Project

Site: Eduvos LMS
Course: Machine Learning Algorithms Assessments
Book: Project

Printed by: Musongela Nyembo
Date: Tuesday, 4 November 2025, 8:15 AM

Table of contents

1. Project
2. Instructions to Students
3. Section A
    3.1. Question 1
    3.2. Question 2
    3.3. Question 3
    3.4. Question 4
    3.5. Appendix A

Project

Faculty:	Information Technology
Module Code:	ITMLA2-44
Module Name:	Machine Learning Algorithms
Content Writer:	Taryn Michael
Internal Moderation:	Community of Practice
Copy Editor:	Kyle Keens
Total Marks:	100
Submission Week:	Week 6

This module is presented on NQF level 6.

5% will be deducted from the student's assignment mark for each calendar day the assignment is submitted late, up to a maximum of three calendar days. The penalty will be based on the official campus submission date.

Projects submitted later than three calendar days after the deadline or not submitted will get 0%. [1]

This is an individual project.

This project contributes 40% towards the final mark.

[1] Under no circumstances will assignments be accepted for marking after the assignments of other students have been marked and returned to the students.

Instructions to Students

1. Please ensure that your answer file (where applicable) is named as follows before submission: **Module Code – Assessment Type – Campus Name – Student Number.**
2. Remember to keep a copy of all submitted assignments.
3. All work must be typed.
4. Please note that you will be evaluated on your writing skills in all your assignments.
5. All work must be submitted through Turnitin. The full originality report will be automatically generated and available for the lecturer to assess. Negative marking will be applied if you are found guilty of plagiarism, poor writing skills, or if you have applied incorrect or insufficient referencing. (See the "instructions to students" book activity before this activity where the application of negative marking is explained.)
6. You are not allowed to offer your work for sale or to purchase the work of other students. This includes the use of professional assignment writers and websites, such as Essay Box. You are also not allowed to make use of artificial intelligence tools, such

as ChatGPT, to create content and submit it as your own work. If this should happen, Eduvos reserves the right not to accept future submissions from you.

Section A

Question 1 — 25 Marks

Study the scenario and complete the questions that follow:

Diagnosing Patients with Dementia

You are a data scientist working for a healthcare organisation that aims to improve early detection of Dementia. The organisation has provided you with a rich dataset containing health information for 2,149 patients. Each patient is uniquely identified by a Patient ID, and the dataset contains a variety of features such as demographic details, lifestyle factors, medical history, clinical measurements, cognitive assessments, and symptoms.

Your task is to build a machine learning model to predict whether a patient has Dementia based on the available data. This classification model will help doctors identify high-risk patients and prioritise them for further diagnostic tests or interventions.

The dataset is named "dementia.csv" and can be downloaded from the “Project Datasets” folder on myLMS.

Appendix A has a detailed description of all the columns within the dataset.

1.1. Performing the necessary preprocessing on the data to get it ready for training the machine learning model. This will include removing missing values, encoding categorical variables, feature standardisation and any necessary feature engineering.

(5 Marks)

1.2. Implement both of the following approaches for feature selection:

a) A filter-based method using Select KBest

b) A wrapper-based method using Recursive Feature Elimination (RFE) with a suitable estimator

Your implementation should clearly show the selected features in each case.

(10 marks)

1.3. For both RFE and SelectKBest:

- Discuss the advantages and limitations of the methods

(5 marks)

- Explain the process for selecting the optimal number of features for each approach

(5 marks)

End of Question 1

Question 2 35 Marks

2.1. Using the features selected from Question 1 (from RFE and SelectKBest), implement the following two machine learning algorithms for predicting dementia using the results from both feature selection methods for comparison:

a) Logistic Regression

(10 marks)

b) Decision tree

(10 marks)

2.2. Evaluate and compare the performance of each algorithm you implemented in 2.1. above. Justify your choice of evaluation metrics used and provide a summary of which algorithm performed better for both feature selection methods, and why you think that is the case.

(15 marks)

End of Question 2

Question 3 — 30 Marks

Study the scenario and complete the questions that follow:

Uber Eats Delivery Times

You are a data scientist working at Uber Eats, one of the world's leading food delivery platforms. The company is focused on improving the accuracy of estimated delivery times to enhance customer satisfaction and operational efficiency. Predicting delivery time more accurately will not only improve the user experience but also help optimise routing and courier management.

