Course Title	**Data Mining**				
Course Code	**DLWSS551**				
Course Type	**Elective**				
Level	Master (2nd Cycle) – Distance Learning				
Year / Semester	1 / 2				
Teacher's Name	**Dr Harris Papadopoulos, Prof. Leonidas Anthopoulos**				
ECTS	10	Lectures / week	3	Laboratories / week	0
Course Purpose	The computerization of modern society facilitated the generation and collection of an ever-increasing volume of raw data. The World Wide Web contains about a trillion web pages! The Large Hadron Collider computers at CERN record about 15 petabytes a year! The amount of data stored in the world's databases doubles every 20 months! These are just a few of the many facts that demonstrate the explosive growth in the quantity of data being recorded. These astonishing amounts of data contain a great deal of potentially important and valuable information waiting to be discovered. The role of Data Mining (DM) is to make this discovery possible. In particular, DM is about automatically extracting implicit, previously unknown and potentially useful information from substantial quantities of data. The constant increase in our ability to record data together with the significant benefits resulting from the application of DM in a variety of diverse areas have made it one of the most promising and flourishing fields of Computer Science. This course will introduce students to the major principles, terminology, problem types and research topics of DM. The main technical basis for DM comes from Machine Learning (ML), which is used for acquiring structural descriptions from data. These descriptions can either be used to predict the outcome of a new situation, or for the explanation and understanding of the behaviour of the source of the data. The course will examine the main ideas behind some of the leading ML techniques being used in practical DM and the issues that should be considered in their application. Additionally, it will consist of hands-on experimentation on real life problems and data to promote the development of critical thinking on the proper application of ML techniques. The course aims to provide you with the knowledge of the essential tools and techniques to: • identify situations in which the application of Data Mining is required or beneficial and • conduct a Data Mining investigation of a practical Machine Learning problem and critically analyse and evaluate the results.					
Learning Outcomes	By the end of the course the students are expected to: • Define and explain the major principles, terminology and problem types of				

<table>
<tr><td></td><td colspan="3">Data Mining
• Describe and discuss the main Machine Learning techniques used in practical Data Mining and their theoretical basis and evaluate their strengths and weaknesses
• Explain and propose ways of dealing with the issues involved in the application of Machine Learning techniques to practical problems
• Apply Machine Learning techniques to a practical problem both in an exploratory or a targeted manner
• Analyse and evaluate the performance of Machine Learning techniques on a supervised Data Mining task
• Define and apply the main data transformation approaches used in practical Data Mining
• Define and explain the main concepts and terminology of Web Mining
• Define, explain and demonstrate the main concepts, approaches and issues for designing a recommendation system
• Describe and explain the two main versions of the Conformal Prediction framework for quantifying uncertainty both in classification and regression and evaluate their outputs</td></tr>
<tr><td>Prerequisites</td><td>None</td><td>Corequisites</td><td>None</td></tr>
<tr><td>Course Content</td><td colspan="3">This course consists of the following ten chapters:
• Chapter 1 (Week 1) is the introductory chapter for the whole course
• Chapter 2 (Week 2) introduces the main components and terminology of a Data Mining task
• Chapter 3 (Weeks 3 & 4) analyses the main ideas behind some of the leading techniques that are used in practical Data Mining
• Chapter 4 (Week 5) deals with the evaluation and comparison of Machine Learning techniques
• Chapter 5 (Weeks 6 & 7) examines some of the most prominent advanced Machine Learning techniques used in practice today
• Chapter 6 (Week 8) introduces the unsupervised learning setting and various approaches for this setting leading to different kinds of representations
• Chapter 7 (Week 9) studies data engineering approaches for transforming the input and output to a suitable or even more effective form
• Chapter 8 (Week 10) examines Web Mining and its three types: Web Content Mining, Web Structure Mining and Web Usage Mining
• Chapter 9 (Week 11) deals with recommendation systems and analyzes the core concepts behind the two main types of recommendation systems: content-based and collaborative filtering
• Chapter 10 (Week 12) introduces a recently developed framework for quantifying the uncertainty of Machine Learning predictions, called Conformal Prediction</td></tr>
<tr><td>Teaching</td><td colspan="3">Mode of Delivery: Distance Learning</td></tr>
</table>

