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

I	D	D_d	E	F	G	C
100	2000.000		T	B	4	**Y**
101	2004.102		F	R	8	**N**
102	2011.454		T	G	12	**Y**
103	2005.666		F	R	9	**Y**
104	2002.128		F	B	7	**N**
105	2014.775		T	G	18	**N**
106	2020.000		T	B	11	**N**
107	2018.245		F	R	3	**Y**
108	2012.243		T	G	19	**N**
109	2019.005		F	G	7	**Y**
110	2009.506		T	R	10	**N**

I: Identifier [0, ∞)
D: Date in KSP format
E: Nominal value {T, F}
F: Nominal value {R, G, B}
G: Numeric value [0, 20]
C: Class; nominal value {**Y**, **N**}

I	D (not KSP)	D_d	E	F	G	C_{NB}	C_{DT}
200	12.2.2013		T	R	11		
201	6.6.2017		T	G	5		
202	**17.5.2012**		F	B	15		

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 4 bins using the equal frequency discretization technique (take into consideration only examples with *I = 100 – 110*)! Denote the values of this new (discretized) attribute **D_d** as D1, D2, D3 and D4. 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 Gini index 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)**