Big Data Analytics

Session 4

- Online Session -

2019 - 2020

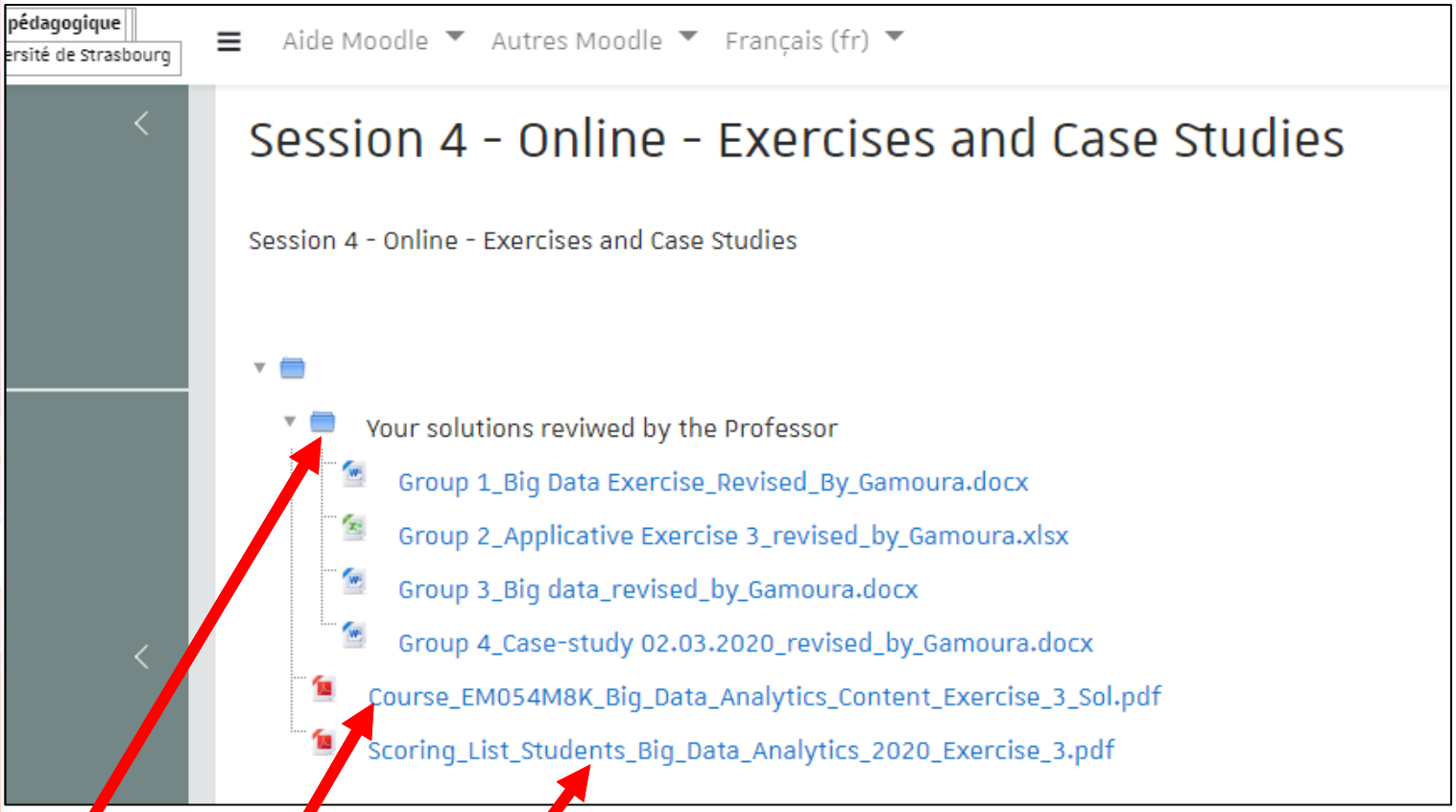
By **PhD. Samia Chehbi Gamoura**
Associate Professor – EM Strasbourg
France

Dear students welcome to my online MOOC platform. I am the avatar of your professor Samia Gamoura of this course, entitled “Big Data Analytics.“ This online session is also available on the website of your professor: www.samiagamoura.com. The last work provided by your professor was Exercise 3 about Machine Learning algorithms in Business cases. Your professor gave you enough time to work on the solution in the previous session. Some of you already sent me their complete exercises. Thanks for everyone.

