Project description for Master Data Science Dissertations

(*MSc Applied Data Science* and *MSc Data Science and Its Applications* – January start 2023)
Prof Berthold Lausen blausen@essex.ac.uk , 7th / 8th September 2023)

Master students: Balakishan, Dhananjay, Gaurav A, Gaurav B, Gyan, Jaykumar, Priyadharshini, Sourav, Rohit, Syed.

Project group allocation:

[Group B]	**Data Breast:**	**Gaurav A, Gaurav B, Sourav, Syed;**
[Group C]	**Data Colorectal:**	**Balakishan, Dhananjay, Rohit;**
[Group L]	**Data Lung:**	**Gyan, Jaykumar, Priyadharshini.**

As explained in person and communicated on Monday, 4 September 2023, the detailed guidance on your projects and next steps as follows:

Research Questions and Methodology:

A. **Data analysis of clinical baseline data** of one cancer (identified to each of the three master student groups [one per group]: Breast, Colorectal, Lung) from **The Cancer Genome Atlas (TCGA)** with the **response variables** – if available - *cancer staging*, especially invasive vs non-invasive cancer (Logistic regression, …) and **overall survival (OS)** as response variables (Kaplan-Meier-Estimator, etc.). Descriptive summary of all clinical predicator variables as well. Make an informed decision which clinical variables (lymph node status, number of nodes look at, …) are available (clinical importance, percentage missing values, etc.) for a meaningful analysis (Research questions: A, B, C, D, E, if choosen F as well).

B. **Supervised learning/predictive modelling.** Response variables *cancer staging*, especially invasive vs non-invasive cancer and *overall survival (OS)* (identify a method of interest each master student individually by suggesting a recent research publication (peer reviewed journal) – to be agreed with supervisor online by Thursday, 14 Sept 2023 - latest), comparison with at least two alternatives, e.g. neural network/Deep Learning approach and random (survival) forest. Resampling for confidence / stability assessment of results. Genomic predicator variables: Gene expression, you may like to consider Proteome expression (if available – optional. Make a careful decision, if you have enough time to research Proteome together with Genome).

C. **Unsupervised learning/clustering of Gene expression** (about 22000 genes), if available you may include Proteome expression. Resampling for confidence / stability assessment of results.

D. **Use cluster (from C) based new engineered predictors** in the supervised learning/predictive modelling (from B). You may use other methods for unsupervised

dimension reduction (PCA, etc.) in addition to cluster analysis (optional. Make a careful decision, if you have enough time to include this in your research).

E. Identify and apply the 'best' data processing pipeline (A to D) to **data from a second cancer** (identified to and agreed with each master student individually by 27th September 2023). Evaluate and assess predictors found the second cancer data sets on the data of the first cancer data sets (B, C and L - depending on your project group).

F. **Optional:** Further Data Science methodology answering an additional research question. For example (optional) NLP can be used to derive further predictors based on textual data. It may be a good idea to identify textual data for a selected set of important genes … say 1000 of 22000. You may need to code how to read in the textual information of these – say – 1000 Genes. Of course if you think of a Deep Learning Approach … you may go for all Genes. ***Commit time to F only after you have done (and written up!) successfully A to E.***

Data and next steps:

1. Access TCGA data (start to work as a group of three or four master students on this) from

   The Cancer Genome Atlas (TCGA):
   https://www.cancer.gov/ccg/research/genome-sequencing/tcga

   Data base: Genomic Data Commons Data Portal (GDC):
   https://portal.gdc.cancer.gov/

   Use Bioconductor to generate clinical and gene expression data sets using R package TCGAbiolinks:
   https://bioconductor.org/packages/devel/bioc/vignettes/TCGAbiolinks/inst/doc/index.html

2. Email a suggestions of names to work as group of three master students (two groups) or four master students (one group) by 5 September 2023 (2 pm).Groups will be finalised on **Thursday 7 September 2023 (meeting 14:00 to 16:00 – meeting at STEM 4.1.**).

3. Email individually a suggestion for one research paper (as explained above) by **Wednesday 13 September 2023 (2 pm).** Two research papers (at least one in addition to your suggestion) as starting point of literature review will be agreed and aim to be finalised meeting online by Thursday 14 September 2023.

4. Email a first draft of your project and progress with accessing and describing the data (descriptive statistics, basic visualisation) by **Thursday 14 September 2023 (1 pm).**

Further timeline:

Monday 25th September 2023, 14:00 BST – email individually draft of three chapters/sections description of your project, literature review and data description (descriptive statistics, basic visualisation); R, Python, etc. code as appendix.

Wednesday, 27th September 2023, 14:00 – 16:00 third group meeting with feedback to each of the three groups. Room to be confirmed in due course.

Tuesday, 24th October 2023, 14:00 – 16:00 fourth group meeting with presentations of and feedback to each of the three groups. Room to be confirmed in due course.

Monday, 6th November 2023, 14:00 email individually your full draft (with drafted appendix and drafted supplement) of your dissertation.

Individual Feedback aim by Friday 10 November 2023 noon.

FASER submission of dissertation, R/Python code supplement document by the deadline Friday 24th November 2023, noon.