University of Kent

2024/2025

Machine Learning and Forecasting (BUSN9108)

There are 9 questions with 100 marks in total.

Student name:

Login:

Student declaration: I confirm that: This is an original assessment and is entirely my own work. Where I have used ideas, tables, figures of other authors, I have acknowledged the source in every case. This assignment was not submitted previously as assessed work for any other academic course.

Plagiarism:	The electronic version will be analysed using the *Turnitin* plagiarism-detection software. You will have the opportunity to submit drafts of your report to *Turnitin* where you will be able to see the originality report (but you may have to wait until the next day to see this). There will thus be no excuse for including any unreferenced material from other sources. You will probably include tables and graphs; it is **essential** that you reference the source at the bottom of every such entry. Plagiarism is most common when students copy and paste descriptive material from some source (and remember that doing this and then changing the occasional word still constitutes plagiarism). Please also see https://www.kent.ac.uk/education/academic-integrity/guide-for-students/what-is-plagiarism Please note that a report that consists mainly of material you have not written yourself is not going to be acceptable, even if it is properly referenced. We are interested in what *you* can write about the information you have found. **Please ensure the similarity score of your report is no more than 25%.** We advise that you keep a copy of this assignment.
Submission methods	• Only submit your PDF-formatted report to the Turnitin system on Moodle; • Do not print your Word-formatted file to a PDF file. Instead, you should save your Word-formatted file to a PDF file. The Turnitin system does not accept a printed PDF file; • Do not copy and paste any part of this document onto your report; • Submit your work via moodle.kent.ac.uk by 20:00 pm, 12th February 2025 (Hong Kong time). • You should receive a confirmation email after your submission. Please ensure your submission is successful.
Total marks:	100%
Weighting:	100%
Length	• Your answer sheet should include up to 10 pages (excluding the cover page). The font size is 11. If it is over ten pages, the markers reserve the right not to mark the extra pages.
Attention	• Any uninterpreted figures, tables, or equations will not be marked. • Do not simply copy and paste SPSS (or Weka, or Python) interface/outputs onto your assignment without any interpretation. The focus of this assignment is not on how well you use the SPSS (or Weka or Python). • You can use any data analysis software package in answering the questions in this assignment.

Question 1.
Download dataset 2025_Box-Plot_Dataset.csv from module BUSN9108 on moodle.kent.ac.uk. Answer the following three questions.

(1). Use a boxplot to compare the *Miles_Per_Gallon* of the three groups Origin=1, 2, 3. Provide two of your findings.

[4 marks]

(2). Plot the Q-Q plot of *Acceleration* to check if it follows the normal distribution. What is your conclusion?

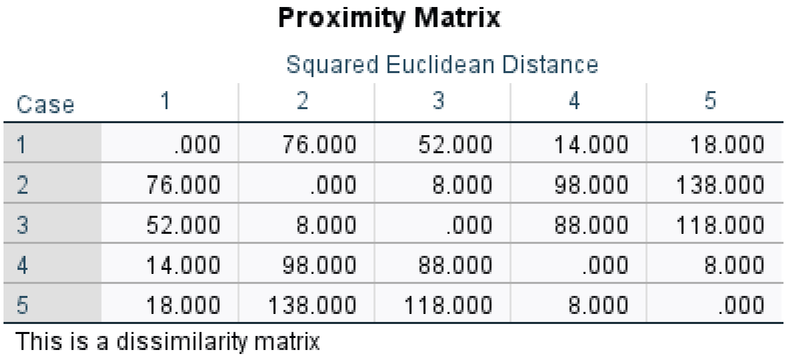
[2 marks]

(3). Use both the Kolmogorov-Smirnov(K-S) test and the Shapiro-Wilk (S-W) test to test whether the variable *Acceleration* is normally distributed. What is your conclusion and why?

[4 marks]

Question 2.
Answer the following questions.

(1). Assume you use the nearest neighbour method to cluster a dataset that contains five cases and obtain the following proximity matrix. If you have determined that the number of clusters is 2, then which two cases are in the same cluster?