Big Data Analytics → Advanced Analytics → Machine Learning

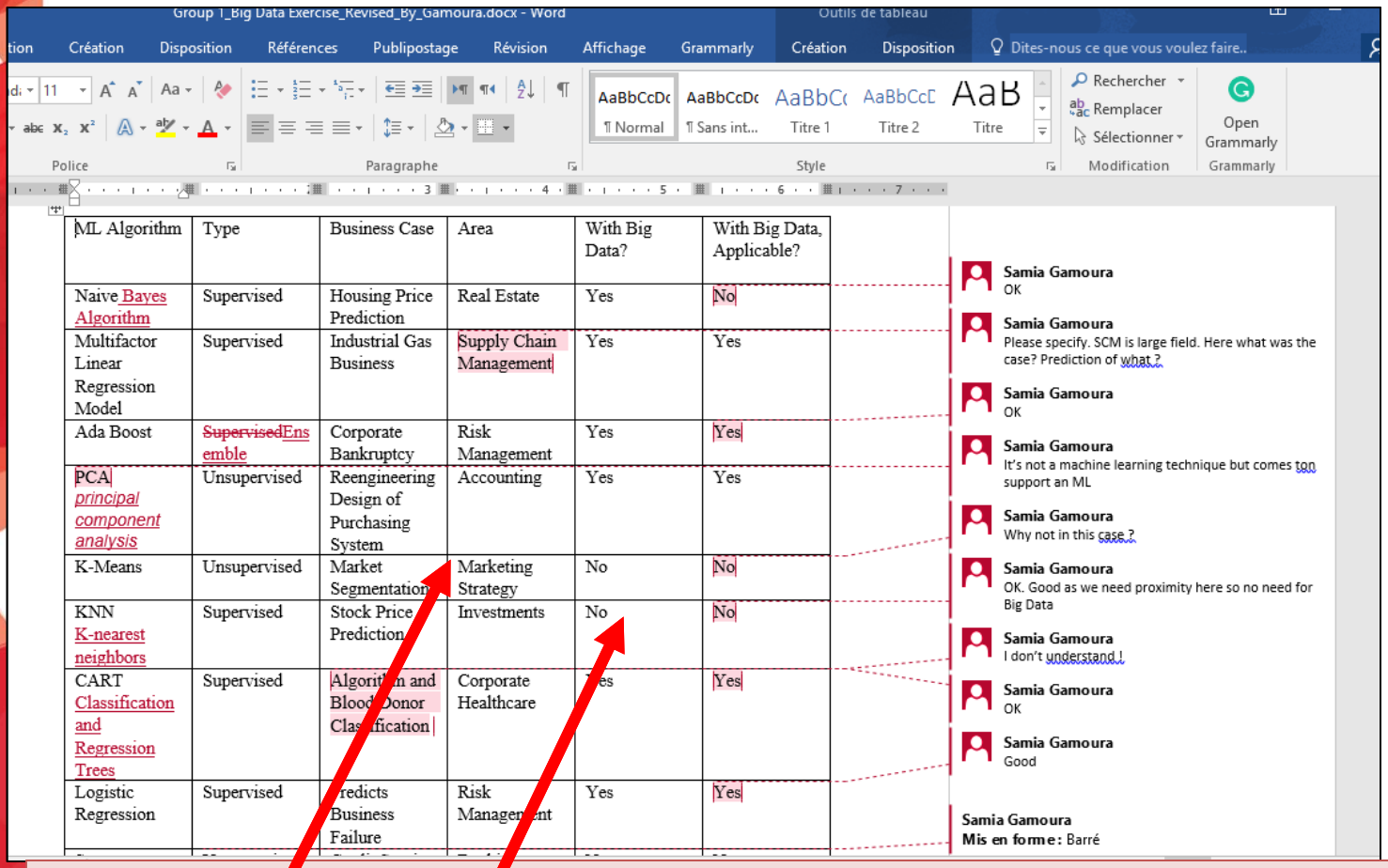
ML Algorithm	Type	Business Case	Area	With Big Data?	With Big Data, Applicable?
Association Rule	Unsupervised	Market basket Analysis	E-marketing	No	Yes
Naive Bayes Algorithm	Supervised	Housing Price Prediction	Real Estate	Yes	No
AdaBoost	Ensemble	Corporate Bankruptcy	Risk Management	Yes	Yes
K-Means Clustering	Unsupervised	Market Segmentation	Marketing Strategy	No	Yes
Logistic Regression	Supervised	Predicts Business Failure	Risk Management	Yes	Yes
Artificial Neural Network	Supervised	Tourism Demand Forecasts	Tourism and hospitality industry	Yes	Yes
Decision tree	Supervised	Financial crisis prediction	Financial industry and budgeting	Yes	Yes
Version space learning	Supervised	Recommender system	Marketing	Yes	No
Rough set rules	Unsupervised	Product classification	Production in Supply chain management	Yes	Yes
K-Means	Unsupervised	Stock market forecasting	Finance, investments	Yes	Yes

Please check this list of the selected correct answers in your solutions. Don't worry, you can upload this solution afterward. You can find it in PDF format in Moodle in the Exercises part, as you can see in the screenshot in the subsequent slide.



