Health Informatics and Analytics Programmes

Core

Health Data Collection and Preparation

(HIA302)

4 credits

Module Guide

TITLE: HEALTH DATA COLLECTION AND PREPARATION
COURSE CODE: HIA 302
PROGRAMME: Health Informatics and Analytics

CONTENT SYNOPSIS:

This module focuses on the fundamentals of data collection and preparation (part of "data engineering") in a healthcare environment.

Topics covered include:

- introduction to data science,
- introduction to data storage and management,
- introduction to data engineering, data exploration and data sourcing,
- data transformation and cleaning,
- Python for health data collection and preparation.

LEARNING OUTCOMES:

On completion of this module, students will be able to:

- Demonstrate the concepts of data collection and preparation in a healthcare environment. (C1, PLO2)
- Develop skills to deal with data collection, storage, and processing of data. (C3, PLO3; C3, PLO6)
- Propose a solution to a health-related data science problem using the Python programming language and libraries. (C2, PLO1; C3, PLO3)

LEARNING HOURS

Teaching Modality	Contact Learning (Hour)		Independent Learning (NF2F)	SLT
	Guided Learning (F2F)	Guided Learning (NF2F) eg: e-Learning		
Lectures	7	7	42	**56**
Workshops	14	7	27	**48**
Case Studies	3	3	11	**20**
Group Project			32	**32**
Presentation	4			**4**
TOTAL				**160**

ASSESSMENT

Group Project 60%
- Written Report 40%
- Presentation 20%

Individual Project (report) 40%

MODULE LEADER: Dr Tan Ee Xion

ASSOCIATED LECTURERS: Prof Patrice Boursier, Dr Chuah Tong Kuan, Dr Wong Siaw Ming, Thinaharan Ramachandran

SYLLABUS

1. Introduction to data science *(Lecture- 1 hour)*
   This section will provide a historical perspective on data science.
   Students will be introduced to "big data" and the different phases of data science:
   a) data collection and preparation, or data engineering,
   b) data analytics,
   c) data visualisation.

2. Introduction to data storage and management *(Lecture -3 hours, Case Study – 3 hours)*
   This section will provide a historical perspective on data management. Students will be introduced to:

a) data base management systems (DBMS),
b) relational DBMS and SQL, NoSQL databases,
c) cloud storage (public, private, hybrid),
d) data warehouses vs data lakes,
e) big data frameworks (Hadoop, Amazon AWS, Microsoft Azure).

3. Introduction to data engineering, data exploration and data sourcing *(Lecture- 3 hours; Workshop - 2 hours)*
This section will provide a historical perspective on data collection and preparation. Students will be introduced to:
a) data sourcing and data sources,
b) contents and formats,
c) data quality.

4. Data transformation and cleaning (*Workshop- 2 hours)*
This section will discuss the scope of data preparation as well as methods for improving, enriching and formatting data. Students will be introduced to:
a) data extraction methods,
b) data cleaning,
c) data integration / blending,
d) data transformation.

5. Python for health data collection and preparation (Workshops- 10 hours)
This section will introduce students to using the Python programming environment and libraries for data collection and preparation. It includes hands-on workshops that are based on small assignments which relate to real-world problems.

6. Individual Project (Python, 20 hours)
Students will work individually on a selected case topic and submit a report (5-10 pages with cover page, summary, discussion, conclusion and references).

7. Group Project (32 hours)
Students will work in group on a selected project, each student developing his/her own part. Each group will submit a written report (10-20 pages with cover page, summary, discussion, conclusion and references). It shall also include a section at the beginning explaining what has been done by each individual. There will be a group presentation (about 30-45 minutes) with specific questions for each individual.

READING LIST

1. Eric Topol. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books, 2019.

2. Lillian Pierson. Data Science for Dummies. For Dummies, 3rd Edition, 2021.

3. Sergio Consoli, Diego Reforgiato Recupero, Milan Petković. Data Science for Healthcare: Methodologies and Applications. 2019 Edition.

4. Mowafa Househ, Andre W. Kushniruk, Elizabeth M. Borycki. Big Data, Big Challenges: A Healthcare Perspective: Background, Issues, Solutions and Research Directions. Lecture Notes in Bioengineering, Springer, 2019.

5. Robert Hoyt, Robert Muenchen. Data Preparation and Exploration: Applied to Healthcare Data. Informatics Education, 2020.

6. Michael Walker. Python Data Cleaning Cookbook: Modern techniques and Python tools to detect and remove dirty data and extract key insights. Packt Publication, 2020.

7. Stephen Klosterman. Data Science Projects with Python: A case study approach to successful data science projects using Python, pandas, and scikit-learn. Packt Publishing, 2019.

8. Ethan Williams. Python for Data Science: The Ultimate Beginners' Guide to Learning Python Data Science Step by Step. Independently published, 2019.

9. David Mertz. Cleaning Data for Effective Data Science: Doing the other 80% of the work with Python, R, and command-line tools. Packt Publishing, 2021.