DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY
COURSEWORK ASSESSMENT DESCRIPTION 2019/20

UNIVERSITY OF HULL

MODULE DETAILS:

Module Number:	600092	Trimester:	T1
Module Title:	Data Mining and Decision Systems		
Lecturer:	Mr Ashley Williamson		

COURSEWORK DETAILS:

Assessment Number:	1	of	1		
Title of Assessment:	Data Mining of Legacy Data				
Format:		Jupyter Notebook (Python Script)	Report		
Method of Working:	Individual				
Workload Guidance:	Typically, you should expect to spend between	40	and	60	hours on this assessment
Length of Submission:	This assessment should be **no** more than: *(over length submissions **will be** penalised as per University policy)*	**8 pages** *(excluding diagrams, appendices, references, code)*			

PUBLICATION:

Date of issue:	14th October 2019

SUBMISSION:

ONE copy of this assessment should be handed in via:	Canvas		If Other (state method)	
Time and date for submission:	**Time**	2pm	**Date**	12th December 2019
If **multiple hand–ins** please provide details*:*	Code submission will be either a single .ipynb file, or a ZIP file containing python scripts. (Canvas) Report submission will be standard PDF (Canvas) Code is the physical artifact to the report writing, and is required as part of the holistic assessment; with the same due date.			
Will submission be scanned via TurnitinUK?	Yes	If submission is to be scanned byTurnitin, these should be one of the allowed types e.g. Word, RT, PDF, PPT, XLS etc. Specify any particular requirements in the submission details Students MUST NOT submit ZIP or other archive formats unless specified. Students are reminded they can **ONLY** submit **ONE** file and must ensure they upload the correct file. Normally only the LAST submission will be considered (and if late incur a late penalty).		

The assessment must be submitted **no later** than the time and date shown above, unless an extension has been authorised on a *Coursework Extension Form: s*ee the Canvas site: Help & Support > Student Forms

MARKING:

Marking will be by:	Student Number

ASSESSMENT:

The assessment is marked out of:	100% for code(40% ACW weighting) 100% for report(60% ACW weighting)	and is worth	60	% of the module marks
N.B If multiple hand-ins please indicate the marks and % apportioned to each stage above (i.e. Stage 1 – 50, Stage 2 – 50). It is these marks that will be presented to the exam board.				

ASSESSMENT STRATEGY AND LEARNING OUTCOMES:

The overall assessment strategy is designed to evaluate the student's achievement of the module learning outcomes, and is subdivided as follows:

LO	Learning Outcome	Method of Assessment *{e.g. report, demo}*
1	*Demonstrate a critical knowledge and understanding of data warehousing, mining and reclamation, data mining systems and expert, decision support systems.*	Report & Code
2	*Critically analyse, research and report on the concepts of data, information and knowledge within a decision support system.*	Report & Code
3	*Develop an appropriate decision support tool knowledge using a rigorous data mining methodology for a complex information scenario*	Code
4	*Select, justify and use appropriate approaches, including some at the forefront of the subject / profession, to identity the impact that data mining and decision systems have on an organisation*	Report

Assessment Criteria	Contributes to Learning Outcome	Mark
ACW Report - PDF	1, 2, 4	100%(36% of module)
ACW Code - Python Script/Notebook (Zip)	1, 2, 3	100%(24% of module)

FEEDBACK

Feedback will be given via:		Feedback will be given via:	Canvas
Exemption	None		

(staff to explain why)	
Feedback will be provided no later than 4 ‘teaching weeks’ after the submission date.	

This assessment is set in the context of the learning outcomes for the module and does not by itself constitute a definitive specification of the assessment. If you are in any doubt as to the relationship between what you have been asked to do and the module content you should take this matter up with the member of staff who set the assessment as soon as possible.

You are advised to read the **NOTES** regarding late penalties, over-length assignments, unfair means and quality assurance in your student handbook, which is available on Canvas.

In particular, please be aware that:

- Up to and including 24 hours after the deadline, a penalty of 10%
- More than 24 hours and up to and including 7 days after the deadline; either a penalty of 10% or the mark awarded is reduced to the pass mark, **whichever results in the lower mark**
- More than 7 days after the deadline, a mark of zero is awarded.