Only 68% of students provided their solutions. All of them have +0.5 point as an additional reward in the practical part of the exam. To check the list of names, please report to the PDF file of scoring uploaded in the Exercises part in Moodle. Your professor reviewed all your solutions and uploaded them in a separated directory as you can see in the screenshot here.

Big Data Analytics → Advanced Analytics → Machine Learning



When you download and open your file in your given solution, you find the comments and corrections of your professor. Please get the corrections and accept or reject the modification by using the track tool in your word file. For the Excel files, please open your comments. Now, let's switch to the course to continue with the Machine Learning part.

4 categories (meta – approaches) following the input Data:

- **Information-Based Machine Learning**
- **Erro-Based Machine Learning**
- **Similarity-Based Machine Learning**
- **Probability-Based Machine Learning**

There are four meta-categories if we base the categorization on the input data. You do not need to know all the technical aspects of these categories and algorithms as you do not have the technical background of algorithmic and computer science in this class. However, you need to know, at least, these approaches of Machine Learning are used in Business and management applications in enterprises.

Information-Based Machine Learning

- Shannon measurement theory presented in 1948 by Claude Shannon
- The extraction of a distinctive measurement from the Data Descriptive Features through the information contents.
- Need of use of information content measurement and metrics

There are four categories of Machine Learning meta algorithms if we base the categorization on the input Data. Information-based category of algorithms in Machine Learning are founded on Shannon measurement Theory of 1948. In this class of algorithms, the distinction is based on measurement from descriptive features in the information content. Keep in mind here that you need measurement metrics to perform in this category.

Information-Based Machine Learning

Approach	Types	Sub-Types
Decision Trees (DTR)	Supervised	Classification & Prediction
Conditional Random Fields (CRF)	Supervised	Classification & Prediction
Linear Regression (LRE)	Supervised	Classification & Prediction
Non Linear Regression (NLR)	Supervised	Classification & Prediction
Polynomial regression (PRE)	Supervised	Classification & Prediction
Maximum Entropy (MEN)	Supervised	Classification & Prediction
Frequent Pattern Growth (FRG)	Unsupervised	Association Rule
Apriori Algorithm (AAL)	Unsupervised	Association Rule
Eclat Algorithm (EAL)	Unsupervised	Association Rule
Random Decision Forest (RDF)	Ensemble	Bagging
AdaBoost Algorithm (ABA)	Ensemble	Bagging
LogitBoost Algorithm (LBA)	Ensemble	Bagging

In Business application of Machine Learning. These are some of the well-known ML algorithms in information-based learning like Decision Trees and Linear Regression that are supervised. Others like Eclat, and Apriori are unsupervised. We can find also some Ensemble meta algorithms like Ada Boost and Random Forest.

Error-Based Machine Learning

- Mathematical models that are founded on the idea of getting performance through error minimization
- Need error measure and the error surface (space)

The second category gathers the Error-Based Machine Learning approaches. They perform by computing the error as the measurement of the difference between the desired class and the obtained class. Backpropagation is known as the widely used method of error computing and correcting in the learning model. The algorithms of this category always need the mathematical measurement of the error.

Error-Based Machine Learning

Approach	Types	Sub-types
Artificial Neural Network (ANN)	Supoervised	Classification & prediction
Fisher Linear Discriminant (FLD)	Supoervised	Classification & prediction
Logistic Regression (LRE)	Supervised	Classification & prediction
Support Vector Machine (SVM)	Supoervised	Classification & prediction

Algorithms in Error-Based models could be Support Vector Machines (SVM) and Artificial Neural Networks including a Deep Learning algorithms such as Conv Networks. Error-based algorithms are mainly supervised.

Similarity-Based Machine Learning

- The best way of predicting the future is to simply compare and find similarities of features from the past
- Need techniques of metrics in determining similarities degrees in the space of descriptive features

Similarity-based Machine Learning algorithms are in the third category. They are based on the hypothesis that the best way to predict the future is to simply compare and find similarities with the past. Similarities key features and comparison metrics are then needed for these approaches.

Similarity-Based Machine Learning

Approach	Types	Sub-Types
Case-Bases Reasoning (CBR)	Supervised	Classification & Prediction
Hidden Markov Models (HMM)	Supervised	Classification & Prediction
K-Means (KME)	Unsupervised	Clustering
Mixture Models (MMO)	Unsupervised	Clustering
Hierarchical Cluster Analysis (HCA)	Unsupervised	Clustering
Hebbian Learning (HLE)	Unsupervised	Unsupervised Artificial Neural Networks
Generative Adversarial Networks (GAN)	Unsupervised	Unsupervised Artificial Neural Networks
Sarsa (SAR)	Réinforcement	Model-free Reinforcement
Q-Learning (QLE)	Réinforcement	Model-free Reinforcement
Divergence de Kullback-Leibler (DKL)	Réinforcement	Model-based Reinforcement

In similarity-based approaches, we find the different modes of learning classes: supervised learning with hidden Markov models for example, unsupervised learning with clustering, and reinforcement learning with Q-learning and Sarsa algorithm.

