|  |  |
| --- | --- |
| . | DS1 Introduction to Data Science  Study Guide Bachelor Applied Artificial Intelligence & Data Management  Academic Year 2021-2022 |

Index

[1 General information 2](#_Toc40102014)

[2 Content of the module 2](#_Toc40102015)

[**2.1** **Module description** 2](#_Toc40102016)

[**2.2** **Relation B AAI&DM Competency Profile** 2](#_Toc40102017)

[**2.3** **Learning objectives** 2](#_Toc40102018)

[3 Teaching and Learning Activities 3](#_Toc40102019)

[**3.1** **Teaching and learning activities** 3](#_Toc40102020)

[**3.2** **Assignments** 3](#_Toc40102021)

[**3.3** **Planning** 3](#_Toc40102022)

[4 Study load 4](#_Toc40102023)

[5 Literature 4](#_Toc40102024)

[6 Assessment 4](#_Toc40102025)

[**6.1** **Assessment methods, type of evaluation and weight** 4](#_Toc40102026)

[**6.2** **Grading: Minimum grade, Compensation, and Duration of Validity** 4](#_Toc40102027)

[**6.3** **Assessment criteria** 5](#_Toc40102028)

[**6.4** **Requirements for handing in assignments** 5](#_Toc40102029)

[**6.5** **Grading** 5](#_Toc40102030)

[**6.6** **Resit** 6](#_Toc40102031)

1. General information

Module name: Introduction to Data Science

Module code: DS1

Semester: Block A

Number of ECTS credits: 3 ECTS credits (or 84 hours)

Year: 1

Lecturer(s): Bram Heijligers; Nitin Bhushan

Contact details: [heijligers.b@buas.nl](mailto:heijligers.b@buas.nl); [bhushan.n@buas.nl](mailto:bhushan.n@buas.nl)

1. Content of the module
   1. **Module description**

This module introduces students to the fundamental concepts and techniques for extracting useful knowledge from data. These concepts are of three types. First, it will discuss data-analytic thinking, and introduce data science standards that are commonly used in the industry. Second, the module introduces students to concepts in statistics and probability theory that form a basis for modern data science. Lastly, the module will zoom in on reporting data analysis and visualization.

* 1. **Relation B AAI&DM Competency Profile**
* **Question articulation:** In particular, the student can translate business requirements into data science problems that can be answered by data description, data visualization, correlation and/or regressions. Further, the student can identify different forms of tabular data and perform appropriate analyses on them.
* **Data collection & management:** Specifically, students can identify what data needs to be collected and in what format, to create value for individuals, organizations and domains.
* **Visualization & Reporting**: Students will understand different data visualization methods available and choose the appropriate graphical representation to gain insight from data.
* **Researching and reflecting attitude:** Students will develop an investigative, sceptical attitude. In specific, students will learn to critically interpret the results of their data science projects.
* **Responsibility:** Students will understand the concept of bias. They will learn issues arising from selection bias and how to resolve such issues.

The BoKS elements covered are:  2K1, 2K4, 2K5, 2S1, 2S2, 2S5, 2S7.

dule addresses the competency ‘………………’

* 1. **Learning objectives**
* Students demonstrate an understanding of different data types. This module introduces students to tabular data and the different data types normally encountered in tabular data. Students learn to distinguish between continuous and categorical data types. Further, they learn the difference between ordinal and nominal data types.
* Students can effectively describe data using statistics and graphs. Students learn to summarize and describe continuous and categorical data types using summary statistics including measures of location (mean, median, mode) and measures of dispersion (standard deviation, variance, range, IQR), and learn how to compute these measures by hand. Further, students learn the appropriate data visualization techniques applicable for each data type.
* Students demonstrate an understanding of statistical inference and basic probability theory. Students learn that data science often involves samples of data. Next, they understand that samples originate from populations which often have a probability model associated with them. They are introduced to basic concepts in statistical inference (to infer population parameters using sample statistics) using the standard normal distribution. Last, students learn the concept of conditional independence and Bayes rule from probability theory.
* Students demonstrate an understanding of correlation and simple linear regression. Further, they learn how to compute a correlation coefficient by hand using sigma notation.
* Students are familiar with CRISP-DM. They understand the different roles and responsibilities involved.
* Students are familiar with sigma-notation and linear transformations. They demonstrate their understanding by computing descriptive statistics and correlation coefficients by hand.
* Students can translate a business requirement into a data science problem. In sum, the student can identify different forms of tabular data and perform appropriate analyses (data description, data visualization, correlation and/or regression).

1. Teaching and Learning Activities
   1. **Teaching and learning activities**

Instruction methods used in this module are lectures, e-learning, workshops and case studies. Students are expected to self-study online study materials (mainly chapters from the online textbook and educational videos) before coming to class. In the lectures the main concepts discussed in the textbook will be reviewed, and illustrated using examples and short cases. In order to deepen understanding, and train students in the application of the concepts. In the same lecture, a case study will be introduced, to be analyzed in groups of maximum four students. One of the groups will present their analysis of the case in the next session, and open the floor for a plenary discussion.

* 1. **Assignments**

Each week students have to prepare classes by studying designated chapters of the textbook and other study materials. After each lecture, students will work in groups on selected case studies that illustrate the concepts discussed in that week.