Proximity Matrix

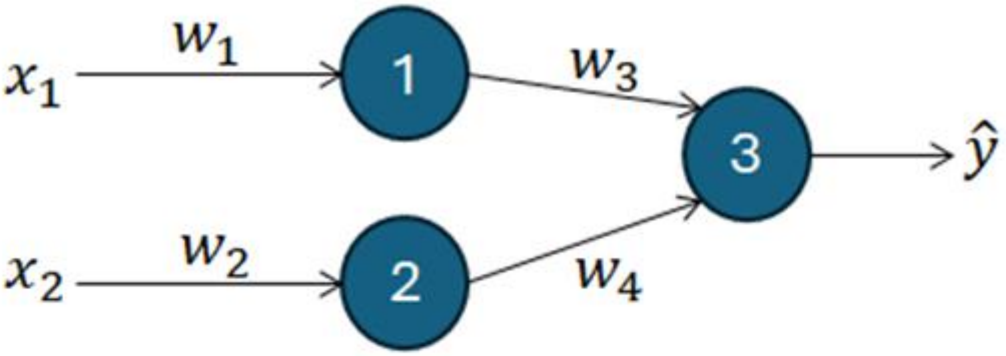
	Squared Euclidean Distance				
Case	1	2	3	4	5
1	.000	76.000	52.000	14.000	18.000
2	76.000	.000	8.000	98.000	138.000
3	52.000	8.000	.000	88.000	118.000
4	14.000	98.000	88.000	.000	8.000
5	18.000	138.000	118.000	8.000	.000

This is a dissimilarity matrix

[4 marks]

(2). The figure below shows a simple neural network. An observation with two variables (x_1, x_2) is presented to the network as shown, where $x_1 = 0.8, x_2 = 0.2, w_1 = 1, w_2 = 1, w_3 = 0.2, w_4 = -0.091$, and the bias in each neuron is 0. The activation functions of neurons 1, 2 and 3 are $f_1(x) = x, f_2(x) = x$, and $f_3(x) = (e^x - e^{-x})/(e^x + e^{-x})$, respectively. What is the predicted output $\hat{y}$ from the neural network? (Keep your answer to four decimal places.)

[4 marks]



Question 3.

Mr Simms was asked to analyse a dataset for a regression task. He used the kk-fold cross validation method, set k=3k=3, and then split the original dataset into three subsets: s_1, s_2, and s_3. Two regression modelling methods with the same number of independent variables were built. Their adjusted R2R2 values on the training datasets and the test datasets are given in the following table.

	Adjusted R^2 (%)	
	Modelling method 1	Modelling method 2
Training set($= s_1 + s_2$)	12.34	13.45
Test set ($= s_3$)	13.87	14.38
Training set ($= s_2 + s_3$)	14.62	13.85
Test set ($= s_1$)	15.93	14.43
Training set ($= s_1 + s_3$)	13.49	14.21
Test set ($= s_2$)	14.32	14.25

Which modelling method will you select? Discuss the reasons.

[6 Marks]

Question 4.

Download dataset *2025_Regression_Dataset.csv* from module BUSN9108 on moodle.kent.ac.uk. The dataset is **sampled from** a dataset in the UCI data bank, which implies that the dataset is a subset of the original dataset. You may find the description of the original dataset on https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data .

(1). Develop a **linear regression model** with variable critical_temp as the dependent variable and the other variables as independent variables (you may select some of them). Provide your modelling process and final model.

[18 marks]

(2). Discuss the performance of your model.

[3 marks]

Question 5.

Download dataset *2025_Classification_Dataset.csv* from module BUSN9108 on moodle.kent.ac.uk. The dataset is **sampled from** a dataset in the UCI data bank, which implies that the dataset is a subset of the original dataset. You may find the description of the original dataset on https://archive.ics.uci.edu/ml/datasets/seismic-bumps.

(1). Develop a **logistic regression** model with variable ***class*** as the dependent variable and the other variables as independent variables when the variable selection method is ***Backward: Ward***. Provide your modelling process.

[8 marks]

(2). Provide your final model and interpret the coefficient of an independent variable (i.e., a scale variable) in your model, respectively.

