Course Outline 2020

INFOSYS 722: DATA MINING AND BIG DATA (15 POINTS)

Semester 1 (1203)

Course prescription
Data mining and big data involves storing, processing, analyzing and making sense of huge volumes of data extracted in many formats and from many sources. Using information systems frameworks and knowledge discovery concepts, this project-based and research-oriented course uses latest published research and cutting-edge business intelligence tools for data analytics.

Course advice
None

Goals of the course
The goals of the course are to introduce students to:

1. Decision Making, Big Data, Machine Learning, and Data Mining – foundational concepts.
2. Big Data and Data Mining Computing Environment – hardware, software, distributed systems and analytical tools.
3. Turning data into insights that deliver value - through methodologies, processes, algorithms and approaches for big data analytics.
4. Big Data and Data Mining in Practice – how the world’s most successful companies use big data analytics to deliver extraordinary results.
5. Apply the knowledge gained through the design and implementation of a prototype.

Learning outcomes (LO)

<table>
<thead>
<tr>
<th>#</th>
<th>Learning outcome</th>
<th>Graduate profile capability*</th>
</tr>
</thead>
</table>
| LO1| Understand foundational concepts of Decision Making and Decision Support from a variety of disciplines; also, understand fundamental principles of Data Mining, Machine Learning, and Big Data. | 1. Disciplinary knowledge and practice  
4b. Communication (Written)  
4c. Engagement |
| LO2| Compare, contrast and synthesize a process for Data Mining; also understand the key components of the computing environment for Big Data and Data Mining including hardware, software, distributed systems, and analytical tools. | 1. Disciplinary knowledge and practice  
2. Critical thinking  
3. Solution seeking  
4b. Communication (Written) |
<table>
<thead>
<tr>
<th>#</th>
<th>Learning outcome</th>
<th>Graduate profile capability*</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO 3</td>
<td>Understand the process of <strong>turning data into insights that deliver value</strong> using predictive modelling, segmentation, incremental response modeling, time series data mining, text analytics, and recommendations.</td>
<td>1. Disciplinary knowledge and practice&lt;br&gt;2. Critical thinking&lt;br&gt;3. Solution seeking</td>
</tr>
<tr>
<td>LO 4</td>
<td>Understand, discuss, and reflect on how successful companies have applied big data and data mining <strong>methodologies, algorithms, and enabling technologies</strong> to deliver extraordinary results and value.</td>
<td>2. Critical thinking&lt;br&gt;3. Solution seeking&lt;br&gt;4c. Engagement</td>
</tr>
<tr>
<td>LO 5</td>
<td>Design and implement a prototypical <strong>Big Data Analytics Solution</strong> to address one of the <strong>17 Sustainable Development Goals</strong> (<a href="https://sdgsinaction.com/">https://sdgsinaction.com/</a>) of the UN or a <strong>decision making situation</strong> facing an organization of your choice.</td>
<td>3. Solution seeking&lt;br&gt;6a. Social responsibilities&lt;br&gt;6b. Environmental responsibilities</td>
</tr>
<tr>
<td>LO 6</td>
<td>Write a <strong>research paper</strong> that details (a) the practical problem (b) the research problem (c) the research objectives (d) the literature that explores potential solutions and methodologies that addresses your objectives (e) the research methodology adopted (f) the design of the processes that converts data into insights and (g) the description of the implementation using various algorithms and enabling technologies (h) your interpretation of the patterns and results and (i) your proposed actions based on the discovered knowledge. The research paper needs to be 10-30 pages (single space, 11 font size, with normal margins, APA format). <a href="https://www.library.auckland.ac.nz/subjectguides/bus/topicguides/apa_for_business.htm">https://www.library.auckland.ac.nz/subjectguides/bus/topicguides/apa_for_business.htm</a> Refer to journals such as Decision Support Systems <a href="https://www.journals.elsevier.com/decisionsupport-systems">https://www.journals.elsevier.com/decisionsupport-systems</a> Also the Journal of Big Data for exemplars. <a href="https://journalofbigdata.springeropen.com/">https://journalofbigdata.springeropen.com/</a></td>
<td>2. Critical thinking&lt;br&gt;4b. Communication (Written)&lt;br&gt;5a. Independence</td>
</tr>
</tbody>
</table>

* See the graduate profile this course belongs to at the end of this course outline.
## Content outline