Probability-Based Machine Learning

- Based on Bayes theorem
- The future is a random event based on relative frequencies with calculation of conditional probabilities based on the sequence of present and the past

The fourth category is probability-based learning technique. The key idea in these approaches is the use of Bayes theorem. In Bayes, the future is defined as a random event based on relative frequencies with calculation of conditional probabilities based on the sequence of present and the past.

Probability-Based Machine Learning

Approach	Types	Sub-types
Naive Bayes Classifier (NBC)	Supervised	Classification & prediction
Bayesian Networks (BNE)	Supervised	Classification & prediction

In business problems using probability-based approaches in Machine Learning, we find the majority of use in supervised algorithms such as Naive Bayes classifiers and Bayesian networks in classification.

Case Study 2

Exigency	Mandatory
Type	Individual Work
Mode	Homework
Estimated duration	120 minutes
Evaluation	included in the practical part of the Exam
Repository	**Session 4 - Case Study 2 to do (Mandatory - Part of the exam evaluation)**
Deadline	**20th March**

This online is now finished. Now, you have homework to do. Now. You have this case study number 2 to do. This case is individual. It's a homework to perform by yourself. The estimated duration to do it is about two hours (120 minutes). This work will be evaluated by your professor afterward. The evaluation is included in the practical part of the exam. The deadline to provide this work is 20th March.

Case Study 2

- **Realize a research investigation in literature on one of the** following topics:
  (* please select only one topic in the list):
    1. Chatbot in Marketing
    2. Opinion mining in Customer Relationship Management
    3. Additive Manufacturing
    4. Recommender system in web Marketplace
    5. Social Computing in customer prospection
    6. Augmented product in Sales Management
    7. Business Process Automation

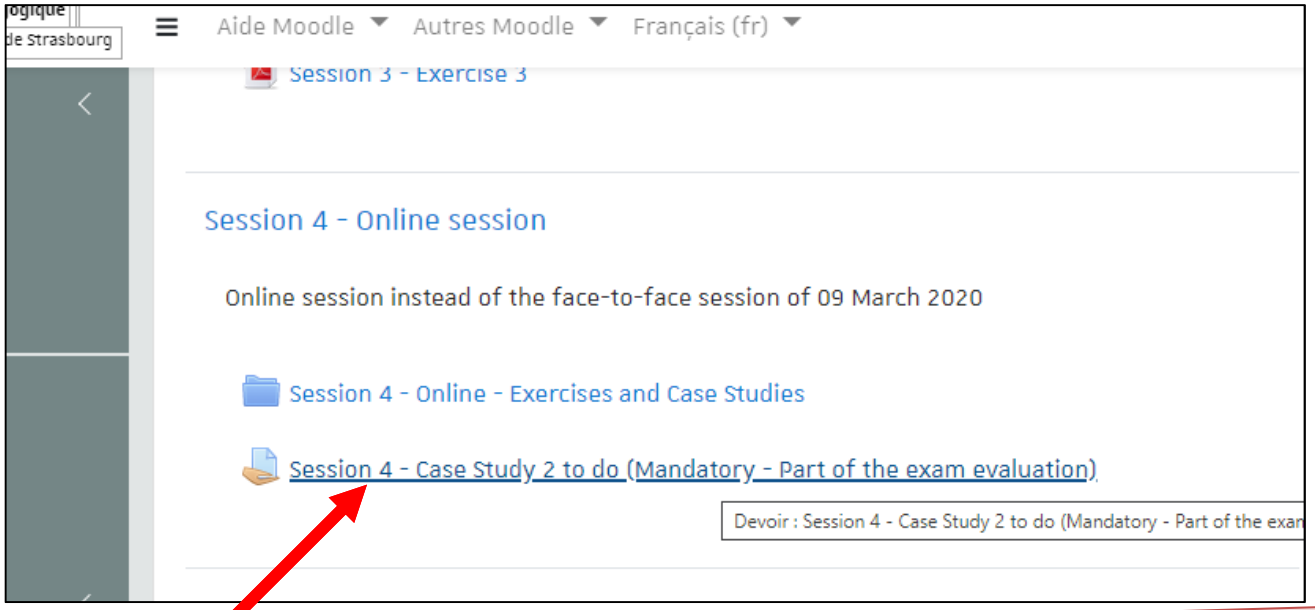
- by following the structure here below.