- The overlength penalty applies to your written report (which includes bullet points, and lists of text. It does not include contents page, graphs, data tables and appendices). 10-20% over the word count incurs a penalty of 10%. Your mark will be awarded zero if you exceed the word count by more than 20%.

Please be reminded that you are responsible for reading the University Code of Practice on Academic Misconduct through the Assessment section of the Quality Handbook (via the SharePoint site). This govern all forms of illegitimate academic conduct which may be described as cheating, including plagiarism. The term ‘academic misconduct’ is used in the regulations to indicate that a very wide range of behaviour is punishable.

In case of any subsequent dispute, query, or appeal regarding your coursework, you are reminded that it is your responsibility to produce the assignment in question.

Description of assessment task.

Data Mining and Decision Systems (600092) Coursework

Data Mining of Legacy Data

Available: 14th October 2019

Submission (60% of Module Marks)

- Code Submission (Python Script / Jupyter Notebook) - 40% of ACW
- Report PDF Submission (Turnitin) - 60% of ACW

Deadline: 12th December 2019

A template for the report submission is available via Canvas. The template contains headings and guidelines for each section, detailing what needs to be presented. Feedback will be delivered on Canvas for each submission item individually, with reference to the criterion reference grid.

Aims

This coursework provides assessed practical experience in handling a data-mining project using industry tooling for legacy data. The Data provided as part of this assessment is from the domain of cardio-vascular medicine. The work herein requires the description and analysis of data for the given domain, including manipulation of data in various forms, as well as creation of classifiers. Skills developed towards this assessment provide experience on applying data mining techniques to real world (albeit simplified) data.

Software utilised as part of this assessment submission, Python and Jupyter, are installed on Fenner A, B, C machines and are commonly available on the internet.

Raw data for the project is available from the module Canvas site as a CSV file, with an accompanying Data Description. The given data is synthetic, but derived from a real-world data set.

You are expected to filter, clean, and transform data as appropriate such that it can be used to produce optimal classification for patient risk. {NoRisk, Risk}.

Coursework Specification

Code submission

Code submission represents the physical artefact demonstrating application of Data Mining techniques towards the provided dataset. This submission should be a Jupyter Notebook (.ipynb) responsible for following a given Data Mining Methodology towards the aims above. This will include any data reading, manipulation, selection, visualisation, model creation, and evaluation.

The notebook, when assessed, will be cleared of output and re-ran from the first cell; ensure that the submitted notebook is capable of this, as any existing output will not be considered. It is expected that code be sufficiently commented, especially where documentation was referenced. Markdown cells may be used to differentiate sections of the notebook for ease of reading and use.

Report submission

A report outlining the followed methodology, detailing steps taken towards processing of the data, including the results of models generated. Each phase of the implemented methodology should be evidenced within the report, making reference and use of elements of the code submission. Figure 1 outlines the CRISP-DM methodology shown as part of the lecture content.

