CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Department of Mathematics

MSc in Data Science and Analytics

Application for Granting of Exit Award
August 2020

Contents

1. DEPARTMENTAL OVERVIEW

1.1 INTRODUCTION & BACKGROUND

The Department of Mathematics is one of four departments within the School of Science and Informatics (SOSI), which forms part of the Faculty of Science and Engineering (FES) at Cork Institute of Technology (CIT). The Department of Mathematics was established in 2003 following the reorganisation of the then Department of Mathematics and Computing.

The department provides a teaching service across the Institute. Its lecturers deliver a portfolio of Mathematics and Statistics modules in the vast majority of the institute's academic programmes, in the Faculty of Engineering and Science (including the National Maritime College of Ireland) and the Faculty of Business and Humanities. This delivery ranges from modules in Mathematics for Craftspersons to Statistical Research Skills and Methods for postgraduate engineering programmes. Modules are delivered in subject areas including Calculus, Linear Algebra, Transform Methods, Probability, Statistics, Numerical Methods and Operations Research among others. The department is also very mindful of the career-focussed approach of the Institutes of Technology sector (and, indeed, of the emerging Technological University sector) and, to that end, constantly strives to maintain a practical and contextualised approach to delivery. Many undergraduate modules therefore include exposure to relevant software packages such as Maple, Minitab, Excel, VBA, SPSS and R.

The Department of Mathematics typically engages with approximately 5500 students during an academic year through its service teaching. As students in CIT are credited to their home departments, the number of full-time equivalent students of the Department of Mathematics would therefore be in the region of 500, when calculated on the basis of credits delivered and the size of class groups. For 2019/20 this figure also includes five PhD students (including a former graduate of the *HDip in Data Science and Analytics programme*), one Masters by Research student, 39 students on the *HDip in Data Science and Analytics* programme (part time) and 41 students currently enrolled in the full-time *MSc in Data Science and Analytics programme*.

1.2 Staff Overview

The department currently comprises of 29 full-time lecturing staff with 48.3% female, significantly above the faculty average of 27%. Approximately 65.5% (19) of the staff in the department currently hold PhD degrees. In addition to this, there are a further 7 staff members working on a part-time basis, teaching across a number of full-time and part-time module deliveries, of whom 3 hold PhD degrees.

	Name	Role
1	Ms Denise O'Keeffe	Administrator
2	Ms Maretta Brennan	Lecturer
3	Dr Michael Brennan	Lecturer
4	Ms Katie Bullen	Lecturer
5	Dr Clodagh Carroll	Lecturer
6	Ms Patricia Cogan	Lecturer
7	Dr Tadhg Creedon	Lecturer
8	Dr Julie Crowley	Lecturer
9	Dr Vincent Cregan	Lecturer
10	Mr Aengus Daly	Lecturer
11	Ms Jackie English	Lecturer
12	Dr David Goulding	Acting Head of Department
13	Dr David Hawe	Lecturer
14	D Mark Hartnett	Lecturer
15	Dr Robert Heffernan	Lecturer
16	Dr Seán Lacey	Lecturer
17	Dr Maryna Lishchynska	Lecturer
18	Ms Hannah Lordan	Lecturer
19	Dr Jeremiah McCarthy	Lecturer
20	Dr Justin McGuinness	Lecturer
21	Dr Violeta Morari	Lecturer
22	Ms Sarah Murphy	Lecturer
23	Dr Marie Nicholson	Lecturer
24	Mr Adrian O'Connor	Lecturer
25	Dr Declan O'Connor	Lecturer
26	Dr Shane O'Rourke	Lecturer
27	Mr Donal O'Shea	Lecturer
28	Dr Catherine Palmer	Lecturer
29	Dr Noreen Quinn	Lecturer
30	Ms Gráinne Read	Lecturer

2. CURRENT PROGRAMMES

Title	Higher Diploma in Science in Data Science and Analytics
Sponsor	Department of Mathematics
Head of Department	Dr David Goulding (Acting)
School	School of Science and Informatics
Head of School	Dr Brendan O'Connell
Faculty	Engineering & Science
Head of Faculty	Mr Tim Horgan
Entry Requirements	Level 8 (Honours) Degree
Duration (full-time)	1 (3) semester(s) with 30 (15) taught credits and 1 semester with 20 (5) taught credits and a 10 credit research project
Mode	Full-time (Part-time)

Title	MSc in Science in Data Science and Analytics
Sponsor	Department of Mathematics
Head of Department	Dr David Goulding (Acting)
School	School of Science and Informatics
Head of School	Dr Brendan O'Connell
Faculty	Engineering & Science
Head of Faculty	Mr Tim Horgan
Entry Requirements	2H1 Level 8 (Honours) Degree
Duration (full-time)	3 semesters: 2 semesters with 30 taught credits in each semester and 1 semester with a 30 credit research project
Mode	Full-time

2.1 INTRODUCTION TO DATA SCIENCE AND ANALYTICS PROGRAMMES IN CIT

Nationally, the Irish Government, in its various action plans for jobs (beginning with the *Action Plan for Jobs 2013*) has committed to exploit the business potential of Data Analytics and Big Data in Ireland[3]. Indeed, in *Ireland's National Skills Strategy 2025*, the Department of Education and Skills forecasted an extra 18,000 jobs in Data Analytics alone between 2013 and 2020[1]. National Skills Bulletins from SOLAS[2] continue to highlight data science as an emerging, difficult to fill and/or in high demand skillset. The increased growth in the field of Data Science and Analytics has seen the requirement for data savvy skillsets in many fields of application.