1. Context and Definitions
2. Examples and cases from literature
3. Your criticism(s) and your opinion,
4. Perspectives of use and conclusion

By using Internet documentation and academic references such as Google Scholar, try to realize a research investigation in literature. You have to select one topic to develop in a report (MS Word for example), by following the structure here below. Content and Definitions. Examples and cases from literature. Your criticism(s) and your opinion. Perspectives of use and conclusion. You have a list of topics proposed by your professor. Each student has to select one topic to develop. There is no issue if two students select the same topic. But this work still individual and each student should provide his own work.

Case Study 2

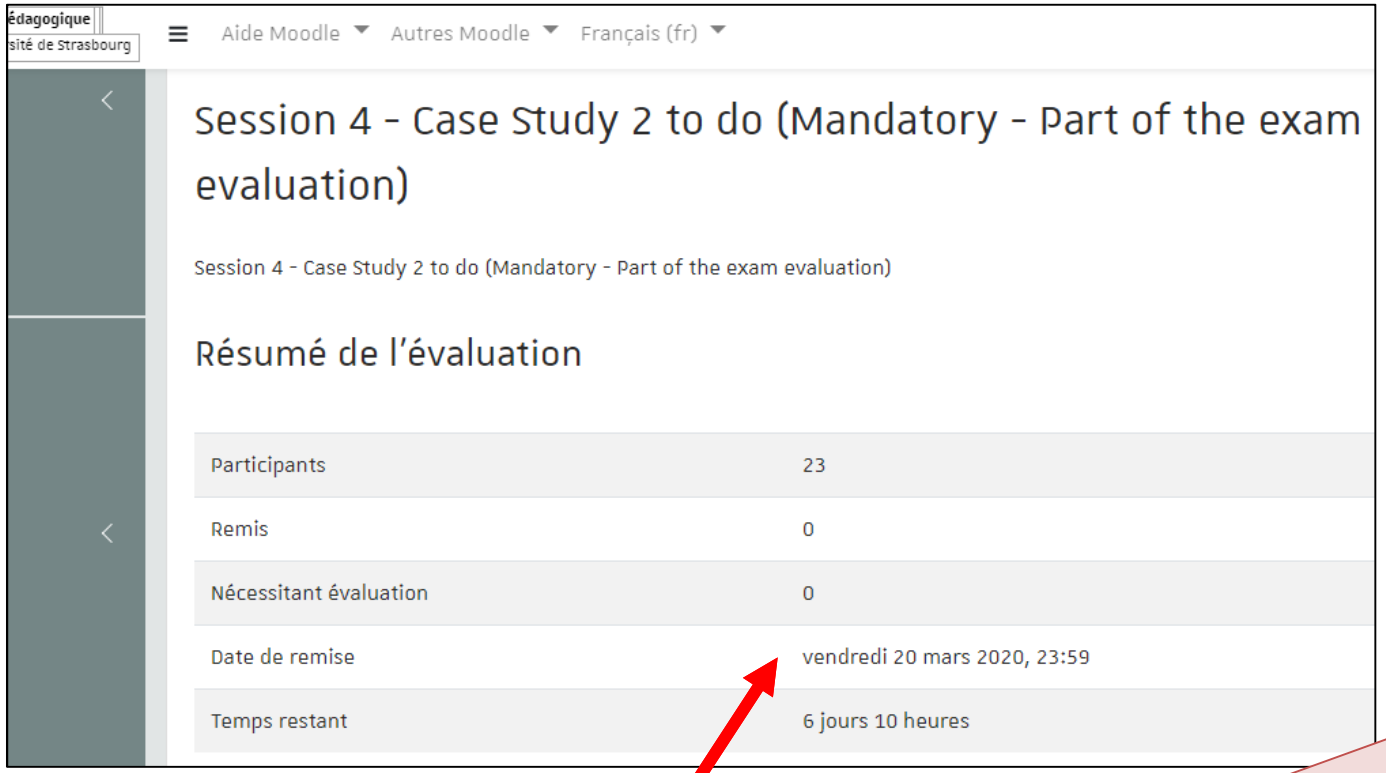
- **Repository: Session 4 - Case Study 2 to do (Mandatory - Part of the exam evaluation)**



As already mentioned, you have to upload your files in the repository named Session 4 – Case Study 2.