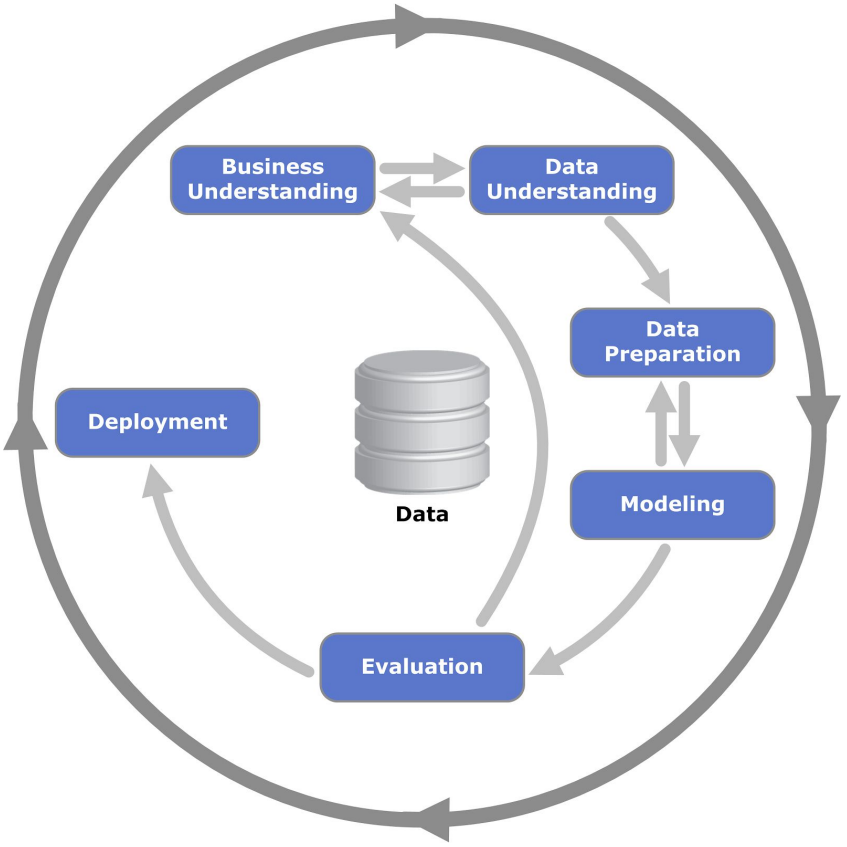

Figure 1 - The Standard CRISP-DM Data Mining Methodology.

This report contains a 8 page maximum limit, excluding front page, references, and appendices. This limit constitutes a maximum, and is no indication on the number of pages that is expected to be submitted. It is sufficiently large to provide flexibility in the use of figures, comparison, reporting, and discussion.
Further information can be found in the template provided on Canvas.

Given Data Description

Table 2 outlines the given data description for the legacy data provided. It is worth noting that the actual data may not conform to this fully, and exceptions may exist. Knowledge from lectures surrounding legacy systems, data corruption, etc may prove useful in identifying this.

Table 2 - Given Data Description for the Legacy Data Provided.

Attribute	**Value Type**	**NumberOfValues**	**Values**	**Comment**
Random	Real	Number of Records	Unique	Real number of help in randomly sorting the data records
Id	Integer	Max of Number of Records	Unique to patient	Anonymous patient record identifier: Should be unique values unless patient has multiple sessions
Indication	Nominal	Four	{a-f, asx, cva, tia}	What type of Cardiovascular event triggered the hospitalisation?
Diabetes	Nominal	Two	{no, yes}	Does the patient suffer from Diabetes?
IHD	Nominal	Two	{no, yes}	Does the patient suffer from Coronary artery disease (CAD), also known as ischemic heart disease (IHD)?
Hypertension	Nominal	Two	{no, yes}	Does the patient suffer from Hypertension?
Arrhythmia	Nominal	Two	{no, yes}	Does the patient suffer from Arrhythmia (i.e. erratic heart beat)?
History	Nominal	Two	{no, yes}	Has the patient a history of Cardiovascular interventions?
IPSI	Integer	Potentially 101	[0, 100]	Percentage figure for cerebral ischemic lesions defined as ipsilateral
Contra	Integer	Potentially 101	[0, 100]	Percentage figure for contralateral cerebral ischemic lesions
Label	Nominal	Two	{risk, norisk}	Is the patient at risk (Mortality)?

In short, a non-clinical description follows:

Random - Real number of help in randomly sorting the data records: Should be unique values.
Session - Anonymous patient session identifier: Should be unique value. Patient can have multiple sessions
Id - Anonymous patient record identifier: Should be unique value per patient. Patient can have multiple sessions
Indication - What type of Cardiovascular event triggered the hospitalisation?

- a-f : Atrial-Fibrillation
- asx : Asymptomatic Stenosis
- cva : Cardiovascular Arrest
- tia : Transient Ischemic Attack ("mini-heart attack")

Diabetes - Does the patient suffer from Diabetes?
IHD - Does the patient suffer from Coronary artery disease (CAD), also known as ischemic heart disease (IHD)?
Hypertension - Does the patient suffer from Hypertension?
Arrhythmia - Does the patient suffer from Arrhythmia (i.e. erratic heart beat)?
History - Has the patient a history of Cardiovascular interventions?
IPSI - Percentage figure for cerebral ischemic lesions defined as ipsilateral
Contra - Percentage figure for contralateral cerebral ischemic lesions
Label - Is the patient at risk (Mortality)?

