Course guide
270418 - PMAAD - Preprocessing and Advanced Models of Data Analysis

Last modified: 02/02/2024

Unit in charge: Barcelona School of Informatics
Teaching unit: 715 - EIO - Department of Statistics and Operations Research.

Degree: BACHELOR'S DEGREE IN ARTIFICIAL INTELLIGENCE (Syllabus 2021). (Compulsory subject).

Academic year: 2023 **ECTS Credits:** 6.0 **Languages:** Catalan, English

LECTURER

Coordinating lecturer: CARINA GIBERT OLIVERAS

Others:

PRIOR SKILLS

The courses of Statistical Modeling and Probability and Statistics

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE09. To ideate, design and integrate intelligent data analysis systems with their application in production and service environments.
CE17. To develop and evaluate interactive systems and presentation of complex information and its application to solving human-computer and human-robot interaction design problems.
CE18. To acquire and develop computational learning techniques and to design and implement applications and systems that use them, including those dedicated to the automatic extraction of information and knowledge from large volumes of data.
CE20. To select and put to use techniques of statistical modeling and data analysis, assessing the quality of the models, validating and interpreting.

Generical:
CG4. Reasoning, analyzing reality and designing algorithms and formulations that model it. To identify problems and construct valid algorithmic or mathematical solutions, eventually new, integrating the necessary multidisciplinary knowledge, evaluating different alternatives with a critical spirit, justifying the decisions taken, interpreting and synthesizing the results in the context of the application domain and establishing methodological generalizations based on specific applications.
CG8. Perform an ethical exercise of the profession in all its facets, applying ethical criteria in the design of systems, algorithms, experiments, use of data, in accordance with the ethical systems recommended by national and international organizations, with special emphasis on security, robustness , privacy, transparency, traceability, prevention of bias (race, gender, religion, territory, etc.) and respect for human rights.
CG9. To face new challenges with a broad vision of the possibilities of a professional career in the field of Artificial Intelligence. Develop the activity applying quality criteria and continuous improvement, and act rigorously in professional development. Adapt to organizational or technological changes. Work in situations of lack of information and / or with time and / or resource restrictions.

Transversal:
CT3. Efficient oral and written communication. Communicate in an oral and written way with other people about the results of learning, thinking and decision making; Participate in debates on topics of the specialty itself.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.
CT8. (ENG) Perspectiva de gènere. Conèixer i comprendre, des del propi àmbit de la titulació, les desigualtats per raó de sexe i gènere a la societat; Integrar les diferents necessitats i preferències per raó de sexe i de gènere en el disseny de solucions i resolució de problemes.

Basic:
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

TEACHING METHODOLOGY

The 12 topics suggested will be developed in 12 theoretical class sessions (2 hours per week) with their respective practices or laboratory (also 2 hours per week). The 3 sessions that are missing from the 15 sessions per semester established in the FIB, will be used for theoretical evaluations (quiz or similar) and practical evaluations (defense of practical work in the middle of the semester and at the end of the semester), remembering also that there are a couple of weeks where there are no lectures to be a week of partial exams and/or final exams, during which advice, support and guidance can be offered to students as reinforcement or preparation for their assessments.

LEARNING OBJECTIVES OF THE SUBJECT

1.Familiarize yourself with the tools and techniques of advanced data analysis to be able to treat data correctly and internalize the data and information obtained as a source of support for decision-making processes.
2.Select, treat and adapt the relevant data to support a specific question.
3.Perform advanced data preprocessing
4.Obtain profiles or patterns from mixed databases from advanced clustering techniques and interpret the results with profiling and post-processing tools
5.Apply multivariate data analysis, especially to categorical data, mixed data and unstructured data
6.Treat semi or unstructured data type text for text mining, sentiment analysis and Topic Modelling
7.Analyze spatiotemporal data. Model data or problems with latent variables.
8.Build the statistical models correctly from the data the context of the reference problem and present it publicly.
9.Develop practical work and projects with a gender perspective
10.Integrate teamwork mechanisms in the performance of practical work.
11.Handle with skill the computer tools necessary to solve the real problems raised with the techniques seen in class
12.Interpret and contextualize the models built from data
13.Validate the models obtained and make a critical interpretation of the results from a technical point of view, contextualizing the results in the framework, reference or understanding of the problem addressed
14.Make a report or final report with the practical assignments or subject project
15.Publicly present a report with the results of the project or practical assignment of the subject

STUDY LOAD

Type	Hours	Percentage
Self study	90,0	60.00
Hours small group	30,0	20.00
Hours large group	30,0	20.00

Total learning time: 150 h

CONTENTS

Introduction

Description:
Data quality, Importance of Data Preprocessing, Introduction to advanced data analysis techniques, Relationship between Multivariate Analysis, Automatic Learning and data science