[1] Department of Education & Skills, *Ireland's National Skills Strategy 2025,* www.education.ie/en/Publications/Policy-Reports/pub_national_skills_strategy_2025.pdf

[2] National Skills Bulletin, 2016 http://www.skillsireland.ie/Publications/2016/National-Skills-Bulletin-2016-Web-Final.pdf

This demand is also made clear in the "Future Skills Needs of the Biopharma Industry in Ireland" report[3], in which it is highlighted that 'the application and use of data analytics is a key emerging skills area' within the industry. As the growth in the BioPharma industry continues this is therefore allied with increased demand in areas of data collection, storage and analysis. Therefore, the report highlights that skillsets in data science and analytics are 'in growing demand across several roles including for health product development, quality control/validation and process optimisation'. It is also noted that in SAS's "The Enterprise AI Promise: Path to Value" report of August 2017[4] over 25% of the respondents highlighted the lack of data scientists as the most significant skill challenge faced in the deployment of Artificial Intelligence within their companies. In the local region, in 2012, both Cork Chamber and Limerick Chamber identified Data Analytics as an area in which a programme should be developed by a HEI in the region[5].

In an effort to address the clear skills gap in the area of Data Science and Analytics, CIT's Department of Mathematics, in partnership with the Department of Computer Science, developed both the *HDip in Data Science and Analytics* (NFQ Level 8) in 2013 and the *MSc in Data Science and Analytics* (NFQ Level 9) in 2017. The department has been successful in applying for Springboard funding under the ICT Skills initiative from 2013-2018 to run the full-time one year HDip conversion programme and for HEA funding to run the MSc in 2017. In recent years with the changing employment landscape in Ireland, provision for the HDip was moved solely to a part-time delivery while the MSc remains a full-time offering in the department.

The *MSc in Data Science and Analytics* is a 90 credit programme, in which core strands – Statistics, Mathematics, Computer Science, Data Science and Data Analytics– are developed and interleaved over the course of the programme. Significant opportunities exist throughout the course for learners to apply their theoretical knowledge and to develop problem solving skills through practical and laboratory sessions. The learners undertake 60 taught credits followed by a 30-credit research project in their final semester, which is an opportunity to

[3] Expert Group on Future Skills Needs, "Future Skills Needs of the Biopharma Industry in Ireland", 2016 http://www.skillsireland.ie/Publications/2016/Biopharma-Skills-Report-FINAL-WEB-VERSION.pdf
[4] SAS "The Enterprise AI Promise: Path to Value", 2017 http://retailbankinginnovation.fintecnet.com/uploads/2/4/3/8/24384857/sas_ai_readiness_survey.pdf
[5] Cork Chamber & Limerick Chamber, *Regional Employer Skills & Education Needs Survey, March 2012*

demonstrate their ability to synthesise the learning acquired in the programme and to apply it in the solution of an authentic research problem in the field of Data Science and Analytics.

The *HDip in Data Science and Analytics*, on the other hand, is a 60 credit programme in which a similar approach to developing expertise within graduates in the core themes of Data Science an Analytics is also employed. During the HDip programme, the students undertake 50 taught credits and a capstone 10 credit project.

As the demand for the Level 9 MSc programme has increased in recent years, it has become clear that an exit award for the programme should be considered. The remainder of this document lays out the educational outcomes of both the HDip and MSc programme and aims to demonstrate the the HDip in Data Science and Analytics could be approved as a suitable exit award from the MSc in Data Science and Analytics in CIT.

2.2 Overview of Programme of Study

The *HDip in Data Science & Analytics* runs over two (four) semesters for full-time (part-time) students. For the full-time students each semester accrues 30 credits, while in the part-time programme students will typically accrue 15 credits per semester. All of 60 ECTS are at advanced level, in accordance with CIT's academic policy for taught Higher Diploma programmes.

The Semester 1 schedule (30 ECTS) consists of six 5-credit modules which provide the learner with the necessary foundation in Mathematics, Statistics and Computer Science. The Semester 2 schedule for the full-time HDip consists of four 5-credit taught modules and a capstone 10-credit project module.

The *MSc in Data Science and Analytics*, on the other hand, is a 90 credit programme. The MSc programme was designed specifically to cater for graduates of all disciplines and not simply for graduates of Computer Science or Mathematics. In order to facilitate this requirement, the first 30 credits of the MSc were chosen to be shared with the HDip programme. These 30 credits at advanced level are to provide students with a grounding in Statistics, Mathematics, Computer Science and Data Science. Following the first semester, the remaining credits in the

MSc are all at expert level. Table 1 below shows a comparison of the modules in the MSc and the HDip in Data Science and Analytics.