Methodology	The course is designed to introduce and explain the material students are expected to learn through an on-line learning environment. The on-line environment provides an opportunity for receiving on-line feedback from the Course Instructor during their study. In addition, students will be encouraged to interact both with other students and the instructor so as to feel part of an on-line community of learners that belong to the University network. The course content will be delivered through online material/notes, recorded lectures and/or narrated presentations. Therefore, students may be asked to download and study notes, tutorials and numerical exercises as well as watch recorded lectures/demonstrations or narrated presentations posted on the web, addressing the main concepts of a particular unit. Furthermore, the planned communication and the dynamic/online interaction activities between the course instructor and the students will include asynchronous communication tools (Discussion Forum) where students may be asked to participate, wherever appropriate, in an online forum posting their views on certain topics covered in a particular unit; and synchronous communication tools (instant messaging, such as Skype, chat rooms, video-conferencing, etc.), so that students may discuss on-line with the Instructor(s) and/or other students specific issues covered in a given unit.
Bibliography	• Ian H. Witten, Eibe Frank and Mark A. Hall (2011), Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufman, 3rd ed. • Bing Liu (2011), Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer-Verlag, 2nd ed. • Jure Leskovec, Anand Rajaraman and Jeff Ullman (2020), Mining of Massive Datasets, Cambridge University Press, 3rd ed. • H. Papadopoulos. “Inductive Conformal Prediction: Theory and Application to Neural Networks”. Tools in Artificial Intelligence, Chapter 18, 315–330. I Tech, Vienna, Austria, 2008. DOI: 10.5772/6078 (Open Access) • Glenn Shafer and Vladimir Vovk. “A Tutorial on Conformal Prediction”. Journal of Machine Learning Research (JMLR), vol. 9, No. 12, 371−421, 2008. (Open Access) • H. Papadopoulos, V. Vovk, A. Gammerman. “Regression Conformal Prediction with Nearest Neighbours”. Journal of Artificial Intelligence Research, vol. 40, 815–840. 2011. DOI: 10.1613/jair.3198. (Open Access) • V. Vovk, V. Fedorova, I. Nouretdinov and A. Gammerman A. “Criteria of Efficiency for Conformal Prediction”. In Conformal and Probabilistic Prediction with Applications. COPA 2016. Lecture Notes in Computer Science, vol. 9653. Springer. 2016. Additional / Complimentary Bibliography: • Gregory Piatetsky-Shapiro (2007). Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from “university” to “business” and “analytics”. Data Mining and Knowledge Discovery, vol. 15, no. 1, pp. 99-105. Springer-Verlag. • Nada Lavrač, Hiroshi Motoda, Tom Fawcett, Robert Holte, Pat Langley

	and Pieter Adriaans (2004). “Introduction: Lessons Learned from Data Mining Applications and Collaborative Problem Solving”, Machine Learning, vol. 57, no. 1, pp. 13-34. Springer-Verlag. • H. Papadopoulos and H. Haralambous. “Reliable Prediction Intervals with Regression Neural Networks”. Neural Networks, vol. 24, no. 8, 842–851. Elsevier, 2011. DOI: 10.1016/j.neunet.2011.05.008. • H. Papadopoulos, N. Georgiou, C. Eliades and A. Konstantinidis. “Android Malware Detection with Unbiased Confidence Guarantees”. Neurocomputing, Vol. 280. Elsevier, 2018. DOI: 10.1016/j.neucom.2017.08.072. • Lambrou, H. Papadopoulos, A. Gammerman. “Reliable Confidence Measures for Medical Diagnosis with Evolutionary Algorithms”. IEEE Transactions on Information Technology in Biomedicine, vol. 15, no. 1, 93–99. IEEE, 2011. DOI: 10.1109/TITB.2010.2091144. • H. Papadopoulos, V. Vovk and A. Gammerman. “Special Issue on Conformal Prediction and its Applications”. Annals of Mathematics and Artificial Intelligence, vol. 74, no. 1–2, Springer, 2015. • Gammerman, V. Vovk, H. Boström and L. Carlsson. “Conformal and probabilistic prediction with applications”. Machine Learning, vol. 108, Springer, 2019.
Assessment	The Students are assessed via continuous assessment throughout the duration of the Semester, which forms the Coursework grade, and the final written exam. The coursework and the final exam grades are weighted 50% and 50%, respectively, and compose the final grade of the course. The assessment weight, date and time of each type of continuous assessment is being set at the beginning of the semester via the course outline. An indicative weighted continuous assessment of the course is shown below: • **Two marked assignments/projects** (40% of total marks for module) • **Two dynamic interactive activities** (10% of total marks for module) • **One closed-book exam** (50% of total marks for module)
Language	English