Preprocessing

Description:
Data acquisition and homogenization, Selection of variables (feature Selection, feature weighting and reduction of variables), Lost data: MICE, MIMMI, Derivation of variables, Transformation of variables, Anomalous Dades (outliers)

Advanced Clustering methods

Description:
Scalability: CURE strategy, Mixed distances and metrics, Ontology-based distances, Clustering on mixed data, DBSCAN, OPTICS, Time series classification

Multiple correspondence analysis and multiple factorial analysis

Description:
CMA

Data analysis - spatiotemporal models

Description:
Conceptes bàsics, dades geolocalitzades, distància geodèsica, components dels models espai-temporals i mètodes bàsics

Text mining

Description:
Sentiment Analysis, Latent Semantic Analysis, Topic Modelling

Modeling based on latent variables

Description:
Modeling based on latent variables

ACTIVITIES

Teamwork

Description:
The students organize themselves into groups and look for real data that meet certain requirements set by the teacher. They use them to apply the techniques and methodologies that are seen throughout the course. At the end, they present a report with the results and make an oral presentation with the most relevant results of the study.

Specific objectives:
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

Related competencies :
CG4. Reasoning, analyzing reality and designing algorithms and formulations that model it. To identify problems and construct valid algorithmic or mathematical solutions, eventually new, integrating the necessary multidisciplinary knowledge, evaluating different alternatives with a critical spirit, justifying the decisions taken, interpreting and synthesizing the results in the context of the application domain and establishing methodological generalizations based on specific applications.
CG9. To face new challenges with a broad vision of the possibilities of a professional career in the field of Artificial Intelligence. Develop the activity applying quality criteria and continuous improvement, and act rigorously in professional development. Adapt to organizational or technological changes. Work in situations of lack of information and / or with time and / or resource restrictions.
CG8. Perform an ethical exercise of the profession in all its facets, applying ethical criteria in the design of systems, algorithms, experiments, use of data, in accordance with the ethical systems recommended by national and international organizations, with special emphasis on security, robustness , privacy, transparency, traceability, prevention of bias (race, gender, religion, territory, etc.) and respect for human rights.
CE18. To acquire and develop computational learning techniques and to design and implement applications and systems that use them, including those dedicated to the automatic extraction of information and knowledge from large volumes of data.
CE09. To ideate, design and integrate intelligent data analysis systems with their application in production and service environments.
CE20. To select and put to use techniques of statistical modeling and data analysis, assessing the quality of the models, validating and interpreting.
CE17. To develop and evaluate interactive systems and presentation of complex information and its application to solving human-computer and human-robot interaction design problems.
CT8. (ENG) Perspectiva de gènere. Conèixer i comprendre, des del propi àmbit de la titulació, les desigualtats per raó de sexe i gènere a la societat; Integrar les diferents necessitats i preferències per raó de sexe i de gènere en el disseny de solucions i resolució de problemes.
CT3. Efficient oral and written communication. Communicate in an oral and written way with other people about the results of learning, thinking and decision making; Participate in debates on topics of the specialty itself.
CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 78h
Laboratory classes: 28h
Self study: 50h

Initial presentation of the practical work

Description:
Initial presentation of the practical work

Specific objectives:
2, 3, 4, 5, 6, 9, 14, 15

Related competencies :
CG4. Reasoning, analyzing reality and designing algorithms and formulations that model it. To identify problems and construct valid algorithmic or mathematical solutions, eventually new, integrating the necessary multidisciplinary knowledge, evaluating different alternatives with a critical spirit, justifying the decisions taken, interpreting and synthesizing the results in the context of the application domain and establishing methodological generalizations based on specific applications.
CG9. To face new challenges with a broad vision of the possibilities of a professional career in the field of Artificial Intelligence. Develop the activity applying quality criteria and continuous improvement, and act rigorously in professional development. Adapt to organizational or technological changes. Work in situations of lack of information and / or with time and / or resource restrictions.
CG8. Perform an ethical exercise of the profession in all its facets, applying ethical criteria in the design of systems, algorithms, experiments, use of data, in accordance with the ethical systems recommended by national and international organizations, with special emphasis on security, robustness , privacy, transparency, traceability, prevention of bias (race, gender, religion, territory, etc.) and respect for human rights.
CE18. To acquire and develop computational learning techniques and to design and implement applications and systems that use them, including those dedicated to the automatic extraction of information and knowledge from large volumes of data.
CE09. To ideate, design and integrate intelligent data analysis systems with their application in production and service environments.
CE20. To select and put to use techniques of statistical modeling and data analysis, assessing the quality of the models, validating and interpreting.
CE17. To develop and evaluate interactive systems and presentation of complex information and its application to solving human-computer and human-robot interaction design problems.
CT8. (ENG) Perspectiva de gènere. Conèixer i comprendre, des del propi àmbit de la titulació, les desigualtats per raó de sexe i gènere a la societat; Integrar les diferents necessitats i preferències per raó de sexe i de gènere en el disseny de solucions i resolució de problemes.
CT3. Efficient oral and written communication. Communicate in an oral and written way with other people about the results of learning, thinking and decision making; Participate in debates on topics of the specialty itself.
CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 6h
Laboratory classes: 2h
Self study: 4h

