	4
User5			4		4	
User6	4	5		4		3

Table 4. User-user similarity matrix.

	User1	**User2**	**User3**	**User4**	**User5**	**User6**
User1	1.00	0.22	0.75	0.00	0.30	0.63
User2	0.22	1.00	0.00	0.64	0.97	0.00
User3	0.75	0.00	1.00	0.16	0.00	0.91
User4	0.00	0.64	0.16	1.00	0.53	0.40
User5	0.30	0.97	0.00	0.53	1.00	0.00
User6	0.63	0.00	0.91	0.40	0.00	1.00

Table 5. Movie-movie similarity matrix.

	Movie1	**Movie2**	**Movie3**	**Movie4**	**Movie5**	**Movie6**
Movie1	1.00	0.79	0.27	0.98	0.00	0.32
Movie2	0.79	1.00	0.00	0.71	0.18	0.69
Movie3	0.27	0.00	1.00	0.34	0.65	0.00
Movie4	0.98	0.71	0.34	1.00	0.00	0.32
Movie5	0.00	0.18	0.65	0.00	1.00	0.49
Movie6	0.32	0.69	0.00	0.32	0.49	1.00

In both user-based and item-based neighborhood collaborative filtering below, consider **only the top 2** most similar users to **User1** and the **only top 2** most similar movies to the movies previously rated by **User1**, respectively.

a. **(10 marks)** Apply the user-based neighbourhood collaborative filtering method to calculate movie recommendation scores for **User1**. Which one of the movies not rated by **User1** yet would you recommend to **User1** first?

b. **(10 marks)** Apply the item-based neighbourhood collaborative filtering method to calculate movie recommendation scores for **User1**. Which one of the movies not rated by **User1** yet would you recommend to **User1** first?

END OF EXAM