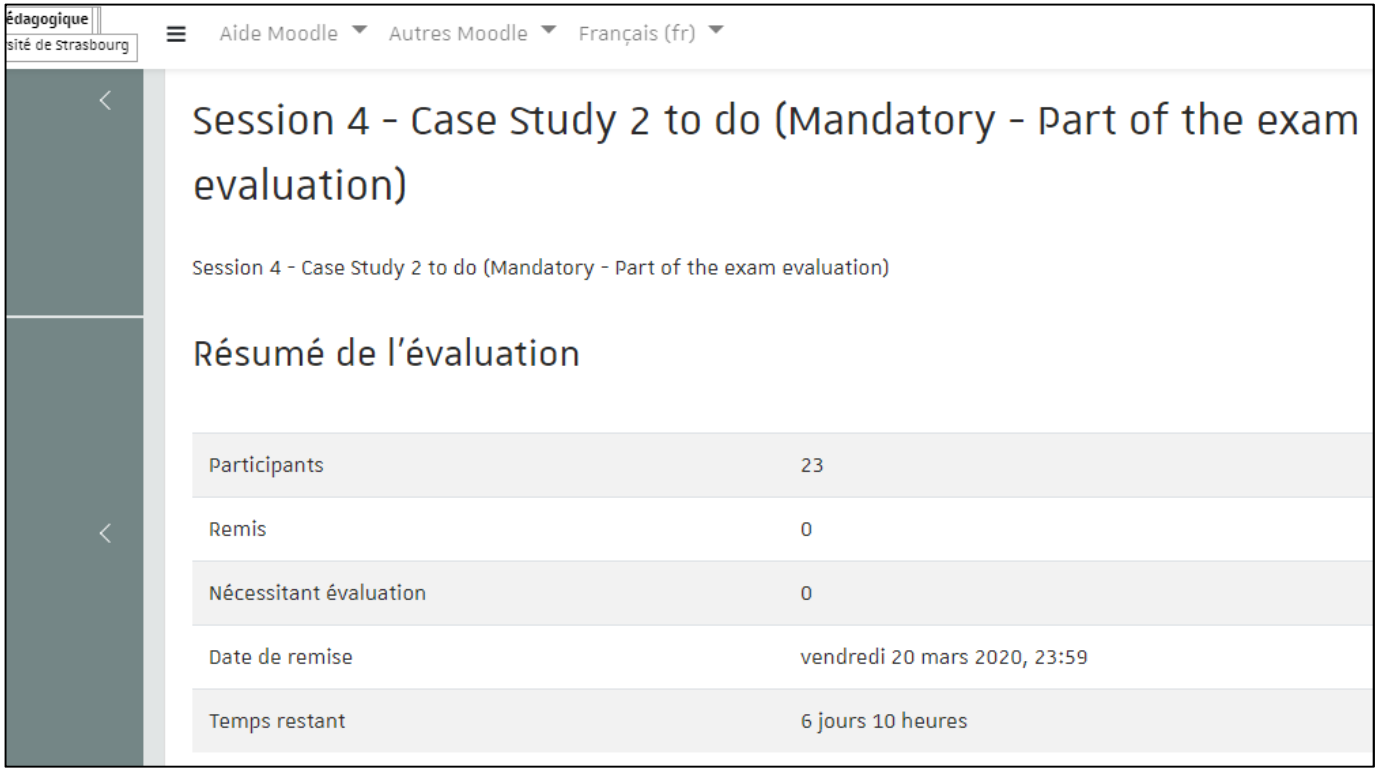
Case Study 2

- **Deadline: 20th March**



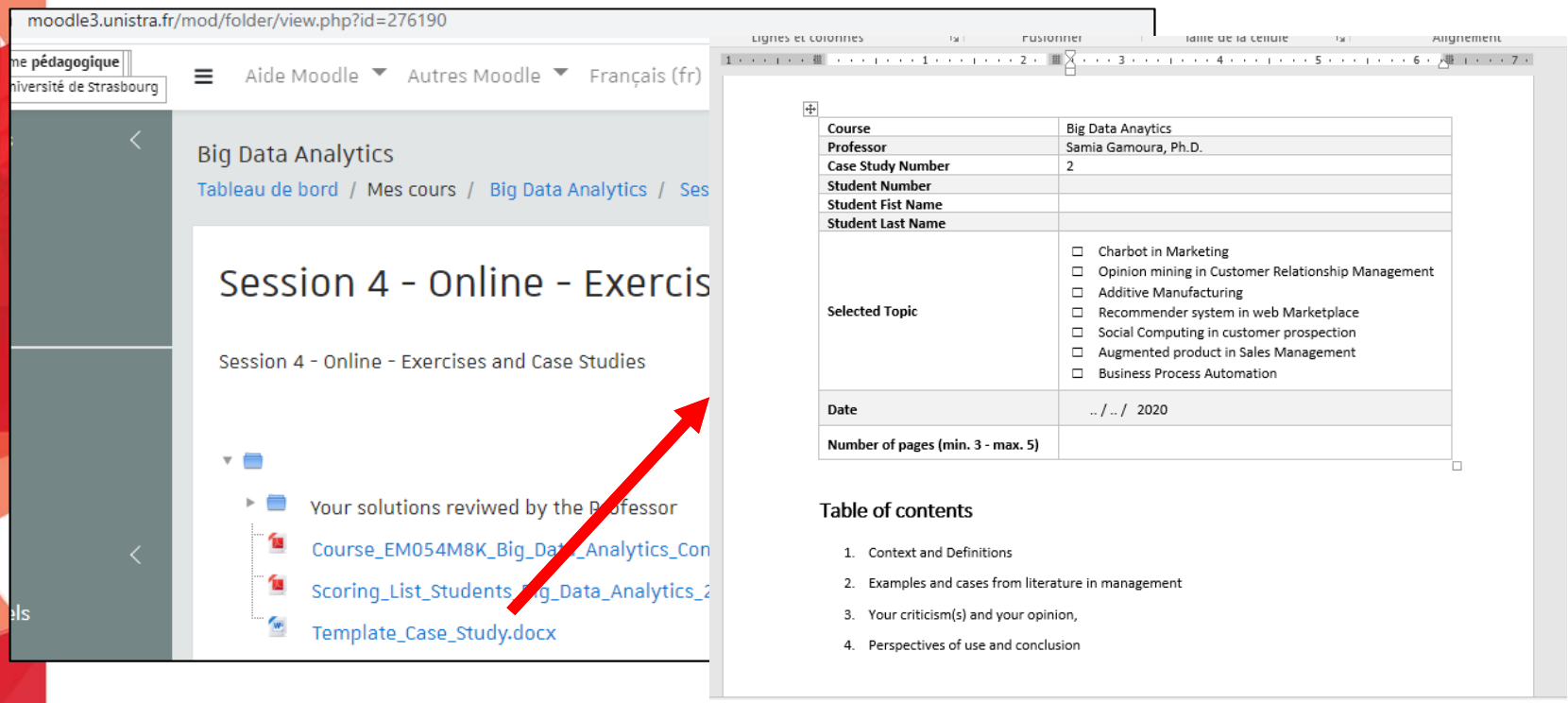
The deadline is Friday 20th March. The system will close the repository automatically when deadline is over. No file will be accepter afterward. Files sent by emails are not accepted.

