First name: _____

Last name: _____

Student ID number: | | | | | | | | |

1st Midterm Exam

course name

INTRODUCTION TO MACHINE LEARNING AND DATA MINING

Instructions:

- Write your **FIRST NAME**, **LAST NAME** and **STUDENT ID NO.** on each piece of paper with solutions;
- This midterm is composed of **6 assignments** for the total amount of **100 points**;
- Solving time is **90 minutes**;
- Only a calculator and 1 piece of paper (A4 format) – with written notes and formulas is allowed;
- All other literature, the use of Internet, laptops, mobile phones and other electronic devices is strictly forbidden!

Koper, 29th of November, 2018

the red color is for task number 3.

I	D	D_d	E	F	G	C
100	2000.000	D1	1	c	1	**a**
101	2003.210	D1	0	a	4	**b**
102	2012.104	D2	1	b	11	**a**
103	2004.443	D1	0	a	10	**a**
104	2002.508	D1	0	c	8	**b**
105	2013.675	D2	1	b	15	**b**
106	2020.000	D3	1	c	4	**c**
107	2016.345	D3	0	a	8	**a**
108	2011.248	D2	1	b	17	**b**
109	2017.565	D3	0	b	6	**a**
110	2008.225	D2	1	a	14	**c**

I: Identifier [0, ∞)
D: Date in KSP format
E: Nominal value {0, 1}
F: Nominal value {a, b, c}
G: Numeric value [0, 20]
C: Class; nominal value {**a**, **b**, **c**}

green part for task 5

blue part task 6

I	D (not KSP)	D_d	E	F	G	C_{NB}	C_{DT}
200	11.2.2017	D3	0	a	17	a	a
201	1.5.2011	D2	1	c	1	c	b
202	**21.6.2012**	D2	1	b	9	b	b

1. Transform the values of attribute **D** for the examples with *I = 200, 201 and 202* into the KSP format (leap year is **bolded**)! Round your results to 3 decimal places! **(10 points)**

2. Draw a boxplot that will represent the values of attribute **G** (take into consideration only examples with *I = 100 – 110*)! **(10 points)**

3. Discretize attribute **D** into 3 bins using the equal height discretization technique (take into consideration only examples with *I = 100 – 110*)! Denote the values of this new (discretized) attribute **D_d** as D1, D2 and D3. Draw the histogram! **(10 points)**

4. Use the **OneR** algorithm to classify the examples with known class value (examples with *I = 100 – 110*)! Check just the attributes **E** and **F**. Sketch the (one level) decision tree! What is the error of this classifier? **(15 points)**

5. Use the **Naïve Bayes** classifier to classify the examples with unknown class value (examples with *I = 200, 201 and 202*)! Build the probability tables by using just the attributes **E** and **F**. Use the Laplace correction to calculate the probabilities! **(25 points)**

6. Build a one level decision tree (root node only) by using the TDIDT principle (ID3 algorithm). Check just attributes **E**, **F** and **D_d** (as potential candidates for the root node). Use the information gain as the »impurity measure« for ranking the attributes. Draw this »partially constructed« decision tree and use it to classify the examples with unknown class value (examples with *I = 200, 201 and 202*)! **(30 points)**