TABLE 1 MODULE LIST FOR DATA SCIENCE PROGRAMMES IN CIT

	HDip Data Science and Analytics	MSc Data Science and Analytics
Semester 1	STAT8006 – Applied Stats and Probability COMP8060 - Scientific Prog in Python STAT8010 - Intro to R for Data Science DATA800 2 - Data Management Systems MATH8009 - Maths Methods and Modelling DATA8001 - Data Science and Analytics	
Semester 2	STAT8011 - Regression Analysis DATA8005 - Distributed Data Management DATA8008 - Data Visualisation & Analytics STAT8008 - Time Series & PCA COMP8043 - Machine Learning DATA8006 - Data Science Analytics Project	STAT9004 - Statistical Data Analysis DATA9002 - Distributed Data Management DATA9001 - Data Analytics & Visualisation STAT9005 - Time Series & Factor Analysis COMP9060 - Applied Machine Learning MATH9001 - Research Methods
Semester 3		DATA9003 - Research Project – Data Science

Table 1 provides a breakdown of the modules taught in both programmes. The modules listed in red for the Higher Diploma are two elective modules and from these students select one module to complete. For the MSc programme, the modules listed in blue in Table 1 are the modules completed at expert level – please note that there are no electives in the MSc programme.

2.2 Educational Aims & Programme Outcomes

TABLE 2 PROGRAMME OUTCOMES FOR HDIP IN DATA SCIENCE AND ANALYTICS

HDip in DSA

Programme Outcomes		
PO1	**Knowledge – Breadth**	
	(a)	Demonstrate detailed knowledge and understanding of areas of Mathematics, Statistics, Computer Science and Business Intelligence relevant to a Data Scientist.
PO2	**Knowledge – Kind**	
	(a)	Demonstrate understanding of the terminology, defining concepts and theories underlying the Data Science and Analytics field; demonstrate knowledge of the advanced methods and technologies for acquiring, interpreting and analysing big data, with a critical understanding of the appropriate contexts for their use; relate current issues in Data Science to society; understand current knowledge of the Data Science field, including current limits of theoretical and applied knowledge.
PO3	**Skill – Range**	
	(a)	Demonstrate mastery of relevant skills and tools in Statistics, Mathematics, Computer Science and Business Intelligence; use these to solve complex problems involving big data sets; interpret and apply appropriate and referenced literature and other information sources; work independently within defined time and resource boundaries; communicate scientific information in a variety of forms to specialist and non-specialist audiences.
PO4	**Skill – Selectivity**	
	(a)	Formulate and test hypotheses; design experiments; appreciate current limits of knowledge in the Data Science field and respond appropriately; think independently and make effective decisions; contribute fully to the day-to-day operations of the Data Science work setting.
PO5	**Competence – Context**	
	(a)	Apply data analysis skills and technologies in a range of contexts in order to critically interpret existing knowledge and apply in new situations; make and report appropriate decisions in a responsible and ethical manner.
PO6	**Competence – Role**	
	(a)	Act effectively under guidance in a peer relationship with qualified practitioners; participate constructively in a complex interdisciplinary team environment; plan for effective project implementation; reflect on own practices.
PO7	**Competence - Learning to Learn**	
	(a)	Learn to act in variable and unfamiliar learning contexts; identify learning needs and undertake continuous learning in the Data Science field; assimilate and apply new learning.
PO8	**Competence – Insight**	
	(a)	Demonstrate an understanding of the wider social, political, business and economic contexts of Data Science, including an appreciation of the philosophical and ethical issues involved.

TABLE 3 PROGRAMME OUTCOMES FOR MSC IN DATA SCIENCE AND ANALYTICS

MSc in DSA

Programme Outcomes		
PO1	**Knowledge – Breadth**	
	(a)	Demonstrate detailed knowledge and understanding of areas of Mathematics, Statistics, Computer Science and Business Intelligence relevant to the Data Analyst.
	(b)	Identify and articulate the key considerations of a "Big Data" problem; draw complex information together; critically comment on the technical, social, economic, environmental and political implications of own work and the work of others in Data Science, including an appreciation of the philosophical and ethical issues involved.
PO2	**Knowledge – Kind**	
	(a)	Demonstrate comprehensive knowledge and detailed understanding of: the theories, paradigms, defining concepts and underlying principles of the rapidly evolving Data Science and Analytics field; demonstrate knowledge and experience of advanced and new methods and technologies for acquiring, interpreting and analysing big data, with a critical awareness of the appropriate contexts for their use through the study of original papers, reports, journals, and data sets; demonstrate comprehensive knowledge and understanding of: the identification, definition and resolution of novel, complex research problems; relevant legal and regulatory frameworks; aspects of the defining elements and the inter-relationships of Data Science & Analytics as a result of in-depth study and research; demonstrate comprehensive knowledge and appreciation of the current limits of theoretical and applied knowledge in interdisciplinary field of Data Science and Analytics.
PO3	**Skill – Range**	
	(a)	Demonstrate mastery of standard and specialised research tools in Statistics, Mathematics, Computer Science and Business Intelligence; use these to proactively model, troubleshoot and solve original technical problems in the "Big Data" space; source relevant information, critically interpret and apply appropriate referenced literature from a wide range of information sources; maintain detailed records of activities; present and defend scientific research findings in a variety of forms to data scientists, "data savvy" practitioners, and non-specialists; formulate a hypothesis and design a relevant programme of investigation; work independently within defined time and resource boundaries; write accurately and in a manner consistent with scientific publications in Data Science or related disciplines.
PO4	**Skill – Selectivity**	
	(a)	Design, develop and test novel hypotheses; design experiments; select from a range of scientific skills, in particular those which draw from Mathematics/Statistics and Computer Science, so as to apply the most appropriate in a range of situations; think independently and make informed effective decisions; make decisions in the Data Science work setting; develop new skills either independently or with minimal mentoring.
PO5	**Competence – Context**	
	(a)	Apply advanced research skills and Big Data technologies; act autonomously and think independently; constructively criticise, draw conclusions and offer recommendations in a wide range of contexts, including unpredictable situations; formulate and communicate judgements, with incomplete or limited information.

