RHODES UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

EXAMINATION: NOVEMBER 2020

COMPUTER SCIENCE HONOURS
PAPER 2 – MACHINE LEARNING

Internal Examiner:	Mr. J Connan Dr. D Brown	**MARKS**: 120 **DURATION**: 3 hours
External Examiners:	Prof. I Sanders	

GENERAL INSTRUCTIONS TO CANDIDATES

1. This paper consists of 6 questions and 3 pages. ***Please ensure that you have a complete paper.***
2. State any assumptions and show all workings.
3. Diagrams are encouraged and should be labelled.
4. Provide answers that are concise, legible and clearly numbered.
5. Use the mark allocation as a guide to the depth of your answer.
6. The Concise Oxford English Dictionary may be used during this examination.
7. You may use a calculator (though it should not be needed).

PLEASE DO NOT TURN OVER THIS PAGE UNTIL TOLD TO DO SO.

Section A: Theory

Question 1 **(6 marks)**

Define Machine Learning and give a simple example of it that illustrates the terms you used to define it?

Question 2 **(4 marks)**

What can PCA be used for in the context of Machine Learning? Further, provide an example where it can improve classifier performance on images.

Question 3 **(6 + 6 + 3 = 15 marks)**

a. What K-Fold cross validation?

b. What are the advantages and disadvantages of having a large or small number of folds?

c. What are typical choices for K in K-Fold Cross Validation for very large, very small or typical datasets?

Question 4 **(7 + 5 + 6 = 18 marks)**

a. Provide a brief overview of Support Vector Machines, how they work and the type of problems that they are best suited to.

b. With reference to SVMs, explain what the parameter C does.

c. With reference to SVMs, explain how Gamma affects our model.

Section B: Practical

Read the entire question before answering each sub question.

Question 5 **(3 + 8 + 4 + 6 = 21 marks)**

a. Write a program `breast_cancer.py` that imports the `breast_cancer` dataset and split the training and test set to **50:50**.

b. Visualise the data by plotting the points, including a print statement appropriate for determining a scaling method for the next question.

c. Apply the min-max scaling method in a way that is appropriate for the data, including a one line comment for why you applied it that way.

d. Replace that scaling method with one that uses interquartile range.

Question 6 **(4 + 8 + 4 + 20 + [10 + 4 + 6] = 56 marks)**

Locate all relevant resources for this question in the `image_classification` directory. Complete this question inside the file `class_model.py`.

a. Correct runtime errors (if any) in `class_model.py`, and add code to split the training and test set to **20:80**.

b. Add a Python function that fits **SVM** models with linear and Gaussian kernels. The function should allow for parameter estimation of the best f1-score using grid search with cross-validation. Your output should look similar to this, but repeated for each kernel:

```
0.986 (+/-0.016) for {'C': 1, 'gamma': 0.001, 'kernel': RBF}
0.959 (+/-0.029) for {'C': 1, 'gamma': 0.0001, 'kernel': RBF}
0.988 (+/-0.017) for {'C': 10, 'gamma': 0.001, 'kernel': RBF}
```

c. Repeat sub question **c** (in the new Python function), but fitting an **MLP** model instead. The function should allow for parameter estimation of the best f1-score using grid search with cross-validation.

d. Add Python code to perform any preprocessing techniques that you may deem necessary to improve the performance of both the **SVM** and **MLP** models above. Parameters to be included in the grid search are up to you. Note that you should select the combination of preprocessing techniques and classifier parameters that yield the highest f1-score.

e. Using the best models obtained above:

   i. Compare the best performing **SVM** and **MLP** models by visualising the f1-score for each model using an appropriate plotting technique.

   ii. Add appropriate labels and legends to the output above.

   iii. At the bottom of the program (`class_model.py`), write a comment explaining why you think your best performing model (in combination with its parameters and preprocessing) performed the way it did for this **image classification problem**.

END OF EXAMINATION