Overall notes for the assessments:

1. Justify your answers and support your work with relevant literature and practical examples, linking theory and practice to substantiate arguments showing evidence of research surrounding the main argument or topics of the question(s)*

2. Any use of AI should be clearly identified, including the prompts used, at the end of the assignment (not part of the word count). When using AI tools, students are expected to critically engage with the information provided by any AI tools (a critical reader always questions the information and points of view presented by the "writer" in a text and the context in which they are presented and engages in further reading / research.) Just using AI to do your work for you will be heavily penalised!

3. Correct paraphrasing of work of other authors, citing in-text and referencing according to the British Harvard Referencing standard throughout the assignment is expected. *

4. Please ensure to include the IDEA Academy cover sheet and adhere to the required assignment presentation formats the clarity and readability. *

5. Students are to refer to guidelines stated in Document 017_23 – Recognising and Avoiding Plagiarism Policy and Procedure and Document 069_22 - Guidelines for the Presentation of Assignments, available on Canvas under IDEA Academy Policies, Documents and Forms.

Task 1: Discussion [20 marks]

General Guidelines

- Please answer only **ONE** of the questions presented below.
- Word count for this discussion is 650 ± 10% words.
- Respond to a minimum of **TWO** maximum **THREE** students within discussion thread (200 ± 10% words).
- Use at least **TWO** references and citations within your posts as per Harvard Referencing style.

Discussion Question 1

Consider the four different types of Artificial Intelligence systems and, answer the following two questions:

- If an Artificial Intelligence algorithm could be applied within your field, which AI type would you see most fit and why?
- How do you think this AI solution could be applied within your professional context?

Discussion Question 2

Consider the two sentences below:

- "Data comes in many formats but it's all numerical in the end"
- "Many ML algorithms use a gradient based optimisation techniques at their core but the hyperparameters are tuned using a gradient free optimisation approach".

What do you think these statements mean? You can refer to a specific dataset or algorithm as an aid to your discussion.

Discussion Question 3

For each of the following model evaluation techniques, specify which techniques can be used to validate a regression or classification problem or both. By referring to a specific case study discuss how these evaluation techniques are useful to measure/compare performance.

- Cross Validation
- Confusion Matrix
- Specificity and Sensitivity
- Bias and Variance
- ROC and AUC

Grading Rubric:

Discussion Question	**Criteria**	**Maximum Marks**
1	Justification of AI algorithm fitness to the field	5
	Detailed description of algorithm.	5
	Clarification of how AI algorithm provides a solution.	5
	Discussion on model validation techniques.	5
2	Detailing techniques used to convert non-numerical data to numeric.	5
	Detailed how numeric data is transformed or normalised to improve algorithms.	5
	Clear distinction between gradient free and gradient based optimisation	5
	Use cases in which algorithms in AI make use of gradient free and gradient based optimisation.	5
3	Clear description of each technique	5
	Clear description of whether they can be used in classification or regression or both.	5
	Correct identification of respective use cases	5
	correctly relating some of the metrics to underfitting and overfitting	5

Task 2: Write up [80 marks]

You are required to use KNIME Analytics Platform to provide a solution to the classification problem described in the following section. Although aiming to get the best classification results is the usual target, you will not be penalised if the model you provide has a low accuracy. Use both the SVM and MLP Learners and compare results achieved.

You are encouraged to start from an attempt that includes the minimum number of nodes for a functional solution and then progress to improve and include:

- Automatic optimisation of hyperparameters.
- Cross validation.
- Export the Learned models in PMML format.

More specifically, you are required to:

1. Provide one functional KNIME workflow with both SVM and MLP solutions.
(10 marks)

2. Write a structured report documenting the process, which includes:
    a. Explore the data by making use of Visualisation tools.
    b. Determine which of the input features are more important for this classification problem.
    c. Discuss which features make the classes linearly separable.
    (20 marks)

    d. Justify the nodes and configurations used.
    e. Discuss why partitioning the data into train and test is important.
    f. Clearly explain the roles of the Learner and the Predictor.
    (10 marks)

    g. Give some detail regarding the inner workings of MLPs and SVMs.
    (10 marks)

    h. Visualise and analyse the results.
    i. Interpret the results obtained by the automatic optimisation of hyperparameters.
    (10 marks)

    j. Discuss how cross validation (k-fold) helps increase confidence in the classification results.
    k. Make use of references to substantiate claims especially for points e, f, g, and j.
    (10 marks)
    l. The web links to both PMML files. (10 marks)