Quiz 1

Description:
Quiz 1

Specific objectives:
2, 3

Related competencies :
CG4. Reasoning, analyzing reality and designing algorithms and formulations that model it. To identify problems and construct valid algorithmic or mathematical solutions, eventually new, integrating the necessary multidisciplinary knowledge, evaluating different alternatives with a critical spirit, justifying the decisions taken, interpreting and synthesizing the results in the context of the application domain and establishing methodological generalizations based on specific applications.
CG8. Perform an ethical exercise of the profession in all its facets, applying ethical criteria in the design of systems, algorithms, experiments, use of data, in accordance with the ethical systems recommended by national and international organizations, with special emphasis on security, robustness , privacy, transparency, traceability, prevention of bias (race, gender, religion, territory, etc.) and respect for human rights.
CE09. To ideate, design and integrate intelligent data analysis systems with their application in production and service environments.
CE20. To select and put to use techniques of statistical modeling and data analysis, assessing the quality of the models, validating and interpreting.
CE17. To develop and evaluate interactive systems and presentation of complex information and its application to solving human-computer and human-robot interaction design problems.
CT8. (ENG) Perspectiva de gènere. Conèixer i comprendre, des del propi àmbit de la titulació, les desigualtats per raó de sexe i gènere a la societat; Integrar les diferents necessitats i preferències per raó de sexe i de gènere en el disseny de solucions i resolució de problemes.
CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.

Theory classes of the subject syllabus

Description:
Theory classes of the subject syllabus

Specific objectives:
2, 3, 4, 5, 6, 7, 8

Related competencies :
CG4. Reasoning, analyzing reality and designing algorithms and formulations that model it. To identify problems and construct valid algorithmic or mathematical solutions, eventually new, integrating the necessary multidisciplinary knowledge, evaluating different alternatives with a critical spirit, justifying the decisions taken, interpreting and synthesizing the results in the context of the application domain and establishing methodological generalizations based on specific applications.
CG8. Perform an ethical exercise of the profession in all its facets, applying ethical criteria in the design of systems, algorithms, experiments, use of data, in accordance with the ethical systems recommended by national and international organizations, with special emphasis on security, robustness , privacy, transparency, traceability, prevention of bias (race, gender, religion, territory, etc.) and respect for human rights.
CE18. To acquire and develop computational learning techniques and to design and implement applications and systems that use them, including those dedicated to the automatic extraction of information and knowledge from large volumes of data.
CE09. To ideate, design and integrate intelligent data analysis systems with their application in production and service environments.
CE20. To select and put to use techniques of statistical modeling and data analysis, assessing the quality of the models, validating and interpreting.
CE17. To develop and evaluate interactive systems and presentation of complex information and its application to solving human-computer and human-robot interaction design problems.
CT8. (ENG) Perspectiva de gènere. Conèixer i comprendre, des del propi àmbit de la titulació, les desigualtats per raó de sexe i gènere a la societat; Integrar les diferents necessitats i preferències per raó de sexe i de gènere en el disseny de solucions i resolució de problemes.
CT3. Efficient oral and written communication. Communicate in an oral and written way with other people about the results of learning, thinking and decision making; Participate in debates on topics of the specialty itself.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Full-or-part-time: 60h
Theory classes: 30h
Self study: 30h

Quiz 2

Description:
During the course there will be short answer tests to set learning pieces. It will be done at the end of certain laboratory classes

Specific objectives:
4, 5, 8

Related competencies :
CG4. Reasoning, analyzing reality and designing algorithms and formulations that model it. To identify problems and construct valid algorithmic or mathematical solutions, eventually new, integrating the necessary multidisciplinary knowledge, evaluating different alternatives with a critical spirit, justifying the decisions taken, interpreting and synthesizing the results in the context of the application domain and establishing methodological generalizations based on specific applications.
CG8. Perform an ethical exercise of the profession in all its facets, applying ethical criteria in the design of systems, algorithms, experiments, use of data, in accordance with the ethical systems recommended by national and international organizations, with special emphasis on security, robustness , privacy, transparency, traceability, prevention of bias (race, gender, religion, territory, etc.) and respect for human rights.
CE09. To ideate, design and integrate intelligent data analysis systems with their application in production and service environments.
CE20. To select and put to use techniques of statistical modeling and data analysis, assessing the quality of the models, validating and interpreting.
CT3. Efficient oral and written communication. Communicate in an oral and written way with other people about the results of learning, thinking and decision making; Participate in debates on topics of the specialty itself.
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of ??study.
CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