PO6	**Competence – Role**	
	(a)	Act effectively, demonstrate initiative, lead and take responsibility in a complex interdisciplinary team environment with data scientists and qualified practitioners of other disciplines; develop and implement novel technical solutions for "Big Data" problems; reflect on own practices.
PO7	**Competence - Learning to Learn**	
	(a)	Learn to act in variable and unfamiliar learning contexts; identify knowledge gaps through effective self-evaluation; source and undertake self-learning as necessary for continued academic and professional self-development as a Data Scientist.
PO8	**Competence – Insight**	
	(a)	Identify and articulate the key considerations of a "Big Data" problem; draw complex information together; critically comment on the technical, social, economic, environmental and political implications of own work and the work of others in Data Science, including an appreciation of the philosophical and ethical issues involved.

Tables 2 and 3 present the Programme Outcomes of both the *HDip in Data Science and Analytics* and the *MSc in Data Science and Analytics* programmes, respectively. These tables present a general overview of graduate learning in terms of knowledge, skills and competencies, with each programme designed to match the skills profile of current data scientists/analysts. From the tables it is clear that the programme outcomes of the *HDip in Data Science and Analytics* form a subset of the programme outcomes of the *MSc in Data Science and Analytics*. It should be noted that the expert level modules in the MSc programme provide the student with a far deeper understanding and appreciation of the topics that they are exposed to.

2.3 Comparison of Second Semester Modules

Given that the *HDip in Data Science and Analytics* and the *MSc in Data Science and Analytics* programmes share a common first semester, this document therefore will focus only on considering the equivalence of the second semester modules.

As can be seen in Table 1, the modules in the second semester for both programmes share common themes. The most significant difference between the two programmes in terms of content in the second semester is that for the HDip, students select one elective (either STAT8008 – Time Series and PCA or COMP8043 – Machine Learning) and complete a 10 credit project whereas the MSc students complete all taught modules including STAT9005 – Time Series and Factor Analysis, COMP9060 – Applied Machine Learning and MATH9001 –

Research Methods. Appendix A contains a comparison of the learning outcomes from each module on the respective programme.

TABLE 4 MODULES FROM SEMESTER 2 THAT CAN BE CONSIDERED EQUIVALENT WITHIN EACH PROGRAMME

HDip in Data Science and Analytics	***MSc in Data Science and Analytics***
STAT8011 - Regression Analysis	STAT9004 - Statistical Data Analysis
DATA8005 - Distributed Data Management	DATA9002 - Distributed Data Management
DATA8008 - Data Visualisation & Analytics	DATA9001 - Data Analytics & Visualisation

Table 4 presents three Semester modules which can be considered as being directly analogous between the two programmes. It can be seen, from the learning outcomes in Appendix A that a direct comparison between the learning achieved in these modules can be made, with again a deeper insight and understanding being attained by learners on the Expert level modules within the MSc programme. Furthermore, Appendix B provides a comparison of both the module content and the assessment schedules for each of the modules again demonstrating that the modules can be considered equivalent for the purpose of aligning the HDip in Data Science and Analytics as an exit award for the MSc in Data Science and Analytics.

The final 15 credits from the two programmes are presented in Table 5. Again, please note that the students in the HDip in Data Science and Analytics select one of the elective modules (STAT8008 or COMP8043) to take and then complete the 10 credit project module DATA8006. It can be seen from the Learning Outcomes in Appendix A and the module content in Appendix B, that the modules COMP8043 and COMP9060 could again be considered equivalent for the purpose of considering an exit award. Similarly while there is a significant difference in the modules STAT8008 and STAT9005 in terms of focus, it can be argued that the time series component of both modules would be considered equivalent and that the additional inclusion of factor analysis in the MSc module provides additional advanced learning at Expert level for these students.

TABLE 5 REMAINING MODULES IN SEMESTER 2

HDip in Data Science and Analytics	***MSc in Data Science and Analytics***
STAT8008 - Time Series & PCA	STAT9005 - Time Series & Factor Analysis
COMP8043 - Machine Learning	COMP9060 - Applied Machine Learning
DATA8006 - Data Science Analytics Project	MATH9001 - Research Methods

Finally it is important to consider the 10 credit project module (DATA8006) which is completed by students in the HDip programme and to seek to map the learning outcomes of this module with a combination of the 5-credit Research Methods module (MATH9001) and the additional 5-credit module at expert level which is not equivalent to the student's chosen elective. From examining the learning outcomes for the modules DATA8006 and MATH9001 it is clear that a direct comparison between the two can be elicited but it should be noted that the research undertaking by the MSc student in MATH9001 tends to be completed in a more structured format than that would be expected in the 10-credit research project, however the Department of Mathematics have ensured that throughout the programme the students are exposed to real world data sets and are assessed on unstructured research problems based on these data sets. It can therefore be claimed that the combination of the two five credit modules in the MSc, taken at expert level, can be considered to be equivalent to the 10-credit project module on the HDip programme.