* 1. **Planning**

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Day** | **Preparation** | **Lecture-Workshop Hybrid** |
| 3 | 1 | None | Introduction to DS (seeing the world in data, attributes) |
| 3 | 2 | Completed workshop 1 | Variables (data frames, continuous, nominal, ordinal etc.) |
| 3 | 3 | Completed workshop 2 | Descriptive analyses (mean, sd, range, IQR) & visualisation (boxplots) |
| 4 | 4 | Completed workshop 3 | Introduction to Probability (random variable, distributions) |
| 4 | 5 | Completed workshop 4 | Introduction to stat. inference (sample, pop, hypothesis testing) |
| 4 | 6 | Completed workshop 5 | Analysing Relationships between variables (e.g., compute correlation by hand) |
| 5 | 7 | Completed workshop 6 | Reporting & visualising (Academic Skills) |
| 5 | 8 | Completed workshop 7 | Introduction to regression & ML (supervised, unsup..., the CRISP DM model) |
| 5 | 9 | Completed all workshops | Conference Poster Presentation |

1. Study load

The study load for the module is divided as follows:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Self-study | 0 |  |
| Workshops (including preparation) | 45 (8x6) |  |
| Workshop assignments (case study) | 16 (8x2) |  |
| Final Poster/Presentation Preparation | 8 |  |
| TOTAL | 72 |  |

1. Literature

Provost, F., & Fawcett, T. (2013). *Data science for business: what you need to know about data mining and data-analytic thinking.* Sebastopol, California: O'Reilly.

**OnlineStatBook:** Online Statistics Education: A Free Resource for Introductory Statistics (onlinestatbook.com)

1. Assessment
   1. **Assessment methods, type of evaluation and weight**

The assessment of the module will take place through:

|  |  |  |
| --- | --- | --- |
| **Type of assessment** | **Type of evaluation** | **Percentage of the final grade** |
| **Individual assessment** | Conference Poster | 100 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

* 1. **Grading: Minimum grade, Compensation, and Duration of Validity**

**Minimum grade**

In order to pass the module, the minimum final grade is 5.5.

*Minimum grades for parts and compensation*

The lowest grade of the constituting elements is 5.

* 1. **Assessment criteria**

**Individual Assignments (100% of total grade)**

**In every interactive workshop the students create building blocks for a data science project proposal in the form of a conference poster. Emphasis is laid on creating a data-driven research question and subsequently describing and visualizing a dataset in a conference poster format.**

* Students can formulate a data-driven research question.
* Students demonstrate an understanding of different data types.
* Students can effectively describe data using statistics and graphs.
* Students demonstrate an understanding of statistical inference and basic probability theory.
* Students demonstrate an understanding of correlation and simple linear regression.
* Students are familiar with CRISP-DM.
* Students are familiar with sigma-notation and linear transformations.
* Students can translate a business requirement into a data science problem.

|  |  |  |
| --- | --- | --- |
| **Component** | **Description of component** | **ILO** |
| Introduction | What is the problem statement? | 1 |
| Dataset | Describe the data. | 6 |
| Data visualisation | Include relevant visuals. | 6 |
| Results | Do we have a data-driven solution to the problem statement? | 1 |
| Discussion | Critically look back at your solution . | 1 |
| References | Cite any sources used. | 6 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | |
| ILO | Missing | Poor | Insufficient | Sufficient | Good | Excellent |
| **1. Question articulation** | The problem statement is not translated to a data science problem. | The problem statement is rephrased without using data science concepts. | The problem statement is rephrased using the wrong data science concept (e.g., correlation instead of regression). | The problem statement is rephrased using the correct data science concept, with a wrong implementation/ interpretation. | The problem statement is rephrased using the correct data science concept, with a correct implementation/ interpretation. | The problem statement is rephrased using the correct data science concept, with a correct implementation, and the problem statement is further improved.  (Above and beyond). |
| **6. Visualization & Reporting** | Visualisation and reporting of findings is missing. | Unsuitable visualisation is used, and findings interpreted wrongly. | Suitable visualisations are included, but not interpreted. | Suitable visualisations are included but interpreted wrongly. | Suitable visualisations are included and interpreted correctly. | Suitable visualisations are included, interpreted correctly, and novel visualisations are proposed. (Above and beyond). |

* 1. **Requirements for handing in assignments**

The individual assignments are to be handed in no later then 5pm on the day before the session in which we present the posters.

* 1. **Grading**

Grading will take place in accordance with the TER. Grades will only be processed if you have registered in Osiris for the examination of the module.

*Missed opportunities*

If you have registered to take the exam but did not deliver (a part of) it, your result will be registered as a missed opportunity (abbreviated as GK).

*Invalid sufficient average grades*

In the case in which the final average grade is sufficient, but one of the parts (for example essay or exam) is insufficient, the letter O will be registered in Osiris to indicate that the results are insufficient.

*Rounding and averaging*

Grades will be entered in Osiris up to one decimal.

If marks consist of more than one decimal, the original mark will be broken off after the first decimal. A few examples:

• 5.49 becomes 5.4

• 6.73 becomes 6.7

*For further regulations on grading, see TER of B AAI&DM.*

* 1. **Resit**

Each student receives two exam opportunities per year. The first opportunity is during or at the end of the module (for assignments). The second opportunity follows later. As a rule, the resits for assignments (both group and individual) also take place in the exam period in which the resit of the exam will take place.

You need to register in Osiris to be able to take part in resits (whether exams, assignments, or reports).