Practical work presentation

Description:
Practical work presentation

Specific objectives:
14, 15

Related competencies :
CG4. Reasoning, analyzing reality and designing algorithms and formulations that model it. To identify problems and construct valid algorithmic or mathematical solutions, eventually new, integrating the necessary multidisciplinary knowledge, evaluating different alternatives with a critical spirit, justifying the decisions taken, interpreting and synthesizing the results in the context of the application domain and establishing methodological generalizations based on specific applications.
CG9. To face new challenges with a broad vision of the possibilities of a professional career in the field of Artificial Intelligence. Develop the activity applying quality criteria and continuous improvement, and act rigorously in professional development. Adapt to organizational or technological changes. Work in situations of lack of information and / or with time and / or resource restrictions.
CG8. Perform an ethical exercise of the profession in all its facets, applying ethical criteria in the design of systems, algorithms, experiments, use of data, in accordance with the ethical systems recommended by national and international organizations, with special emphasis on security, robustness , privacy, transparency, traceability, prevention of bias (race, gender, religion, territory, etc.) and respect for human rights.
CE17. To develop and evaluate interactive systems and presentation of complex information and its application to solving human-computer and human-robot interaction design problems.
CT8. (ENG) Perspectiva de gènere. Conèixer i comprendre, des del propi àmbit de la titulació, les desigualtats per raó de sexe i gènere a la societat; Integrar les diferents necessitats i preferències per raó de sexe i de gènere en el disseny de solucions i resolució de problemes.
CT3. Efficient oral and written communication. Communicate in an oral and written way with other people about the results of learning, thinking and decision making; Participate in debates on topics of the specialty itself.
CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.

Full-or-part-time: 6h
Self study: 6h

Quiz 3

Quiz 4

Description:
During the course there will be short answer tests to set learning pieces. It will be done at the end of certain laboratory classes

GRADING SYSTEM

Propose the following evaluation system:
- Workgroup realized at the end of the course 20%.
- Oral test of knowledge control 10% (discussion between the teacher and the oral presentation of the work in the team).
- Quality and performance of the work team. 10%
- Oral and written communication 10%.
- Ethics of the treball and treball team propiment dit 10%
- Gender perspective of the team and the treball 10%.
- Attendance and participation in classes and laboratories. 10%
- 4 Quiz at the end of the course 20%.

Reassessment

Only those students who had previously taken the final exam and failed it can take the reassessment exam.

BIBLIOGRAPHY

Basic:
- Gibert, Karina; Sànchez-Marré, Mquel; Izquierdo, Joaquin. "A survey on pre-processing techniques: Relevant issues in the context of environmental data Mining". AI communications: the european journal of artificial intelligence [on line]. Desembre 2016, vol. 29, núm. 6, p. 627-663 [Consultation: 15/03/2023]. Available on: https://upcommons.upc.edu/handle/2117/123530.- Angerri, X., & Gibert, K. "Preprocessing and Artificial Intelligence for Increasing Explainability in Mental Health". International Journal on Artificial Intelligence Tools [on line]. [Consultation: 15/03/2023]. Available on: https://www.worldscientific.com/doi/abs/10.1142/S0218213023400110.- Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. The Elements of statistical learning : data mining, inference, and prediction. 2nd ed. New York: Springer, cop. 2009. ISBN 9780387952840.
- Husson, François; Lê, Sébastien; Pagès, Jérôme. Exploratory multivariate analysis by example using R. Second edition. CRC Press, Taylor & Francis Group, 2017. ISBN 9781315301860.
- Johnson, Richard A; Wichern, Dean W. Applied multivariate statistical analysis. Sixth edition. Harlow, Essex: Pearson Education Limited, [2014]. ISBN 9781292024943.
- Agresti, Alan; Franklin, Christine. Statistics: the art and science of learning from data. 4th ed. Harlow: Pearson Education, 2018. ISBN 9781292164779.
- Bruce, Peter; Bruce, Andrew; Gedeck, Peter. Practical statistics for data scientists: 50+ essential concepts using R and Python. 2nd ed. Beijing: O'Reilly, [2020]. ISBN 9781492072942.

Complementary:
- Peña, Daniel. Análisis de datos multivariantes. Madrid: McGraw-Hill, cop. 2002. ISBN 9788448136109.
- Husson, François; Lê, Sébastien; Pagès, Jérôme. Exploratory multivariate analysis by example using R. 2nd ed. Boca Raton: CRC Press, Taylor & Francis, 2017. ISBN 9781315301860.
- Greenacre, Michael. Correspondence Analysis in Practice. 3th ed. Chapman and Hall/CRC, 2016. ISBN 9781315369983.

RESOURCES

Hyperlink:
- https://www-eio.upc.edu/teaching/DocenciaMultivariant/