Assignment

- **Deadline for submission: Monday, 13th May 2019, 12:00 midday**
- **Provisional mark and feedback release date: Monday 10th June 2019, 17:00pm**
- Length: The summative assessment refers to the outcome of the entire module and culminates in a documented paper of 2000-2500 words and the submission of an iPython-notebook.
- Select one dataset in health and biomedicine of your choice from the provided list!

Datasets for the Assignment

Diabetic Retinopathy Debrecen Data Set Data Set

http://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set

ILPD (Indian Liver Patient Dataset) Data Set

http://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29

Mammographic Mass Data Set

http://archive.ics.uci.edu/ml/datasets/mammographic+mass

Assignment

The assignment should follow this outline:

1 Introduction		Words
	Please describe the background, context, and importance of the data in light of related literature. Show a sound interpretation of the medical problems presented in the data. Outline the selected dataset (including features and class labels) and provide descriptive statistics of the contained variables. Visualise the feature space in a plot and explore the underlying characteristics. [20 marks]	500

Assignment

2 Methodology	Words
• Describe the data cleansing, feature selection, feature construction, and feature pre-processing of the chosen dataset. • Select two supervised models of the course. Give a high-level description of both algorithms including their pseudo-code. Describe and demonstrate for both classification algorithms: a) supervised learning on training data b) optimization of hyper-parameters c) model evaluation including but not limited to criteria such as confusion matrix, precision, recall, F1, and AUC Demonstrate your solution with an attached iPython notebook. Ensure reproducibility and transparency by using an URL with the original dataset. [30 marks]	750

Assignment

3 Results		Words
	Present optimized hyper-parameters and reasonable evaluation criteria such as a confusion matrix, precision, recall, F1, AUC and a ROC-plot. Provide a sensitivity analysis for both algorithms with different parameters and give a textual description of the results. [30 marks]	750
4 Discussion and Conclusion		Words
	Compare and discuss your findings with other scientific publications that used the same medical dataset. Discuss how you would improve your methodology. Outline the potential usage of the trained algorithms in healthcare and health service delivery. What benefits might be anticipated from their deployment? [20 marks]	500

Assignment

5 Appendix		
	Attach a reproducible iPython Jupyter notebook	
Total: 100 marks		

Assessment

Markers will look for the following sections in the assignment:

- sound understanding of the provided dataset and appropriate pre-processing to obtain a suitable feature space
- appropriate selection and learning/training of a classification algorithms to address the decision problem in healthcare;
- evaluation of the performance and prediction on unseen health data instances