INSTITUTO SUPERIOR TÉCNICO
DEPARTAMENTO DE MATEMÁTICA

Project – Multivariate Statistical Methods for Engineering and Management & Statistical Methods in Data Mining

(MEGI & MMA, 1^{st} Semester, 2017/2018)

Handed out on October 31, 2017.

To be handed back by December 11, 2017.

1. Make a preliminary analysis of the data and discuss what you have learned from this analysis.
2. Solve your classification problem using supervised learning methods. Have in mind that some of the input variables may be irrelevant to the classification problem and that you may need to do some preprocessing methodologies of your data set e.g. dimensionality reduction techniques.
3. Apply unsupervised methods to your data. Interpret the results. Compare the obtained partitions with the true class each object belongs to.
4. Repeat the classification study using for classes the partition clusters obtained in (3). Compare the results with the ones obtained in 2. Discuss potential advantages and drawbacks of each strategy. Which would you recommend to analyze your dataset?

   Include in your discussion all options that you have made, advantages and disadvantages of each alternative.
5. Imagine you are going to meet the researcher who contacted you. Report to him/her what you have learned about the problem. Discuss limitations of the analysis you have done and provide suggestions for future work.

The datasets and Oral Presentation slots are distributed as follows:

Group	Number	Name	Dataset	Oral Present.
1	80868	Catarina Padrela Loureiro	A. Secondary schools	18 Oct, 9h
	80973	Inês Margarida Ribeiro de Oliveira		
	81027	Manuel Maria Da Costa Lorga Dias Portela		
	81085	Fernando Miguel Carvalho Subtil		
	81208	João Pedro Rodrigues Gois		
2	86132	MichelăTakao KozakiăBessa	Students choice	18 Oct, 9h30
	86302	António Luís Fernandes Pais		
	86303	André do Bem Pina Prata		
	87877	Luis Sánchez Riveiro		
	87895	Joao Paulo Antunes Felicio		
3	69947	António Cardoso	B. Italian Olive Oil	18 Oct, 15h30
	86327	Sara Moreira		
	88654	Matilde Telmon		
	90835	Marta Castro		
4	80869	Miguel Rebocho	Student choice	18 Oct, 10h30
	81194	Beatriz Lourenço		
	81280	Rita Matos		
	81775	Pedro Borralho		
	90820	Maria Beatriz Roseiro		
5	78736	Beatriz Xavier	Student choice	18 Oct, 11h
	78986	Tiago Afonso		
	79178	Leonor Botelho		
	81831	Diogo Carvalho		
	87934	Bernardo Amorim Vieira		
6	78670	Luís Franco	C. Human Thyroid Data	18 Oct, 14h
	81424	António Capela		
	82191	Pedro Ferreira		
7	76483	Rodrigo Macedo	Student choice	18 Oct, 14h30
	77932	Joana Cidade Alves		
	77959	Tomás Morais Bello		
	77963	Manuel Mirante Granate		
	86374	Ana Carolina Fonseca		
8	75917	Patricia Felizardo	Student choice	18 Oct, 15h
	76091	Eduardo Mendes		
	78686	Bruna Mason		
	88551	Luis Angel Espinosa Villalpando		
	88819	Pablo Valdés-Stauber		
9	46258	Rui Costa	D. United Nations	18 Oct, 10h
	81403	Miguel Frazão		

- **About the presentation:**
  - Duration of 20 minutes plus 10 minutes for discussion.
- **About the report:**
  - The report should not exceed 10 pages, in the form of a scientific paper.
- Do not forget topics such as:
  1. Description of the problem under study;
  2. Objectives;
  3. Estimation and validation methods;
  4. Discussion of the results and interpretation of the findings;
  5. Conclusions;
  6. References.
- The R code, the presentation slides as well as the report must be send by email to:
  `rosario.oliveira@tecnico.ulisboa.pt`
-

Datasets Description:

A. **Secondary schools:** Medianas das classificação internal e final e classificação nos exames nacionais de 225 estabelecimentos de ensino, nas seguintes disciplinas do Ensino Secundário: Matemática A, Português, Biologia e Geologia, Geometria Descritiva e Fisica e Química A. Apenas estabelecimentos de ensino com mais de 10 notas em cada disciplina foram considerados.

   - Escola - indentificador dao estabelecimento de ensino
   - Exame.MAT - Classificação Exame de Matemática A;
   - CIF.MAT - Classificação Interna Final a Matemática A;
   - Exame.PT - Classificação Exame de Português;
   - CIF.PT - Classificação Interna Final a Português;
   - Exame.BG - Classificação Exame de Biologia ou Geologia;
   - CIF.BG - Classificação Interna Final a Biologia ou Geologia;
   - Exame.GD - Classificação Exame de Geometria Descritiva;
   - CIF.GD - Classificaçãoo Interna Final a Geometria Descritiva;
   - Exame.FQ - Classificação Exame de Física e Química A;
   - CIF.FQ - Classificação Interna Final a Física e Química A;
   - PUB_PRIV: Tipo de estabelecimentos de ensino: público ou privado;
   - Distrito.

B. **Italian Olive Oil:** In order to characterize the olive oils from different growing regions of Italy, it has determined the concentrations of 8 fatty acids in 572 samples of olive oil from 9 different growing regions.

C. **Human Thyroid Data:** This data set comprises 215 patients form the same hospital. These individuals were divided into 3 groups of known classification, euthyroid patients (EU) for which there were 150 cases, patients suffering from hyperthyroidism (HYPER) for which there were 35 cases from hypothyroidism (HYPO) for which there were 30 cases. Each individual was characterized by the result of 5 laboratory test; total serum thyroxine (T4), total serum tri-iodothyronine (T3) or (T3RIA), T3 resin uptake (RT3U), serum thyroid-stimulating hormone (TSH), and increase TSH after injection of TSH-releasing hormone (TSH). The primary problems here are distinguishing EU cases from either HYPER or HYPO as HYPER and HYPO can be easily distinguished from one another.

D. **United Nations:** In September 2000, leaders from 189 nations agreed on a vision for the future: a world with less poverty, hunger and disease, greater survival prospects for mothers and their infants, better educated children, equal opportunities for women, and a healthier environment; a world in which developed and developing countries worked in partnership for the betterment of all. This vision took the shape of eight Millennium Development Goals, which provide a framework of time-bound targets by which progress can be measured."[1]

   United Nations chose 74 countries to be analysed. Each country has 12 indicators of progress belonging to 4 goals of the Millennium Development, namely:

   - Goal 1: Eradicate extreme poverty and hunger

     v1 Employment-to-female-population ratio (percentage)

     v2 Employment-to-male-population ratio (percentage)

     v3 Growth rate of GDP per person employed (the current values of 1990, in thousands of dollars)

   - Goal 3: Promote gender equality and empower women

     v1 Proportion of seats held by women in national parliament (percentage)

     v2 Ratios of girls to boys in primary, secondary and tertiary education

     v3 Share of women in wage employment in the non-agricultural sector (percentage)

   - Goal 6: Combat HIV/AIDS, malaria and other diseases

     v7 Incidence rate associated with tuberculosis (per 100,000 people)

     v8 Detected rate associated with tuberculosis (all forms)

     v9 Death rate associated with tuberculosis (per 100,000 people)

   - Goal 7: Ensure environmental sustainability

     v10 CO2 emissions (tons per capita)

     v11 Energy use (tons of oil equivalent per capita)

     v12 Proportion of land area covered by forest

[1] `https://mdgs.un.org/unsd/mdg/Host.aspx?Content=Products/Progress2005.htm`