Data Science (Eng) 774/874
Post-block Assignment 1

11 March 2021

Kindly complete the following assignment and submit your assignment as an electronic submission in PDF format on SUNLearn or SUNOnline by **23:50 on 02 April 2021.**

1. **Assignment [45]**

A healthcare facility has recently employed you as a Data Scientist and has tasked you to develop and implement an algorithm that can predict the likelihood of a patient suffering a stroke. For this matter, you have been provided with the Stroke Prediction Dataset, *healthcare-dataset-stroke-data*.

Conduct an Exploratory Data Analysis (EDA) that includes the following tasks:

- Study each feature individually
- Address all data quality issues and provide a data quality report before cleaning the dataset.
- Provide a data quality report after cleaning the dataset.
- Determine if there exists a relationship/correlation between features using a visual representation. (Hint: You can use scatter plots).

After the completion of the EDA, the data are clean and ready to be fed to a predictive algorithm. As a Data Scientist, which modelling technique would you select to achieve your goal and why?

Note: Your report must be aligned with the processes that were covered in the Data Understanding and Data Preparation phases of the CRISP-DM methodology. Ensure that you deal properly with features that contain missing values and explain the strategy you will use solve the issue.

2. Evaluation rubric

Your exploratory data analysis will be marked according to the following rubric:

Criteria		**Excellent - above average**			**Not at desired level of competency**
	Weight	**4**	**3**	**2**	**1**
Individual analysis of features	10	All features are studied, plotted, data types are defined correctly.	Minor errors in 1 aspect or feature.	Major errors in less than 5 aspects or features.	More than 5 major errors.
Data Quality Issues have been addressed	8	All features that possess missing (or NaN) values have been correctly identified.	Minor errors in less than 2 aspects or features.	Major errors in less than 5 aspects or features.	No data quality issue was addressed.
Data Quality Report – Before Cleaning the data.	8	A clear and complete data quality report has been provided. (before the data was cleaned)	Minor errors in 1 aspect or feature.	Major errors in less than 5 aspects or features.	More than 5 major errors.
Data Quality Report – After Cleaning the data	8	A clear and complete data quality report has been provided (after the data was cleaned)	Minor errors pertaining to 1 feature	Major errors in less than 5 aspects or features.	More than 5 major errors.
A modelling technique has been suggested	6	A modelling technique has been suggested and an adequate explanation in support of the choice has been provided	A modelling technique has been suggested with no supporting evidence	N/A	No modelling method was suggested
General impression of the report	5	Well-structured immaculate report.	Well-structured report with minor spelling errors	Disorganized report with several grammatical errors	The report is substandard and not at Postgraduate level