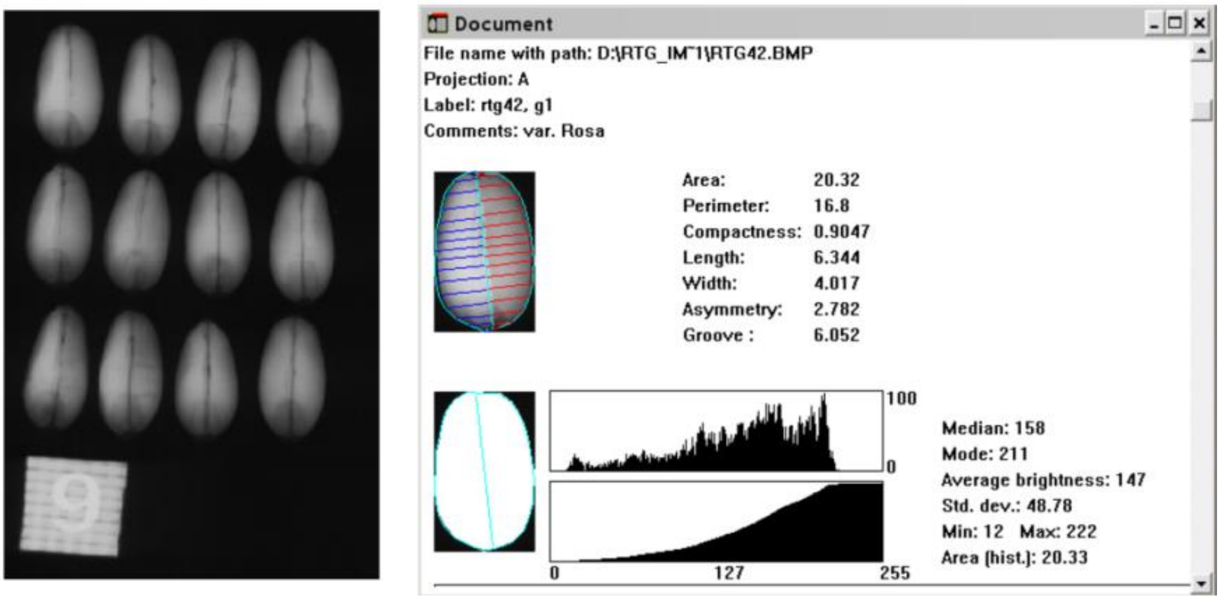
The Classification Problem

The Wheat Seeds Dataset involves the prediction of species given measurements of seeds from different varieties of wheat.

It is a multi-class (3-class) classification problem. The number of observations for each class is balanced. There are 210 observations with 7 input variables and 1 output variable. The variable names are as follows:

The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian (Class 1, 2, 3 respectively), 70 elements each, randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin.[1]

To construct the data, seven geometric parameters of wheat kernels were measured using X-ray imagery as shown in Fig. 1.



File name with path: D:\RTG_IM~1\RTG42.BMP
Projection: A
Label: rtg42, g1
Comments: var. Rosa

Area:	20.32
Perimeter:	16.8
Compactness:	0.9047
Length:	6.344
Width:	4.017
Asymmetry:	2.782
Groove :	6.052

Median: 158
Mode: 211
Average brightness: 147
Std. dev.: 48.78
Min: 12 Max: 222
Area (hist.): 20.33

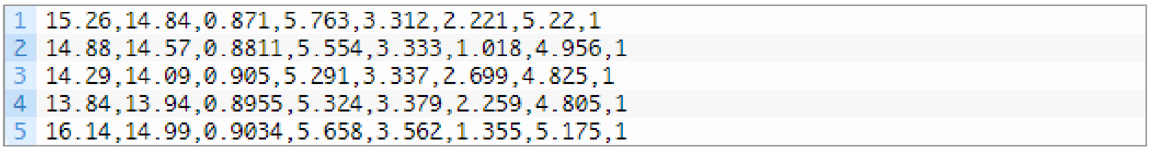
Figure 1: X-ray photogram and geometric and statistical parameters of the kernel image[1]

[1] M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. Zak, 'A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images', in: Information Technologies in Biomedicine, Ewa Pietka, Jacek Kawa (eds.), Springer-Verlag, Berlin-Heidelberg, 2010, pp. 15-24.

The measured features are listed here:

1. Area A
2. Perimeter P,
3. Compactness C = 4*pi*A/P^2,
4. Length of kernel,
5. Width of kernel,
6. Asymmetry coefficient
7. Length of kernel groove.

All these parameters were real-valued and continuous. A sample of the first 5 rows is shown in Fig. 2. The last column of the dataset is the Class (1, 2, 3). For this dataset, the baseline performance (ZeroR) of predicting the most prevalent class is a classification accuracy of approximately 28%.



```
15.26,14.84,0.871,5.763,3.312,2.221,5.22,1
14.88,14.57,0.8811,5.554,3.333,1.018,4.956,1
14.29,14.09,0.905,5.291,3.337,2.699,4.825,1
13.84,13.94,0.8955,5.324,3.379,2.259,4.805,1
16.14,14.99,0.9034,5.658,3.562,1.355,5.175,1
```

Figure 2: The first 5 rows of the dataset

The Wheat Seed Dataset can be accessed here:

https://tinyurl.com/43fjukwu

Deliverables

1. The final KNIME workflow (*.knwf) – Upload this to Google Drive and paste the link in the report.
2. The report (*.pdf).