Your task is to develop a machine learning model that predicts the total delivery time (in minutes) for a food order, based on historical data. This model must use gradient descent to learn the best model parameters for predicting delivery times from the available features.

Dataset Description

You have access to the following data for previous orders:

- **Order_ID**: A unique identifier for each order.
- **Distance_km**: The delivery distance in kilometres.
- **Weather**: Weather conditions during the delivery.
- **Traffic_Level**: Traffic conditions category.
- **Time_of_Day**: The time of day when the delivery took place.
- **Vehicle_Type**: Type of vehicle used for delivery.
- **Preparation_Time_min**: The time required to prepare the order, measured in minutes.
- **Courier_Experience_yrs**: Experience of the courier/driver in years.
- **Delivery_Time_min**: The total delivery time in minutes (target variable - continuous).

The dataset is named "uber_eats.csv" and can be downloaded from the "Project Datasets" folder on myLMS

3.1. Explain why Linear Regression with Gradient Descent is suitable for predicting Uber Eats delivery time.

(3 marks)

3.2. Implement the Linear Regression algorithm using Gradient Descent on the Uber Eats dataset to predict delivery time.

Your implementation should include the following steps:

- Data cleaning (e.g., handling missing values)

(3 marks)

- Encoding categorical variables

(4 marks)

- Feature selection

(5 marks)

- Feature standardisation

(3 marks)

- Splitting data

(2 marks)

- Training model using Gradient Descent

(5 marks)

- Evaluating model performance using Mean Squared Error, mean absolute error, root mean squared error, and R^2

(5 marks)

End of Question 3

Question 4

10 Marks

Study the scenario and complete the questions that follow:

Clustering Analysis

An Italian wine consortium representing vineyards in a specific region has commissioned a study to improve its understanding of the chemical properties of its wines. The region produces three distinct wine varieties (cultivars), and the consortium wants to ensure that each cultivar maintains its unique quality and consistency while exploring new ways to market its products effectively.

You are hired as a data analyst to help the consortium achieve these goals. Your task is to use K-Means clustering to group the wines based on their chemical properties and identify natural clusters.

The dataset is named "wine_clustering.csv" and can be downloaded from the “Project Datasets” folder on myLMS

Implement K-means clustering on the data in the scenario above. Ensure that you standardise your data. You may use either the elbow or silhouette method to determine the optimal number of clusters. After implementing the clustering, create a plot to visualise the clusters you have created.

(10 marks)

End of Question 4

Appendix A

Dataset Features

Below are descriptions of the features of the dataset you will be using for question 1 and 2:

Demographic Details:

- **Age**: Age of the patients (60–90 years).
- **Gender**: Male or Female.
- **Ethnicity**: Categories include Caucasian, African American, Asian, and Other.
- **EducationLevel**: None, High School, Bachelor's, Higher.

Lifestyle Factors:

- **BMI**: Body Mass Index (15–40).
- **Smoking**: Smoking status (No, Yes).
- **AlcoholConsumption**: Weekly alcohol consumption (0–20 units).
- **PhysicalActivity**: Weekly physical activity (0–10 hours).

- **DietQuality**: Diet quality score (0–10).
- **SleepQuality**: Sleep quality score (4–10).

Medical History:

- **FamilyHistoryDementia**: Family history of Dementia (0: No, 1: Yes).
- **CardiovascularDisease**, **Diabetes**, **Depression**, **HeadInjury**, **Hypertension**: (0: No, 1: Yes).

Clinical Measurements:

- **SystolicBP** and **DiastolicBP**: Blood pressure measurements.
- **CholesterolTotal**, **CholesterolLDL**, **CholesterolHDL**, **CholesterolTriglycerides**: Cholesterol levels.

Cognitive and Functional Assessments:

- **MMSE**: Mini-Mental State Examination score (0–30).
- **FunctionalAssessment**: Functional assessment score (0–10).
- **MemoryComplaints**, **BehavioralProblems**: (0: No, 1: Yes).
- **ADL**: Activities of Daily Living score (0–10).

Symptoms:

- **Confusion**, **Disorientation**, **PersonalityChanges**, **DifficultyCompletingTasks**, **Forgetfulness**: (0: No, 1: Yes).

Target Variable:

- **Diagnosis**: Diagnosis of Dementia (0: No, 1: Yes).