3. CONCLUSION

The Department of Mathematics are seeking to ask Council to approve the HDip in Data Science and Analytics as an exit award for the MSc in Data Science and Analytics. This document provides the background to the two programmes in terms of their development and discusses how the two programmes are intertwined. Given that the first 30 credits of both programmes are shared, the Department of Mathematics therefore only focuses on demonstrating the equivalence of the Semester 2 modules on both programmes and seek to demonstrate that the HDip in Data Science and Analytics provides a suitable exit award for the higher degree programme. Through mapping programme and module learning outcomes, in addition to examining the correspondence of indicative content across modules, the Department believe that they have fully demonstrated that the programme learning outcomes of the Higher Diploma programme are indeed met by the completion of the 60 taught credits of the Masters programme. Based on this document and with the agreement of both an industrial and academic extern, the Department now wish to seek the approval of Council for the granting of this exit award.

APPENDIX A – COMPARISON OF LEARNING OUTCOMES FOR SEMESTER 2 MODULES

HDip in DSA	MSc in DSA
DATA8008	**DATA9005**
Data Visualisation & Analytics	**Data Analytics & Visualisation**
5.0	**5.0**
Advanced	**Expert**
1. Describe the concepts, principles and methods of data visualisation.	1. Describe and critique the concepts, principles and methods of data visualisation.
2. Apply data explorative and pre-processing techniques to specified datasets.	2. Research and apply a variety of data explorative and pre-processing techniques to a range of datasets.
3. Design, implement and communicate appropriate data visualisation techniques to solve data analytical problems.	3. Research and appraise a variety of data analytics solutions to current challenges in the area.
4. Interpret and communicate patterns and knowledge discovered as a result of applying data visualisation and analytical techniques to data sets and analytical problems.	4. Interpret, critique and communicate patterns and knowledge discovered as a result of applying data visualisation techniques and analytical techniques to datasets and analytical problems.
5. Assess a variety of data analytics solutions to current challenges in the area.	5. Research and implement appropriate data visualisation techniques to solve data analytical problems.

HDip in DSA	MSc in DSA
STAT8011	**STAT9004**
Regression Analysis	**Statistical Data Analysis**
5.0	**5.0**
Advanced	**Expert**
1. Explore data sets and select appropriate statistical methods for data science problems.	1. Explore data sets and establish a data analysis protocol for data science problems.
2. Interpret the results of statistical analyses performed by a software package or presented in research papers.	2. Interpret the results of statistical analyses performed by a software package or presented in research papers.
3. Analyse data sets with binary response variables using logistic regression.	3. Build and validate statistical models with categorical response variables using logistic regression.
4. Analyse data sets with continuous response variables and multiple predictors (both categorical and continuous) using ANOVA, multiple regression and ANCOVA.	4. Build and validate statistical models with continuous response variables and multiple predictors (both categorical and continuous) using ANOVA, multiple regression and ANCOVA.
5. Apply the concepts of Design of Experiments and analyse associated sets of data.	5. Distinguish between parametric and non-parametric methods and decide when the most commonly used non-parametric methods should be applied.
	6. Explain and apply the statistical concepts relevant to experimental design and data analysis with an emphasis on large data sets.

HDip in DSA	MSc in DSA
COMP8043 (Elective)	**COMP9060**
Machine Learning	**Applied Machine Learning**
5.0	**5.0**
Advanced	**Expert**
1. Select and apply appropriate machine learning algorithms to datasets from a specific application domain.	1. Apply appropriate machine learning methodologies to facilitate pre-processing, dimensionality reduction and model selection.
2. Analyse and evaluate the performance of machine learning algorithms.	2. Evaluate the accuracy of predictive models using standard methods.
3. Apply machine learning methodologies to facilitate pre-processing, dimensionality reduction and model selection.	3. Select and apply appropriate machine learning algorithms to datasets from a specific application domain.
4. Develop a machine learning algorithm for solving a real-world problem.	4. Develop and implement machine learning algorithms for building predictive models.
5. Implement and apply optimization algorithms for solving complex problems with a high dimensional search space.	5. Implement and apply optimization algorithms for solving complex problems with a high dimensional search space.
	6. Apply neural networks and deep learning methods for solving real-world problems.

HDip in DSA	MSc in DSA
STAT8008 (Elective)	**STAT9005**
Time Series & PCA	**Time Series & Factor Analysis**
5.0	**5.0**
Advanced	**Expert**
1. Perform PCA to reduce dimensionality of datasets.	1. Implement factor analysis techniques on a large dataset and interpret the resulting models.
2. Describe the assumptions underlying PCA & time series models.	2. Apply the theoretical principles that govern a time series.
3. Apply the theoretical principles that govern a time series.	3. Apply regression and time series model for prediction. Differentiate between pure and causal time series models.
4. Apply regression and time series models for prediction, and give an account of the paradigm under which the forecasts are being made, along with their reliability.	4. Use statistical packages to generate and analyse models.
5. Perform diagnostic analysis and forecasts for both PCA and time series models, using statistical software.	5. Critically analyse and report on the paradigm under which forecasts are being made, along with their reliability. Perform residuals analysis and tests of fit.