Case Study 2



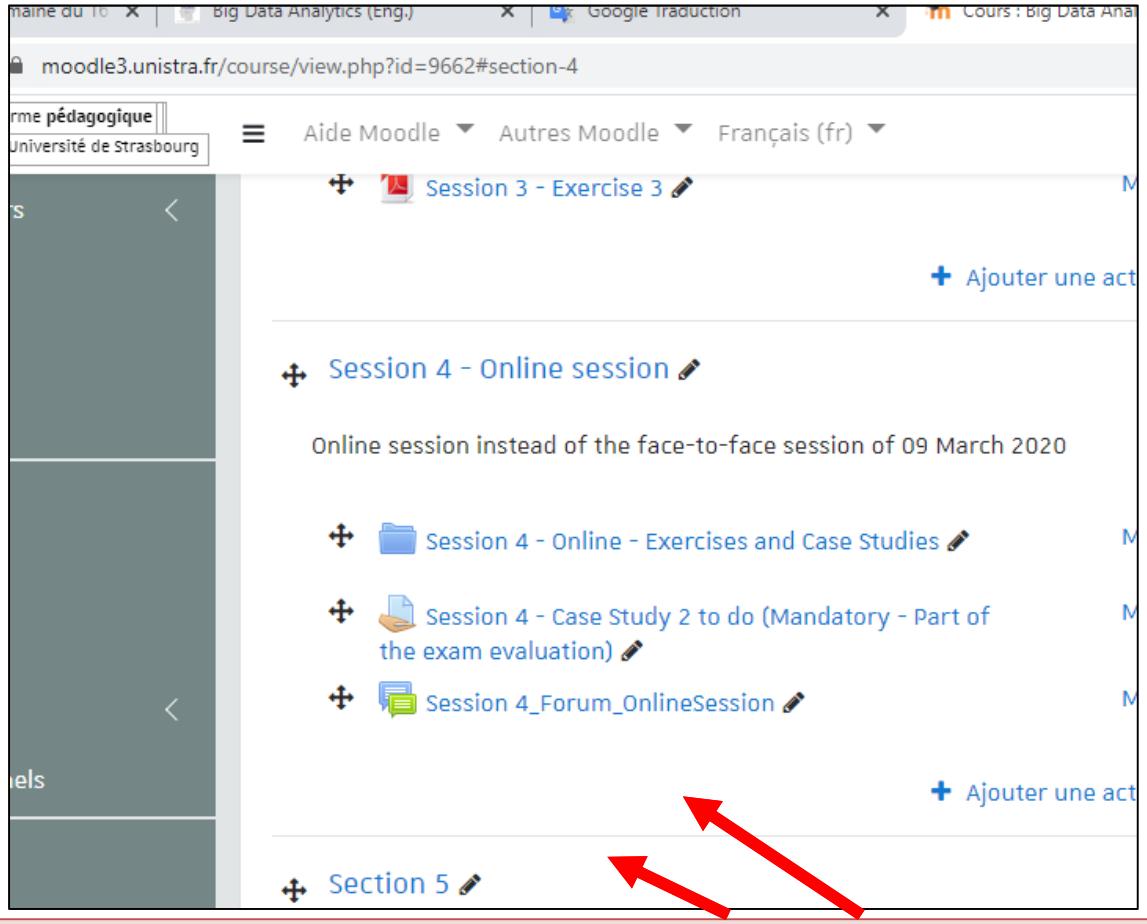
The file you provide must be in Word format (not PDF) and must named with your full name. For example, the file of the student **TOM BRYAN** must have the name: **TOM BRYAN _Case_Study_2.docx (or TOM BRYAN _Case_Study_2.doc)**

