UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY

DEPARTMENT OF COMPUTING AND TECHNOLOGY

ADVENT 2025 SEMESTER **EXAM**

PROGRAM: *BACHELOR OF SCIENCE IN COMPUTER SCIENCE (**BSCS**)*
YEAR: 3 SEMESTER: 1
COURSE CODE: *CSC3116* EXAMINATION TYPE: PROJECT-BASED
COURSE NAME: *MACHINE LEARNING*
EXAM DATE: DECEMBER 2025
TIME ALLOWED: *2 Weeks*

INSTRUCTIONS TO CANDIDATES

1. This paper consists of **four (4)** questions. Attempt **all** questions.
2. Prepare and submit one consolidated report (PDF or Word) that clearly presents your methodology, results, analysis, and conclusions for each question.
3. For each question, submit all supporting files (datasets, code, trained models, and output files) as specified at the end of each question.
4. Upload your complete submission (report + code + models + output files) to your Moodle account within **2 weeks** of receiving this exam. Late submissions will not be accepted.
5. This is an **individual and formal examination**. Collaboration, discussion, or sharing of code/results with others is strictly prohibited. Any form of plagiarism or collusion will result in loss of marks for all parties involved.
6. Use **k-fold cross-validation** for model evaluations to ensure robustness.
7. Visualize your results (for example, confusion matrices, error plots, learning curves, etc.) whenever possible.
8. Include clear explanations of design choices, hyperparameter tuning, encountered challenges, and solutions.
9. Bonus marks will be awarded for comprehensive and well-structured reports, high model performance, and clarity and technical coherence in writing.

QUESTION 1 (25%)

In this question, you are required to design and evaluate a random forest classifier for the task of fish disease classification. The dataset contains features extracted from fish images, including texture features such as entropy, contrast, energy, homogeneity, correlation, and dissimilarity, color features including average RGB pixel values, as well as statistical features such as mean, standard deviation, variance, kurtosis, and skewness. The target variable represents ten fish disease classes, encoded as integers from 0 to 9. Your task is to train a random forest classifier on this dataset and evaluate its performance using standard classification metrics. Additionally, you are required to compare the performance of the random forest model with that of other classifiers. You are also expected to analyze the effect of dimensionality reduction or feature selection on classifier performance, model interpretability, and computational efficiency.

a) Use the training dataset *fish_disease_train*.csv (last column = class label) to train a random forest classifier. Save the trained model file as *model_1.pkl*. **(5 marks)**
b) Load the trained model and test it on dataset *fish_disease_test.csv* to obtain model predictions. **(2 marks)**
c) Evaluate the trained model on precision, recall, F1-score, and accuracy. **(4 marks)**
d) Apply PCA or any filter feature selection method of your choice and compare the model's performance on test data, before and after dimensionality reduction. **(4 marks)**
e) Compare the classification results of the random forest on test data with decision tree and KNN models, in terms of accuracy, precision, recall, F1-score and receiver operating characteristic - area under the curve (ROC-AUC), with visualization. **(5 marks)**
f) Discuss the results in detail including class-wise performance and the implications of the observed performance trend. **(3 marks)**
g) Discuss any model hyperparameter settings, observations, and conclusion. **(2 marks)**
h) Submit the following files:
    1) mode_1.pkl (trained random forest model).

2) code_source1.ipynb or code_source1.py (source code).
3) Report section for question 1.

QUESTION 2 (25%)

In this question, you are required to design and evaluate linear regression models for a multi-output regression task using the *energy efficiency dataset*. Specifically, you are required to predict the *heating load* and *cooling load* of buildings based on eight building features, including relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution. You are required to evaluate the model performance separately for each target variable, using appropriate regression metrics, and compare the linear regression model with several other regression techniques.

a) Use the *energy efficiency dataset_train.csv* (last two columns = target values) to train two linear regression models to predict *heating load* and *cooling load*. Save the trained models as *model1_2.pkl* and *model2_2.pkl*. **(4 marks)**
b) Load and test the trained models on the test set, the *energy efficiency dataset_test.csv*. **(3 marks)**
c) Visualize two plots of the predicted data against one of the input features and include the regression line. **(3 marks)**
d) Evaluate the two models on test data using MSE, MAE, RMSE, MAPE, and R^2 score and report the metrics in a results table. **(5 marks)**
e) Discuss the results obtained in part (d), highlighting their implications for the target variables. Identify which features have the strongest impact on each target and describe any notable patterns or insights you observe. (*Hint:* you may use the Pearson correlation coefficient to determine which features most strongly influence each target). **(4 marks)**
f) Compare the performance of the linear regressors with the following models:
    1) Polynomial regression
    2) Ridge regression
    3) Lasso regression
    4) Support vector regression (SVR) (**4 marks)**
g) Discuss the model hyperparameter settings, observations, and conclusion. **(2 marks)**

h) Submit the following files:
    1) Model1_2.pkl and Model2_2.pkl (trained linear regressors).
    2) code_source2.ipynb or code_source2.py (source code).
    3) Report section for question 2.

QUESTION 3 (25%)

In this question, you are required to design and evaluate an XGBoost ensemble classifier for the MNIST digit recognition dataset.

a) Use the training dataset "mnist_train.csv" (first column = class label) to train an XGBoost Classifier. Save the model as "model_3.pkl". **(5 marks)**
b) Fine-tune the critical hyperparameters such as learning rate, max_depth, n_estimators, and justify your choices. **(5 marks)**
c) Test your trained classifier on "mnist_test.csv" and evaluate your model using precision, recall, F1-score, and accuracy. **(5 marks)**
d) Compare results with any baseline models such as logistic regression and decision tree, and ensemble models such as Adaptive Boosting (AdaBoost) and Gradient Boosting Machines (GBM). **(6 marks)**
e) Discuss any observations made and conclusion. **(4 marks)**
f) Submit the following files:
    1) model_3.pkl (trained XGBoost model).
    2) code_source3.ipynb or code_source3.py (source code).
    3) Report section for question 3.

QUESTION 4 (25%)

In this question, you are required to apply feature selection using mutual information to identify the most relevant predictors for ovarian cancer classification on the *ovarian cancer dataset.csv*. You will then use the selected features to train and evaluate a decision tree classifier, and analyze how feature selection affects the model's performance.

a) Apply mutual information to select the top 44 most relevant features, and save the resulting dataset as *selected_subset.csv*. **(4 marks)**
b) Train a decision tree classifier on *selected_subset.csv* dataset. Save the model as *model1_4.pkl*. **(4 marks)**

c) Evaluate the model's performance (precision, recall, F1-score, accuracy) on a held-out test split. **(4 marks)**

d) Train another decision tree classifier on the overall *ovarian_cancer_train.csv* dataset. Save the model as *model2_4.pkl*. **(4 marks)**

e) Test *model2_4.pkl* on *ovarian_cancer_test.csv* and compare its performance with that obtained for *model1_4.pkl* in (c) above. **(5 marks)**

f) Discuss your findings and interpret whether feature selection improved model generalization. **(4 marks)**

g) Submit the following files:

1) selected_subset.csv (file containing selected 42 features).
2) model1_4.pkl and model2_4.pkl (trained decision tree models).
3) code_source4.ipynb or code_source4.py – source code.
4) Report section for question 4.

~End of Examination~