<table>
<thead>
<tr>
<th>Week / Module</th>
<th>Lectures</th>
<th>Labs</th>
<th>Assessment due this period</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 02/03</td>
<td><strong>Lecture:</strong> Decision Making and Support. Intelligence Density. Big Data and Data Mining.</td>
<td>Data Mining Basics: Nine Step Process using ISAS: Introduction and Application</td>
<td></td>
</tr>
<tr>
<td>2 09/03</td>
<td><strong>Lecture:</strong> Data Mining Processes (KDD, SEMMA, and CRISP-DM). Exploring the Nine Step Process.</td>
<td>ISAS - Advanced Use Case</td>
<td>Iteration 1 – Proposal - Steps 1 – 2 (Due 5pm, Wednesday 11/03)</td>
</tr>
<tr>
<td>3 16/03</td>
<td><strong>Lecture:</strong> Passive Data Mining (Browsing, Visualization, Statistics, and Hypothesis testing)</td>
<td>ISAS - Advanced Use Case</td>
<td></td>
</tr>
<tr>
<td>4 23/03</td>
<td><strong>Lecture:</strong> Active Data Mining (Neural Networks, Rule Induction, Regression)</td>
<td>OSAS - Introduction to Python for Machine Learning</td>
<td>Iteration 2 – ISAS - Steps 1 – 8 (Due 5pm, Friday 27/03)</td>
</tr>
<tr>
<td>5 30/03</td>
<td><strong>Lecture:</strong> Modelling</td>
<td>OSAS - Python for Machine Learning</td>
<td></td>
</tr>
<tr>
<td>6 06/04</td>
<td><strong>Lecture:</strong> Interpretation</td>
<td>OSAS - Python Machine Learning Workflow</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Mid Semester Break</strong></td>
</tr>
<tr>
<td>7 27/04</td>
<td><strong>Lecture:</strong> Visualization, Digital Twins, and Data Mining in the Real World</td>
<td>OSAS - Recap on Tools and Technology and Assignment Help</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Guest Lecture:</strong> Mr Dulantha Peiris (BEng)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 04/05</td>
<td><strong>Lecture:</strong> The AI Startup. From Vision to Action.</td>
<td>BDAS - Introduction to GitHub, Amazon Web Services, Jupyter, Secure Shell (SSH), PySpark and Spark</td>
<td>Iteration 3 – OSAS - Steps 1 – 8 (Due 5pm Wednesday 06/05)</td>
</tr>
<tr>
<td>9 11/05</td>
<td><strong>Lecture:</strong> AI and DS</td>
<td>BDAS - PySpark, Spark, DataFrames and Data Cleaning Application</td>
<td>Resubmission of Iteration 2 (Due 5pm, Wednesday 13/05)</td>
</tr>
<tr>
<td>10 18/05</td>
<td><strong>Lecture:</strong> Data Ethics</td>
<td>BDAS - PySpark Algorithms and Visualisations</td>
<td></td>
</tr>
<tr>
<td>11 25/05</td>
<td><strong>Conclusion</strong></td>
<td>BDAS – Recap on Tools and Technology and Assignment Help</td>
<td></td>
</tr>
</tbody>
</table>

1 Refer to the nine steps of the assignment specification at the end of this document
Learning and teaching

The class will meet for three hours each week. Class time will be used for a combination of lectures and discussions. In addition to attending classes, students should be prepared to spend at least about another ten hours per week on activities related to this course. These activities include carrying out the required readings, labs and research relevant to this course, and preparing for assignments.

150 hours learning over a single semester including:
- 36 contact hours through lectures
- 22 contact hours through laboratories/tutorials
- 16 contact hours through technical workshop
- 76 hours of reading, programming, and self-study

Teaching staff

Academic Director and Lecturer
Dr Ami Peiris  Office: OGGB Room 476  Office Hour: Thursdays  4-5 PM
Email: a.peiris@auckland.ac.nz

Coordinators and Tutors

Shahzad Khan (Course Coordinator)  Office Hour: TBA
Email: Shahzad.khan@auckland.ac.nz

Shahab Bayati (Lead Tutor)  Office Hour: TBA
Email: s.bayati@auckland.ac.nz

Ankit Gandhi (Tutor)  Office Hour: Tuesdays  2-3 PM
Email: shahzad.khan@auckland.ac.nz

Di Zhao (Tutor)  Office Hour: TBA
Email: zd277607529@gmail.com

Shohil Kishore (Technical Coordinator) Email:
s.kishore@auckland.ac.nz

Course Material

There are two primary textbooks useful for the course. These textbooks can be downloaded free of cost from the University of Auckland library.


