University of Westminster
Department of Computer Science

7BUIS008W	Data Mining & Machine Learning
Module leader	Panagiotis Chountas
Unit	Coursework 2
Weighting:	50%
Qualifying mark	40%
Description	Students will be given a large data set (which may come from a range of different types of data sets) and they will be asked to fully implement data mining/machine learning techniques, focused on problem analysis, data pre- processing, data post-processing. Students must choose and implement appropriate algorithms; using the programming language (such as Python) and standard packages and toolkits (such as R). Students also need to perform critical evaluation of the performance of data mining and machine learning algorithms for a given domain/application. The students will be expected to produce a 2500- word project report on their analysis of the data set resulting from applying their own implementation of the algorithm(s).
Learning Outcomes Covered in this Assignment:	This assignment contributes towards the following Learning Outcomes (LOs): • LO2 fully implement data mining/machine learning projects, focused on problem analysis, data pre-processing, data post-processing by choosing and implementing appropriate algorithms; • LO4 fully implement encode and test data mining and machine learning algorithms using the programming language (such as Python) and standard packages and toolkits (such as R); • LO6 perform critical evaluation of performance metrics for data mining and machine learning algorithms for a given domain/application.
Handed Out:	29th November 2022
Due Date	10th January 2023 Submission by 13:00 hours
Expected deliverables	Submit on Blackboard a zip file containing the required documentation (either in docx or pdf format). All implemented codes should be included in your documentation together with the results/analysis.
Method of Submission:	Electronic submission on BB via a provided link close to the submission time.
Type of Feedback and Due Date:	Feedback will be provided on BB, on 25th January 2023
BCS CRITERIA MEETING IN THIS ASSIGNMENT	• **7.1.6 Use appropriate processes** • **7.1.7 Investigate and define a problem** • **7.1.8 Apply principles of supporting disciplines** • **8.1.1 Systematic understanding of knowledge of the domain with depth in particular areas** • **8.1.2 Comprehensive understanding of essential principles and practices** • **8.2.2 Tackling a significant technical problem**

	• **10.1.2 Comprehensive understanding of the scientific techniques**

Assessment regulations

Refer to section 4 of the "How you study" guide for undergraduate students for a clarification of how you are assessed, penalties and late submissions, what constitutes plagiarism etc.

Penalty for Late Submission

If you submit your coursework late but within 24 hours or one working day of the specified deadline, 10 marks will be deducted from the final mark, as a penalty for late submission, except for work which obtains a mark in the range 50 – 59%, in which case the mark will be capped at the pass mark (50%). If you submit your coursework more than 24 hours or more than one working day after the specified deadline you will be given a mark of zero for the work in question unless a claim of Mitigating Circumstances has been submitted and accepted as valid.

It is recognised that on occasion, illness or a personal crisis can mean that you fail to submit a piece of work on time. In such cases you must inform the Campus Office in writing on a mitigating circumstances form, giving the reason for your late or non-submission. You must provide relevant documentary evidence with the form. This information will be reported to the relevant Assessment Board that will decide whether the mark of zero shall stand. For more detailed information regarding University Assessment Regulations, please refer to the following website:**http://www.westminster.ac.uk/study/current-students/resources/academic-regulations**

Coursework Tasks

Predict the price of Brent oil Using Linear Regression

By observing the markets you learn everything about people, and most importantly this knowledge is provided in such a way where it is perfect for data scientists to put their hands on it. The main problem that tends to appear with markets, is that they are often unpredictable.
The mathematical models might all show that the value of a certain commodity will go up and then something unpredictable happens (e.x. COVID-19) and everything changes.Thus, it becomes obvious that the markets are extremely prone to external influence and factors. Nevertheless, your task is to attempt, by using linear regression, to predict the price of Brent oil.
We now have all Brent oil price information for the last 20 years under "BrentOilPrices.csv" on BB. It is now time to see how our data looks in order to determine what *features* we want to keep and which ones we want to discard. It appears that we are quite lucky! The dataset only contains the date and the price, thus there is no sort of tampering needed on our part.