Case Study 2



You find also a template word file to use for your report. You have to fill the head and develop the content as required.

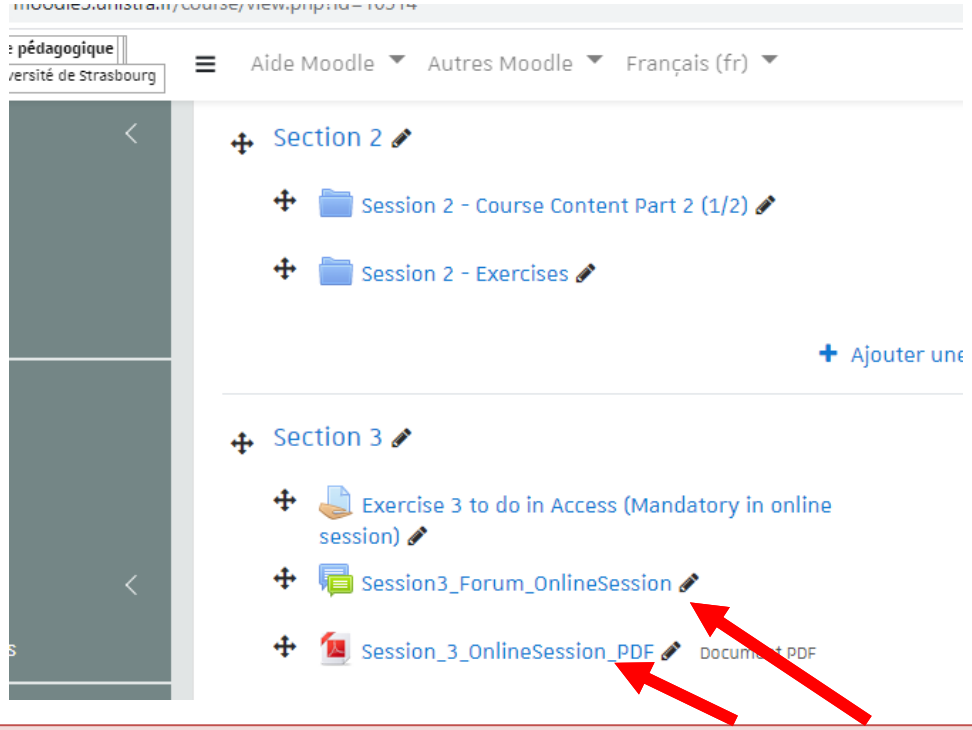
Case Study 2



Now you can download and review the PDF file of this online session in Moodle. If you have any questions, please use the Forum platform in Moodle to follow exchanges and ask your questions. Your professor will exchange with you through this Forum to provide you with answers about this session.

Individual Work. Duration 60 minutes. Mandatory

Each student must create the Database Cinemas under MS Acces



Now you can download and review the PDF file of this online session in Moodle. If you have any questions, please use the Forum platform in Moodle to follow exchanges and ask your questions. Your professor will exchange with you through this Forum to provide you with answers about this session.

Bibliography

- Book: ‘Data Analytics Made Accessible’. 2018. by Anil Maheshwari
- Book: ‘Too Big to Ignore: The Business Case for Big Data’. by award-winning
- Book: ‘Data Smart: Using Data Science to Transform Information into Insight’, by J. W. Foreman’.
- Paper: ‘Almeida, F. (2018). Big Data: Concept, Potentialities and Vulnerabilities’. Emerging Science Journal, 2(1).
- McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.
- Zikopoulos, P., & Eaton, C. (2011). Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.
- Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. International Journal of Information Management, 34(3), 387-394.

10:00

CONFÉRENCE DES GRANDES ÉCOLES
EXCELLENCE FOR A COMPLEX WORLD

EM Strasbourg
BUSINESS SCHOOL
UNIVERSITY OF STRASBOURG

End of this online session…

AACSB ACCREDITED

EFMD EPAS ACCREDITED