HDip in DSA	MSc in DSA
DATA8005	**DATA9002**
Distributed Data Management	**Distributed Data Management**
5.0	**5.0**
Advanced	**Expert**
1. Appraise the challenges posed by big data and the new infrastructure, data models and processing techniques it demands.	1. Appraise the challenges posed by big data and the new infrastructure, data models and processing techniques it demands.
2. Survey the main NoSQL-based data models, exploring the best-fit for different use-cases.	2. Compare and contrast the main NoSQL-based data models, discriminating the best fit for different use-cases.
3. Query a range of NoSQL databases using a high-level programming language.	3. Combine document-oriented and graph-based data models for a fit for purpose multi-component system.
4. Explore the scalability, flexibility and reliability of a distributed data cluster supporting large data sets.	4. Demonstrate the scalability, flexibility and reliability of a distributed data cluster supporting large data sets.
5. Implement an analytical solution over a large-scale dataset using MapReduce and Spark.	5. Compare and contrast the MapReduce and Spark large-scale analytics libraries in terms of their expressiveness and efficiency.

HDip in DSA	MSc in DSA
DATA8006	**MATH9001**
Data Science Analytics Project	**Research Methods**
10.0	**5.0**
Advanced	**Expert**
1. Conduct a feasibility study of the proposed data science methodologies and technologies.	1. Propose a research question, develop the research methodology and project a plan for the research project.
2. Systematically review and adapt the employed data science methodologies during implementation in response to practical, real-world constraints.	2. Undertake preliminary experimental/design/analytical/modelling work as appropriate.
3. Critically assess the project outcomes.	3. Evaluate critically a number of solutions to the identified problem.
4. Use appropriate written and oral communication skills required of a professional practitioner, with a particular emphasis on conveying the underlying message of the research to stakeholders at all stages of the data science/analytics project.	4. Communicate effectively the idea and contribution of the proposed research project.
5. Undertake a review of relevant and appropriate literature to determine current knowledge in a field of data science and analytics.	5. Select a potential peer-review conference/journal paper for the research work and identify how this work may contribute to furthering knowledge in the specific field.
6. Plan the creation of effective final deliverables for a data science/analytics project that will meet the needs of stakeholders and others.	6. Plan the creation of effective final deliverables for a data science/analytics project that will meet the needs of stakeholders and others.

APPENDIX B – COMPARISON OF INDICATIVE CONTENT FOR SEMESTER 2 MODULES

STAT8011 - Regression Analysis
Module Indicative Content
Data Analysis Protocol
Consolidate prior knowledge of graphical and numerical descriptive statistics to explore categorical and continuous data sets. Outliers, missing values, testing of assumptions and transformation of variables. Model fitting and model interpretation. Model diagnostics.
Design of Experiments
Observational (vs) experimental data. The fundamentals of experimental design. Analysis of variance. Factorial design.
Multiple Regression
Assumptions, collinearity, interpreting coefficients, model fitting, model diagnostics, confidence intervals of coefficients, Analysis of covariance (ANCOVA).
Logistic Regression
Overview of different types of generalised linear models and their uses with a focus on logistic regression for binary data.
Software analysis
SPSS, R, Excel

STAT9004 - Statistical Data Analysis
Module Indicative Content
Data Analysis Protocol
Exploratory data analysis: graphical and numerical methods to explore categorical and continuous data sets, outlier detection, missing values, testing of assumptions and transformation of variables. Model fitting and model interpretation. Model diagnostics.
Design of Experiments
Observational (vs) experimental data. The fundamentals of experimental design. Analysis of variance. Factorial design. Statistical power and multiple comparisons. Non-parametric alternatives.
Multiple Regression
Assumptions, collinearity, interpreting coefficients, model fitting, model diagnostics, confidence intervals of coefficients, Analysis of covariance (ANCOVA).
Generalised Linear Models
Definition of a generalized linear model: link functions. Overview of different types of generalised linear models and their uses with a focus on logistic regression for binary data.
Software analysis
SPSS, R, Excel

DATA8005 - Distributed Data Management
Module Indicative Content
The Big Data Revolution.
Data storage and data process: Historical evolution. New infrastructure, data models and processing techniques required to deal with big data. Main challenges: Capture, store, search, analyse and visualise the data.
Big Data Framework.
Dataset characterisation: Variety, velocity and volume. Data Framework ecosystem overview: Tools to ingest, store, analyse and manage data. Data integration: Extracting, transforming and loading relational and nonrelational-based data.
Data Storage.
Distributed File System. Data nodes vs. name nodes. Data replication and fault tolerance. Cluster manager: component and roles. Large files splitting and distribution algorithms.
NoSQL Data Models.
NoSQL databases arising to tackle problem RDBMS is not good at: Schema-less, high level data representation, scale-out distributed-based infrastructure. CAP theorem. Lost of transactional properties: ACID relational properties vs BASE for NoSQL. Wide range of data models: Pure key/value, columnn-based, document oriented and graph-based. Trade-off between their expressiveness and efficiency. Polyglot persistance: On combining different NoSQL data models for a fit for purpose multi-component system.
Data Processing: Large-scale Analytics.
Text, temporal and geospatial-based datasets. Execution plan: Cluster nodes collaboaration, parallel processing, job scheduling, network transferrence, key/value-based communication. Large-scale analytics libraries: MapReduce and Spark. On comparing their expressiveness and efficiency.