Other readings and supplemental material will be distributed in class as needed. Students are also advised to take advantage of the extensive software resources made available for this course.

**Assessment information**

| ISAS - IBM Software Analytics Solution – SPSS Modeller. |
| OSAS - Open Source Analytics Solution – Python, Jupyter, MySQL, MySQL Workbench, Kettle/Spoon, Tableau, & Weka. |
| BDAS - Big Data Analytics Solutions – GitHub, AWS (EC2/AMI), Jupyter, PySpark, Spark. |
| MSAS - Microsoft Software Analytics Solution – Microsoft SQL Server, Azure Machine Learning and Power BI. **MSAS is offered to the student for the student’s own self learning and does not constitute any portion of the final grade.** |

<table>
<thead>
<tr>
<th>Assessment task</th>
<th>Weight %</th>
<th>Group and/or individual</th>
<th>Submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Iteration 1 Proposal (Steps 1 – 2)</td>
<td>0</td>
<td>Individual</td>
<td>Week 2 – 5pm Wed, 11th March</td>
</tr>
<tr>
<td>2. Iteration 2 ISAS (Steps 1 – 8)</td>
<td>22</td>
<td>Individual</td>
<td>Week 4 – 5pm Fri, 27th March</td>
</tr>
<tr>
<td>3. Iteration 3 OSAS (Steps 1 – 8)</td>
<td>22</td>
<td>Individual</td>
<td>Week 8 – 5pm Wed, 06th May</td>
</tr>
<tr>
<td>4. Resubmission of Iteration 2</td>
<td></td>
<td>Individual</td>
<td>Week 9 – 5pm Wed, 13th May</td>
</tr>
<tr>
<td>5. Iteration 4 BDAS (Steps 1 – 8)</td>
<td>28</td>
<td>Individual</td>
<td>Week 12 – 5pm Wed, 03rd June</td>
</tr>
<tr>
<td>6. Research Paper (Details of Steps 1 – 9)</td>
<td>20</td>
<td>Individual</td>
<td>Week 13 – 5pm Wed, 10th June</td>
</tr>
<tr>
<td>7. Discussion Questions</td>
<td>8</td>
<td>Individual</td>
<td>Weeks 2, 4, 6, 8 8 pm Mondays</td>
</tr>
</tbody>
</table>

**Pass requirements**

Plussage applies for Iteration 2 ONLY. That is if you **re-submit Iteration 2 in Week 9** then we will remark it and if you score a better mark we will take the better mark as your mark for iteration 2.

An overall course grade of 50% or more is required to pass the course.
Description of assessment tasks

<table>
<thead>
<tr>
<th>Assessment task</th>
<th>Learning outcome to be assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Iteration 1 Proposal (Steps 1 – 2)</td>
<td>LO1, LO5</td>
</tr>
<tr>
<td>2. Iteration 2 ISAS (Steps 1 – 8)</td>
<td>LO1 – LO5</td>
</tr>
<tr>
<td>3. Iteration 3 OSAS (Steps 1 – 8)</td>
<td>LO1 – LO5</td>
</tr>
<tr>
<td>4. Resubmission of Iteration 2</td>
<td>LO1 – LO5</td>
</tr>
<tr>
<td>5. Iteration 4 BDAS (Steps 1 – 8)</td>
<td>LO1 – LO5</td>
</tr>
<tr>
<td>6. Research Paper (Details of Steps 1 – 9)</td>
<td>LO6</td>
</tr>
<tr>
<td>7. Discussion Questions</td>
<td>LO1 – LO5</td>
</tr>
</tbody>
</table>

Inclusive learning
Students are urged to discuss privately any impairment-related requirements face-to-face and/or in written form with the courses convenor/lecturer and/or tutor.

Academic integrity
The University of Auckland will not tolerate cheating, or assisting others to cheat, and views cheating in coursework as a serious academic offence. The work that a student submits for grading must be the student’s own work, reflecting his or her learning. Where work from other sources is used, it must be properly acknowledged and referenced. This requirement also applies to sources on the worldwide web. A student’s assessed work may be reviewed against electronic source material using computerised detection to provide an electronic version of their work for computerised review.

Student feedback
Student feedback is important to us and has been used to improve the course from semester to semester. This semester you may be asked to complete evaluations on the teaching of the course, both in lectures and in tutorials. Please note that you do not have to wait until these evaluations are conducted in order to provide feedback. If there is something that you think we could improve then please let us know (via email or in person) as soon as possible.