Task A

a) **Data Visualisation: Define simple line chart to give an idea of the stock price change Brent oil price information for the last 20 years**

b) **Build *explanatory variables* — the *features* we are going to use to predict the price of oil. The variables we will be using at this stage, are the moving averages for the past three (MA3) and nine days (MA9), based on input from the oil stock market.**

A moving average is a technical indicator that market analysts and investors may use to determine the direction of a trend. It sums up the data points of a financial security over a specific time period and divides the total by the number of data points to arrive at an average. It is called a "moving" average because it is continually recalculated based on the latest price data.
Analysts use the moving average to examine support and resistance by evaluating the movements of an asset's price. A moving average reflects the previous price action/movement of a security. Analysts or investors then use the information to determine the potential direction of the asset price. It is known as a lagging indicator because it trails the price action of the underlying asset to produce a signal or show the direction of a given trend.

c) **Define the Train and Test Data:** This step covers the preparation of the train data and the test data.

d) **Build a Linear Regression Model** (LR) using **the moving averages for the past three (MA3) and nine days (MA9) as inputs;**

e) **Prediction Function and Result:** In this step, run the model using the test data we defined in step four. Visualise the predicted versus the actual stock values for the specific time period and calculate the model's accuracy

f) **Calculate the alpha and betas value:** Define the linear regression equation using the alpha and betas values

[30 Marks]

Predicting the Brent oil price Stock with LSTM Neural Networks

Build a Python program that can predict the price t**he price of Brent oil using the data set given for Task A**.

Task B

a) **Define the Train and Test Data:** This step covers the preparation of the train data and the test data. Explain the techniques used to generate the train data and the test data for the given Brent **oil price time series data set**.

b) **Build the Model:** Define the Long Short-Term Memory model (LSTM) and clearly explain the input features as a function of time lag.

c) **Prediction Function and Result:** In this step, we are running the model using the test data we defined in step four. Visualise the predicted versus the actual stock values for the specific time period

[20 Marks]

Guidelines:

You are required to deliver a report (max 20 pages including all figures) describing the methods adopted and the discussion of achieved results with reference to the tasks listed below. Assume that the report is targeted to a *marketing strategist*, who is interested to learn the business insights inferred in your analysis and to receive suggestions on how to take appropriate actions therefore.

Marking Scheme

Due to the nature of the assessment candidates may come up with more than one equally, good solutions. Thus marks will be allocated as follows

Predict the price of Brent oil Using Linear Regression

Task A

a) **Data Visualisation: Define simple line chart to give an idea of the stock price change Brent oil price information for the last 20 years;**

[4 Marks]

b) **Build *explanatory variables* — the *features* we are going to use to predict the price of oil. The variables we will be using at this stage, are the moving averages for the past three (MA3) and nine days (MA9), based on input from the oil stock market;**

[7 Marks]

A moving average is a technical indicator that market analysts and investors may use to determine the direction of a trend. It sums up the data points of a financial security over a specific time period and divides the total by the number of data points to arrive at an average. It is called a "moving" average because it is continually recalculated based on the latest price data.
Analysts use the moving average to examine support and resistance by evaluating the movements of an asset's price. A moving average reflects the previous price action/movement of a security. Analysts or investors then use the information to determine the potential direction of the asset price. It is known as a lagging indicator because it trails the price action of the underlying asset to produce a signal or show the direction of a given trend.

c) **Define the Train and Test Data:** This step covers the preparation of the train data and the test data;

[5 Marks]

d) **Build a Linear Regression Model** (LR) using **the moving averages for the past three (MA3) and nine days (MA9) as inputs;**

[5 Marks]

e) **Prediction Function and Result:** In this step, run the model using the test data we defined in step four. Visualise the predicted versus the actual stock values for the specific time period and calculate the model's accuracy;

[5 Marks]

f) **Calculate the alpha and betas value:** Define the linear regression equation using the alpha and betas values;

[4 Marks]

[30 Marks]

Predicting the Brent oil price Stock with LSTM Neural Networks

Task B

a. **Define the Train and Test Data:** This step covers the preparation of the train data and the test data. Explain the techniques used to generate the train data and the test data for the given **Brent oil price time series data set**.

[7 Marks]

b. **Build the Model:** Define the Long Short-Term Memory model (LSTM) and clearly explain the input features as a function of time lag.

[8 Marks]

c. **Prediction Function and Result:** In this step, we are running the model using the test data we defined in step four. Visualise the predicted versus the actual oil price values for the reported time period

[5 Marks]

[20 Marks]