600092 Data Mining and Decision Systems - CRG

ACW - Data Mining of Legacy Data

UNIVERSITY OF HULL

Learning Outcome	Criterion	Pass	2:2	2:1	1st	Upper 1st
[LO1] Demonstrate a critical knowledge and understanding of data warehousing, mining and reclamation, data mining systems and expert, decision support systems.	Outline of how the methodology was applied to the process of Legacy Data Mining with evidence supported by artefact. (20%)	An outline of the methodology followed is presented, but may be shallow in depth, or contain some factual errors. No evidence supporting methodology phases is presented.	A complete outline of the methodology followed is presented, with some sections evidenced back to code.	In addition to previous. Each phase of the methodology followed is supported by empirical evidence taken from code. Methodology phases are contextualised to the problem domain.	In addition to previous. Methodology is well outlined, with sections broken down into sub-sections with strong evidencing of process. Contextualisation of the process as applied to the problem domain is well-presented.	In addition to previous Deployment and Business Understanding phases are presented as a hypothetical scenario and contextualised to the given domain. Alternative and/or additional methodologies are evidenced, and fully justified, alongside their interactions with the current methodology.
[LO2] Critically analyse, research and report on the concepts of data, information and knowledge within a decision support system.	Presentation of results tables, and any accompanying figures. (10%)	A basic results section is presented. This may cover basic data relationships or metrics.	In addition to previous. Single model classification accuracy is presented.	In addition to previous. Multiple models are evaluated, with TP, TN, FP, FN metrics. Additional figures and/or tables may be present.	In addition to previous. Additional fit-for-purpose metrics are reported, with results tables enabling comparison of different input sets for each model evaluated. Additional figures are presented appropriately to highlight any additional data patterns.	In addition to previous. Results presented utilise a variety of fit-for-purpose metrics, and contain interesting entries to invoke in-depth discussion and comparison.

UNIVERSITY OF HULL

[LO4] Select, justify and use appropriate approaches, including some at the forefront of the subject / profession, to identity the impact that data mining and decision systems have on an organisation.	Critical evaluation of results with discussion surrounding the process. (30%)	Results are interpreted correctly in the context of the assignment.	In addition to previous. Results are interpreted correctly, with some discussion relating back to the domain. Rationale may be provided but not fully justified.	In addition to previous. Rationale is provided for chosen metrics, with correct justification, relating these metrics back to the domain. Comparisons between models, and their results, is provided but may not be fully justified.	In addition to previous. Results are thoroughly evaluated, referencing several metrics, and comparing various data preprocessing approaches towards the objective. Discussion critically reflects on the application of the methodology towards the assignment.	In addition to previous. Discussion is thoughtful and critically reflects on aspects of the methodology highlighting advantages and disadvantages of certain stages. Alternative Methodologies and practices are considered and contextualised to the problem task.
[LO3] Develop an appropriate decision support tool knowledge using a rigorous data mining methodology for a complex information scenario.	Demonstrate the use of Python Jupyter Notebooks in a Data Mining Context towards Mining of Legacy Data (40 %)	Basic data cleaning, and transformation is achieved, but may be erroneous or incomplete in places. A single model is created, but may be incorrectly utilised.	In addition to previous. All Data errors are repaired, with data cleaned. Transformations may not be most appropriate. A single model is created and trained on the processed data.	In addition to previous. Multiple models are trained on the processed data. Evaluation metrics are generated by predicting on a trained model. This is done utilising basic data partitioning. Notebook is laid out logically following the phases of the methodology followed.	In addition to previous. Markdown cells are used to better differentiate sections and introduce code segments. Advanced Data partitioning is performed with data-driven metrics considered for the partition. Python code is well-documented, with reference to documentation where appropriate. Multiple methods of data preprocessing for model creation are implemented, alongside advanced model training techniques.	In addition to previous. Data is processed efficiently and effectively, showing a clear understanding of library use. Model hyper-parameters are explored and expertly determined from data-driven processes.
Weighting	All criteria are weighted as shown by the percentages indicated in the relevant criterion box.					