In the event of an unexpected disruption
We undertake to maintain the continuity and standard of teaching and learning in all your courses throughout the year. If there are unexpected disruptions, the University has contingency plans to ensure that access to your course continues and your assessment is fair, and not compromised. Some adjustments may need to be made in emergencies. In the event of a disruption, the University and your course coordinators will make every effort to provide you with up to date information via Canvas and the University website.

Graduate profile for Master of Commerce
The following six themes represent the capabilities that the Business School seeks to foster in all of its graduates. The development of these capabilities does not come all at once, but rather is expected to build from year to year. Each course is not expected to contribute
to all capabilities, but each course will have its own goals and learning outcomes that relate to the overall development of this profile.

<table>
<thead>
<tr>
<th>Graduate Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Disciplinary knowledge and practice</strong></td>
</tr>
<tr>
<td>Graduates will be able to apply highly specialized knowledge within the discipline to demonstrate an advanced awareness and understanding in a global context.</td>
</tr>
<tr>
<td><strong>2. Critical thinking</strong></td>
</tr>
<tr>
<td>Graduates will be able to analyze and evaluate the relevant literature, and design and develop scholarly arguments that demonstrate advanced and diverse thinking.</td>
</tr>
<tr>
<td><strong>3. Solution seeking</strong></td>
</tr>
<tr>
<td>Graduates will be able to creatively research and analyze complex issues, and develop innovative solutions.</td>
</tr>
<tr>
<td><strong>4. Communication and engagement</strong></td>
</tr>
<tr>
<td>Graduates will be able to engage, communicate, and collaborate with diverse groups using multiple formats and effectively address a range of professional and academic audiences.</td>
</tr>
<tr>
<td><strong>5. Independence and integrity</strong></td>
</tr>
<tr>
<td>Graduates will be able to demonstrate advanced independent thought, self-reflection, ethics, and integrity.</td>
</tr>
<tr>
<td><strong>6. Social and environmental responsibility</strong></td>
</tr>
<tr>
<td>Graduates will consider, in relation to their discipline, the potential significance of the principles underpinning both the Treaty of Waitangi and sustainability.</td>
</tr>
</tbody>
</table>
Design and implement a prototypical **Data Mining** and **Big Data Analytics Solution** to address one of the 17 Sustainable Development Goals of the UN or a decision making situation facing an organization of your choice.

The assignment follows a sequence of steps that is a synthesis of the Cross-Industry Standard Process for Data Mining (CRISP-DM) process (SPSS, 2007) and the KDD process (Fayyad et al., 1996).

1. **Business and/or Situation understanding (10%).** “First is developing an understanding of the application domain and the relevant prior knowledge and Figure 1: CRISP DM Process (SPSS, 2007) identifying the goal of the KDD process from the customer’s viewpoint.” (Fayyad et al., 1996)
   
   1.1 Identify the objectives of the business/situation  
   1.2 Assess the situation  
   1.3 Determine data mining objectives, and  
   1.4 Produce a project plan

2. **Data understanding (10%).** Data provides the “raw materials” of data mining. This phase addresses the need to understand what your data resources are and the characteristics of those resources. “Second is creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.” (Fayyad et al., 1996)
   
   2.1 Collect initial data  
   2.2 Describe the data  
   2.3 Explore the data, and  
   2.4 Verify the data quality
3. **Data preparation (15%).** After cataloguing your data resources, you will need to prepare your data for mining. “Third is data cleaning and pre-processing. Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes” (Fayyad et al., 1996)

- 3.1 Select the data
- 3.2 Clean the data
- 3.3 Construct the data
- 3.4 Integrate various data sources
- 3.5 Format the data as required

4. **Data transformation (5%).** “Fourth is data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or invariant representations for the data can be found.” (Fayyad et al., 1996)

- 4.1 Reduce the data
- 4.2 Project the data

5. **Data-mining method(s) selection (10%).** “Fifth is matching the goals of the KDD process (step 1) to a particular data-mining method. For example, summarization, classification, regression, clustering, and so on, are described later as well as in Fayyad, Piatetsky-Shapiro, and Smyth (1996).” (Fayyad et al., 1996)

- 5.1 Match and discuss the objectives of data mining (1.1) to data mining methods
- 5.2 Select the appropriate data-mining method(s) based on discussion

6. **Data-mining algorithm(s) selection (15%).** “Sixth is exploratory analysis and model and hypothesis selection: choosing the datamining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process includes deciding which models and parameters might be appropriate (for example, models of categorical data are different than models of vectors over the reals) and matching a particular data-mining method with the overall criteria of the KDD process (for example, the end user might be more interested in understanding the model than its predictive capabilities).” (Fayyad et al., 1996)

- 6.1 Conduct exploratory analysis and discuss
- 6.2 Select data-mining algorithms based on discussion
- 6.3 Build/Select appropriate model(s) and choose relevant parameter(s)

7. **Data Mining (15%).** “Seventh is data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps.” (Fayyad et al., 1996) This is, of course, the flashy part of data mining, where sophisticated analysis methods are used to extract information from the data.