DATA9002 - Distributed Data Management
Module Indicative Content
The Big Data Revolution.
Data storage and data process: Historical evolution. New infrastructure, data models and processing techniques required to deal with big data. Main challenges: Capture, store, search, analyse and visualise the data.
NoSQL Databases.
Alternative to relational databases to address big data challenges. Impedance mismatch, scale-out vs. scale-up. Wide range of data models: Pure key/value, columnn-based, document-oriented and graph-based. Polyglot persistance. CAP theorem, partition tolerance, BASE vs. ACID transactions.
Document-oriented DBs.
Efficient, scalable and resilient data storage: Replication and sharding. Clusters, configuration nodes, shards, chunk of data, shard key range, balancing backgroud operators. Expressive and efficient data queries: JSON-based document representation. Aggregation framework: Commands and pipelines.
Graph-based DBs.
Efficient, scalable and resilient data storage: Property graph data model. Nodes, relationships, properties and labels. Expressive and efficient data queries: Cypher declarative SQL-like language. Graph formalism and optimal path-traversal algorithms. Polyglot persistance: On combining document-oriented and graph-based data models for a fit for purpose multi-component system.
Large-Scale Data Framework.
Storage: Distributed File System. Data nodes vs. name nodes. Large files splitting and distribution algorithms. Analysis: Map-Recude. Divide and conquer algorithm schema. Map-sort-reduce process. Parallel processing. Key/value-based communication. Standard I/O file streaming. Spark: Resilient Distributed Dataset. Transformations and actions, basic API. Lazy evaluation. Context, cluster manager and worker nodes.

DATA8008 - Data Visualisation & Analytics	DATA9005 - Data Analytics & Visualisation
Module Indicative Content	**Module Indicative Content**
Data Visualisation Fundamentals	**Theory and Concepts of Data Visualisation**
History of data visualisation. Understand the various categories used in the field e.g. Information/data/scientific visualisation, infographics, visual analytics. Overview of theory and best practice in these fields, e.g. cognitive amplification, perceptual enhancement and ways to encourage inferential processes.	History of data visualisation. Understand the various categories used in the field e.g. Information/data/scientific visualisation, infographics, visual analytics. Investigate theorists and best practice in these fields, e.g. cognitive amplification, perceptual enhancement and ways to encourage inferential processes.
Data visualisation pre-processing techniques	**Data visualisation pre-processing techniques**
Learn data cleaning techniques relevant to data visualisation - data aggregation, data sampling, impute missing data, find inconsistencies. Learn transformation techniques - data normalisation, construct new variables, Investigate how to use regular expressions and data manipulation techniques to pre-process data sets. Implement these processes using R, Excel or similiar computer package.	Learn data cleaning techniques relevant to data visualisation - data aggregation, data sampling, impute missing data, find inconsistencies. Learn transformation techniques - data normalisation, construct new variables. Investigate how to use regular expressions and data manipulation techniques to pre-process data sets. Implement these processes using R, Excel or a similiar computer package.
Advanced visualisation techniques	**Advanced visualisation techniques**
Investigate and implement computer based tools for visualisation, e.g. dashboard creation with RShiny, Tableau/Qlikview; how these packages can be connected to data sources, e.g. databases.	Research and implement computer based tools for visualisation, e.g. dashboard creation with RShiny, Tableau/Qlikview; how these packages can be connected to data sources, e.g. databases.
Geographic Information Systems (GIS)	**Geographic Information Systems (GIS)**
Investigate and implement GIS software, e.g. QGIS, R, ArcGIS; examine and discuss their features - interactivity, panning, zooming; browser based implementations.	Research, implement and critique GIS software, e.g. QGIS, R, ArcGIS; examine and discuss their features - interactivity, panning, zooming; browser based implementations.
Visualisation and Analytics	**Visualisation and Analytics**
Examine a variety of visualisation, analytical and statistical modelling methods that are used to solve data mining and data analytics problems, e.g. anomaly detection, pattern discovery, network analysis. Investigate clustering techniques e.g. partitioning methods, hierarchical clustering and advanced methods - fuzzy clustering, density based and model based clustering.	Appraise and implement a variety of visualisation, analytical and statistical modelling methods that are used to solve data mining and data analytics problems, e.g. anomaly detection, pattern discovery, network analysis.
Data Analytics Techniques	**Data Analytics Techniques**
Investigate the main pitfalls in data visualisation and data analytics in a real-world setting. Compare and contrast various data analytics techniques.	Investigate, discuss and critque the main pitfalls in data visualisation and data analytics in a real-world setting.

