TU Dublin, Tallaght Campus
Enterprise Computing and Digital Transformation

HCAIM – Human Centric Deep Learning 2022/23
CA1 (25% Weighting)

CA Overview:

Students will be given a unique data set, where each student must use the dataset with the student number on it (this can be found on Moodle in the CA). The full details of the dataset can be found in Appendix 1.

The students will be expected to demonstrate and apply fundamental techniques/algorithms for the prediction and training of neural networks, including investigations for model bias. Compare, contrast and demonstrate appropriate hyper-parameters for training artificial neural networks with an emphasis on reproducibility (generalizability), transparency and interpretation. Use Deep Learning frameworks to implement Deep Learning models for classification and or regression tasks. This includes a strong focus on evaluation of the models, for performance and ethical considerations, resulting in transparaency and explainable AI.

This CA requires you to develop the most suitable AI model, where every suitable method for model development be employed to ensure the best generalization.

CA Points:

All steps should be separated in a Jupyter notebook (sections), and two versions of the notebook are required on final upload of the elapsed CA: HTML(or PDF) and the .ipynb Jupyter notebook.

Note: The HTML(or PDF) will be run through URKUND, for plagiarism evaluation, where the institution's plagiarism policy will be applied[1].

Students are expected to discuss each piece of work, expanding on the rationale for selection of techniques and the processes applied to get to the final model, including the use of any visual aids such as tables, plots etc. The word count is a minimum of 3000 words. Code on its own, will not be considered as a passing grade. A random selection of students may be invited to a voice viva for academic integrity.

Please add your name and student number as the first block in markdown on the notebook.

[1] https://www.tudublin.ie/for-students/student-services-and-support/student-wellbeing/disability-support-services/tallaght/open-book-exam-study-tips/policy-on-plagiarism/#:~:text=Plagiarism%20Policy%20Statement%20%2D%20TU%20Dublin%20%2D%20Tallaght&text=Globally%2C%20plagiarism%20is%20recognised%20as,recognition%20to%20the%20original%20source.

CA Headings:

The CA will be graded under the following headings. Where each heading should have (where appropriate) a rationale and description expanding the approach and its findings, code that executes the work, and in some cases a visual aid to further compound your findings. Use standard Jupyter notebook markdown, to crate headings and sections. A grading Rubric is also provided in Appendix 2. The following headings should guide your project:

- Introduction
- Opening the dataset, Brief data exploration and data pre-processing
- Model exploration to determine network topology
- Hyperparameter investigation (learning parameters and optimization)
- Most appropriate model selection
- Grid search
- Final Model presentation and performance evaluation
- Analysis of performance per class
- Analysis of performance for target group
- Summary and Conclusion

Quality of report and model development process will also be graded.

CA Notes:

At all stages, precautions must be taken (checkpointing) to ensure work is competed by the CA deadline.

It is up to the student to select the notebook platform (such as locally, Google CoLab or on Azure Data Science Virtual machines) and it is the student's responsibility to ensure that the platform has the required version of Python, Anaconda and required libraries.

Items such as the grid search results and model epoch training results should be left visible in the notebook (that is verbose=0 is not applied) so that they are in the final uploads.

Appendix 1

Dataset URL:
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. For this project, a csv of the dataset available on Kaggle for the year 2015 was used. We are using the dichotomous dataset (0,1) for the class label **Diabetes_binary.**

Attribute:

Attribute 1 : HighBP
Attribute 2 : HighChol
Attribute 3 : CholCheck
Attribute 4 : BMI
Attribute 5 : Smoker
Attribute 6 : Stroke
Attribute 7 : HeartDiseaseorAttack
Attribute 8 : PhysActivity
Attribute 9 : Fruits
Attribute 10 : Veggies
Attribute 11 : HvyAlcoholConsump
Attribute 12 : AnyHealthcare
Attribute 13 : NoDocbcCost
Attribute 14 : GenHlth
Attribute 15 : MentHlth
Attribute 16 : PhysHlth
Attribute 17 : DiffWalk
Attribute 18 : Sex
Attribute 19 : Age
Attribute 20 : Education
Attribute 21 : Income
Attribute 22 : Diabetes_binary (Class label)

Appendix 2 – Grading Rubric

1. Introduction
2. Opening the dataset, Brief data exploration and data pre-processing
3. Model exploration to determine network topology (8 required)
4. Hyperparameter investigation (8 batch sizes [suitable epochs], 3 optimizers, and 2 or more regularization techniques)
5. Most appropriate model selection
6. Grid search (best 4 batch sizes, best two optimizers, one weight initializer, and three epochs [you can use early stopping])
7. Final Model presentation and performance evaluation
8. Analysis of performance per class
9. Analysis of performance for target group (two groups required)
10. Summary and Conclusion
11. Quality of report and model development process

ID	Marks	Poor example	Average Example	Strong example
1	3	Little to no discussion on the dataset and its overall structure.	Some discussion on the dataset and/or structure.	Strong understanding of the dataset and its structure presented.
2	5	Dataset opened, but issues may exist, little to no data exploration and no pre-processing.	Dataset opened, and some pre-processing completed, little or no discussion on the process.	Daset opened, data was explored, correct preprocessing applied and a strong discussion on the process.
3	10	One or more networks developed, no rationale for the selection of models or the order of development	3-5 networks developed, some rationale for the selection of models or the order of development	8 networks developed, strong rationale for the selection of models and the order of development
4	20	Little or no hyperparameters trailed, no concise discussion on the order of trails and why one was selected over another	Some hyperparameters trailed, some concise discussion on the order of trails and why one was selected over another	All hyperparameters trailed, strong concise discussion on the order of trails and why one was selected over another
5	5	Little to no (or incorrect) rationale for the selection of the best model, or incorrect model selected	Some to no (or partially incorrect) rationale for the selection of the best model, or incorrect (but close) model selected	Strong rationale for the selection of the best model, and best model selected
6	20	Grid search not completed or failed	Grid search completed with less arguments and/or little to no discussion on perfomance	Grid search completed with strong discussion on performance
7	3	Did not rerun the model from the gridsearch	Used gridsearch model but little to no discussion	Used gridsearch model and provided discussion on the performance compared to the gridsearch/model development
8	10	Suitable metrics generated but no discussion on the meaning of the results	Suitable metrics generated with some discussion on the meaning of the results	Suitable metrics generated with strong discussion on the meaning of the results
9	20	Little to no generation of metrics per target group and/or little to no discussion on the findings/meaning	Generation of metrics per target group and little to no discussion on the findings/meaning	Generation of metrics per target group and strong discussion on the findings/meaning
10	4	Little to no conclusion or nor indicative of the findings prior	Some conclusions and in part indicative of the findings prior	Strong conclusions and indicative of the findings prior
Total	**100**			