[4 marks]

(3). Based on the classification table of your final model, what can you suggest on improving the performance of your model in the future?

[2 marks]

Question 6.

Download dataset *2025_credit-card-clients.csv* from module BUSN9108 on moodle.kent.ac.uk. The dataset is **sampled from** a dataset in the UCI data repository, which implies that the dataset is a subset of the original dataset. You may find the description of the original dataset on https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. Use three modelling methods to build three classification models with variable ***Y*** as the dependent variable and the other variables as independent variables (you may select some of them), respectively. Please note: (A) **Logistic regression** should not be used in answering this question, and (B) if you want to use Decision trees, you can only use one of the three decision trees: QUEST, CRT and CHAID. Answer the following questions.

(1). Provide your modelling process for each method.

[6 marks]

(2). Select the modelling method with the best performance and explain how you will use it in the future.

[4 marks]

(3). Develop a random forest and then use SHAP (SHapley Additive exPlanations) to rank the importance of the independent variables in descending order.

[5 marks]

Question 7.

A construction equipment company, CX, started its business about 100 years ago. Recently, one of its product, *compact excavator*, has become the leading product of CX, but the sales team has suffered with repeated over productions and under productions, due to wrong sales forecast. The production team needs to schedule how many compact excavators need to be produced one month in advance. See dataset *2025_TimeSeries_Dataset.csv* from module BUSN9108 on moodle.kent.ac.uk for the sales data.

Ms Chan, the senior manager of the sales team, is very anxious with the accuracy of the current sales forecasting model, which uses the moving average of the last three months. The sales team was quite puzzled by the great variability in the sales every month, and the current forecast model seems to be too simplistic and could not justify why it averages the last three periods. Some of her sales team suggested that the historical data for the number of sales might contain seasonal dependencies, but they have no idea how to confirm the existence of the seasonal patterns and how to model this feature. Some others also suggested using a regression model, but they could not find any relevant external factors that could explain the behaviour of the sales time series, so they decided to enhance the forecast based on the historical time series data at this stage.

(1). Before any modelling, visually analyse each of the systematic patterns (e.g. trend and seasonality) in the time series and discuss their existence or/and patterns.

[3 marks]

(2). Divide the data into the training dataset/period (up to the end of 2015) for estimating forecast models, and the test (hold-out) dataset/period (Jan-2016 onward) to evaluate your model forecasts. Develop a Holt's exponential smoothing (HES) model and a Holt-Winter's model. When estimating the model parameter(s), you are suggested to use the MSE(mean squared error) in the fitting period only.

a) Present the two models, respectively. Note: you need to provide the mathematical models.

[5 marks]

b) Based on Part a), which model do you recommend finally and why?

[2 marks]

Question 8.

Download dataset 2025_Cluster-analysis_Dataset.csv from module BUSN9108 on moodle.kent.ac.uk. The dataset contains the following variables measuring the geometric parameters of a kind of plant:

v1. area A,
v2. perimeter P,
v3. length of kernel,
v4. width of kernel,
v5. asymmetry coefficient
v6. length of kernel groove

Answer the following questions.

(1). Use Ward's method to determine the number of clusters and explain the reason,

[3 marks]

(2). Based on the number of clusters determined from the above step, use the k-means clustering method to cluster the observations and interpret the outcome.

[3 marks]

(3). Find the value of the Dunn index of the k-mean clustering result from Part (2). Note: you will need to write Python code to answer it.

[4 marks]

Question 9.

Download dataset *2025_FactorAnalysis_Dataset.csv* from module BUSN9108 on moodle.kent.ac.uk. This dataset contains responses to a questionnaire on factors related to the quality of a public place. Each observation represents a response from a user. Answer the following questions.

(1). How many factors do you select and how do you select them?

[2 marks]

(2). What is the cumulative percentage of variance accounted for by your selected factors? Interpret it.

[2 marks]

(3). If you use rotation method “Varimax” and use extraction method “Principal component analysis”, which variables are your factor(s) associated with?

[2 marks]