DATA8008 - Time Series & PCA
Module Indicative Content
Principle Component Analysis
PC eigenvalues & eigenvectors, Scree plots, PC Loadings & Scores, Goodness of fit of PC models, Regression and prediction using PCs, Rotations, KMO & Bartlet's test of sphericity
Time series analysis
Decomposition (trend, periodicity, seasonality, white noise), Smoothing Techniques, Stationarity, Autocorrelation, Correlograms, Autoregressive (AR), Moving Average (MA) and mixed (ARIMA) models, R-Square, Stationary R-Square, BIC
Forecasting
Forecast Error, Confidence Intervals, MAE, MAPE, RMSE, Ljung-Box statistic.
Software packages
R, Minitab, Excel, SPSS

STAT9005 - Time Series & Factor Analysis
Module Indicative Content
Factor analysis
Assumptions, Data screening, Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), Structural Equation Modelling (SEM).
Time series analysis
Decomposition (trend, periodicity, seasonality, white noise), Smoothing Techniques, Autoregressive (AR), Moving Average (MA) and mixed (ARIMA) models. Examples of both Pure and Causal Time Series Models. Anomaly Detection in time series (contextual anomalies, anomalous subsequences).
Forecasting
Forecast Error, Confidence Intervals, MAE, MAPE, RMSE, Ljung-Box statistic.
Software packages
R, Minitab, Excel, SPSS

COMP8060 - Machine Learning
Module Indicative Content
Methodology and Evaluation
Application of a standard machine learning methodology using techniques such as dimensionality reduction, model selection, feature selection and hyper-parameter optimization. Overview of evaluation methods such as precision, recall, confusion matrices, learning curves, ROC curves.
Classification Algorithms
Mainstream classification algorithms such as Decision Trees, Ensemble Technique (Bagging and Boosting), Support Vector Machines, Naïve Bayes, Bayesian Networks, Logistical Regression, Instance- Based Learning and Deep Learning.
Regression Algorithms
Introduction to the area of regression. Univariate and multi-variate linear regression, neural networks, ridge regression. How to avoid overfitting through the use of regularization.
Unsupervised Learning Algorithms
Overview of unsupervised learning techniques. Example applications of clustering techniques. Introduction to algorithms such as K-Means, K-Median, DBScan. Optimization and distortion cost function. Random initialization and methods of selecting number of clusters.
Optimization
Introduction to the area of optimization. Categories of optimization such as meta-heuristic and constraint-based optimization. Informed/Uninformed search strategies. Meta-heuristic optimization algorithms. Introduce the concept of heuristic algorithms such as hill climbing, simulated annealing, evolutionary, particle swarm optimization (PSO) and ant colony optimization (ACO).
Case Study
Design and implementation of a relevant case study such as a recommender system.

COMP9060 - Applied Machine Learning
Module Indicative Content
Pre-processing and Evaluation
Application of a standard machine learning pre-processing methodology using techniques such as dimensionality reduction, model selection, feature selection and hyper-parameter optimization. Overview of evaluation methods such as precision, recall, confusion matrices, learning curves, ROC curves.
Classification Algorithms
Learn data cleaning techniques relevant to data visualisation - data aggregation, data sampling, impute missing data, find inconsistencies. Learn transformation techniques - data normalisation, construct new variables. Investigate how to use regular expressions and data manipulation techniques to pre-process data sets. Implement these processes using R, Excel or a similiar computer package.
Regression Algorithms
Research and implement computer based tools for visualisation, e.g. dashboard creation with RShiny, Tableau/Qlikview; how these packages can be connected to data sources, e.g. databases.
Unsupervised Learning Algorithms
Research, implement and critique GIS software, e.g. QGIS, R, ArcGIS; examine and discuss their features - interactivity, panning, zooming; browser based implementations.
Neural Networks and Deep Learning
Appraise and implement a variety of visualisation, analytical and statistical modelling methods that are used to solve data mining and data analytics problems, e.g. anomaly detection, pattern discovery, network analysis.
Optimization
Investigate, discuss and critque the main pitfalls in data visualisation and data analytics in a real-world setting.

DATA8006 - Data Science Analytics Project
Module Indicative Content
Literature Survey
Application of a standard machine learning methodology using techniques such as dimensionality reduction, model selection, feature selection and hyper-parameter optimization. Overview of evaluation methods such as precision, recall, confusion matrices, learning curves, ROC curves.
Core Problem
Formulate the core research question and identify data sets relevant to the chosen data science/analytics area.
Development of Methodology
Formulate and assess viable methodologies and technologies to address the chosen research question with a view to identifying the most appropriate methodologies and technologies.
Project Implementation
Supervised self-directed learning, utilising the identified data science methodologies.
Written Report
Write a professional report that conveys the findings of the research in the chosen area of specialisation. Critically evaluate the results and give appropriate recommendations.
Oral Presentation
Make an oral presentation on the undertaken data science/analytics project.

MATH9001 - Research Methods
Module Indicative Content
Research Skills and Planning Methods
Review research skills including technical writing, research methodology and project planning. Identify a real-world problem using big data and create an appropriate research question. Principles of research and design. Role of hypotheses. Construction and planning of experiments. Plagiarism. Ethics. Intellectual Property. Referencing. Publication Types. Peer Review. Library Catalogue. E-book database and online databases.
Research Principles and Methodologies
Research in a professional context, in support of planning, decision-making and policy analysis; the impact of research. Research principles (eg validity, reliability, generalisability); qualitative and quantitative approaches and rationale (eg empirical, basic, applied, practical and action research).
Reporting of Data Analysis and Management
Data management, cleaning, analysis, interpretation, presentation and preparation.
Thesis Development and Publication
Structure of a Masters Thesis and a peer-reviewed conference/journal. Preparation for conference. Identify an appropriate journal/conference and identify how it may contribute to the wider knowledge in the field.