- 7.1 Create and justify test designs
- 7.2 Conduct data mining – classify, regress, cluster, etc. (models must execute)
- 7.3 Search for patterns

8. **Interpretation (20%).** “Eighth is interpreting mined patterns, possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models.” (Fayyad et al., 1996) We assess and evaluate the models and the results and
their reliability. “You are ready to evaluate how the data mining results can help you to achieve your objectives.” (SPSS, 2007)

8.1 Study and discuss the mined patterns
8.2 Visualize the data, results, models, and patterns
8.3 Interpret the results, models, and patterns
8.4 Assess and evaluate results, models, and patterns
8.5 Iterate prior steps (1 – 7) as required

9. Action: “Ninth is acting on the discovered knowledge: using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting it to interested parties. This process also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge.” (Fayyad et al., 1996) “Now that you've invested all of this effort, it’s time to reap the benefits. This phase focuses on integrating your new knowledge into your everyday business processes to solve your original business problem and/or situation.” (SPSS, 2007)

9.1 Discuss how you would apply the knowledge and deploy the implementation
9.3 Discuss how you would monitor the implementation
9.4 Discuss how you would maintain the implementation
9.5 How could you enhance the solution in the future?

Research Paper Specification

Write a research paper that details (a) the practical problem (b) the research problem (c) the research objectives (d) the literature that explores potential solutions and methodologies that addresses your objectives (e) the research methodology adopted (f) the design of the processes that converts data into insights and (g) the description of the implementation using various algorithms and enabling technologies (h) your interpretation of the patterns and results and (i) your proposed actions based on the discovered knowledge. The research paper needs to be 10-30 pages (single space, 11 font size, with normal margins, APA format). Refer to journals such as Decision Support Systems (https://www.journals.elsevier.com/decision-support-systems) and Journal of Big Data (https://journalofbigdata.springeropen.com/) for exemplars.
## Lecture and Lab Readings, Videos and Materials

### Week 1
- **Data Mining Basics: Steps 1 - 9 using ISAS**

<table>
<thead>
<tr>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building a Data Mining Model</td>
<td>Predictive Analytics on SPSS Modeller</td>
<td>Constructing a Predictive Model</td>
</tr>
<tr>
<td>Building a Data Visualisation Model</td>
<td>Changes to Computer Thinking</td>
<td>Microsoft Course on Data Science Fundamentals (SL)</td>
</tr>
</tbody>
</table>

### Week 4
- **OSAS – Python for Machine Learning**
  - ‘Hello World’ for Machine Learning (SL)
  - Python for Non-Programmers (SL)
  - Python Semantics (SL) / Python Fundamentals (SL)
  - What is Anaconda?
  - What is Jupyter Notebooks? / What is Spyder?

### Week 6
- **OSAS – Machine Learning Workflow**
  - Kettle Fundamentals
  - Weka Fundamentals Course
  - Tableau Fundamentals
  - MySQL Fundamentals / MySQL Workbench
  - Neural Networks that Change Everything

### Week 8
- **BDAS – Introduction and Application**
  - AWS EC2 Documentation / EC2 Masterclass
  - Git Crash Course
  - What is Spark? / Spark 101
  - PySpark Introduction / PySpark with Big Data Course
  - Can Google Predict the Stock Market?

### MSAS – Azure Machine Learning and Power BI (Self-learning; Will not be assessed)
- Getting Started with Microsoft Azure / AML Basics
- Practical AML Experiment / Comparing Regressors on AML
- From CSV to Power BI Dashboard
- SQL Server Data Tools Tutorial / Microsoft SSAS Tutorials
- Using the Cloud to Forecast the Weather
- Machine Learning Overview
- Power BI (SL)
- Big Data Is Better Data
- The Beauty of Data Visualisation
- The Turing Test

SL = Self-Learning. Anything marked with ‘SL’ is a course where you may gain a certification. You can complete these courses at your discretion.