Project

Site: Eduvos Learning Management System
Course: ITDAA4-12
Book: Project

Printed by: Eddie Theron
Date: Saturday, 22 April 2023, 10:13 AM

Table of contents

1. Project

Faculty:	Information Technology
Module Code:	ITDAA4-12
Module Name:	Data Mining and Data Administration
Module Coordinator:	N/A
Internal Moderation:	Community of Practice
Copy Editor:	Kyle Keens
Total Marks:	100
Submission Week:	Block 2 Week 5

This module is presented on NQF level 8

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.

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

This is an individual assignment.

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.

2. Section A

Section A

Learning Objective

[Enter text]

Choose an item: Assignment / Project Topic

[Enter text]

Scope

[Enter text]

Technical Aspects

[Enter text]

Marking Criteria

[Enter text]

2.1. Question 1

Question 1 — 30 Marks

Study the scenario and complete the questions that follow:

Classification: Diabetes Expert System

Sediba MedResearch is a research company that focuses on the study of diseases aiming to provide accurate expert systems that can help medical practitioners treat patients within hospitals. Among the several diseases the research firm works on; you have been assigned as a data scientist to the niche group working on Diabetes detection and classification. Currently, the team is working on a dataset ("**smr_diabetes.csv**") to predict, given a patient profile, the likelihood that the patient has Pima Indian diabetes. The team deem relevant to test several models and identify the model with the highest detection accuracy.

1.1. Divide the datasets into 80/20 per cent of training and test sets and use 10-fold cross-validation during model training to minimize bias. Train the datasets using the classification algorithms learned: Logistic Regression, Naïve Bayes, Decision Trees and Neural Networks using the best hyperparameter configuration possible.

(12 marks)

1.2. Using a properly labelled bar graph, plot each algorithm's average classification accuracy (training and test sets) following the 10-fold cross-validation training process and advise on the most accurate model among the four. Also, provide each algorithm's aggregated confusion matrices (training and test sets) after 10-fold cross-validation.

(18 marks)

End of Question 1

2.2. Question 2

Question 2 **45 Marks**

Study the scenario and complete the questions that follow:

Regression Analysis: Word Life Expectancy

The World Health Organisation has surveyed demographic variables affecting people's life expectancy across all countries in the world from a period of 2000 to 2015. The datasets ("**Life_Expectancy_Data.csv**") are described as follows:

Country	Country name
Year	Year of the data
Status	Country status of developed or developing
Life Expectancy	In age
Adult Mortality	Probability of dying between 15 and 60 years per 1000 population
Infant.deaths	Number of Infant Deaths per 1000 population
Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
percentage.expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita(%)
Hepatitis.B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	number of reported cases per 1000 population
BMI	Average Body Mass Index of entire population
under.five.deaths	Number of under-five deaths per 1000 population
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total.expenditure	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) room
HIV.AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	Gross Domestic Product per capita (in USD)
Population	Population of the country
thinness..1.19.years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
Income.composition.of.resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Number of years of Schooling(years)

2.1. Investigate in the literature wrapper-based feature selection methods, and provide a summary report of these methods, including graphical illustrations.

(8 marks)

2.2. Divide your data into 80/20 training and test sets. Using a **wrapper feature selection method of your choice** - build a linear regression model

in a 10-fold cross-validation scheme to predict life expectancy based on the relevant predicting variables. Provide an elaborate description of the wrapper method used.

(15 marks)

2.3. Compute the coefficient of determination plots and R^2 scores for the best model on both the training and test set

(8 marks)

2.4. List the most relevant variables obtained and comment on your observations.

(5 marks)

2.5. Perform 2D clustering analysis using k-Means clustering and the silhouette method to find existing clusters based on their GDP and Life expectancy attributes. Present your data graphically and elaborate on your findings.

(9 marks)

End of Question 2

2.3. Question 3

Question 3 — 25 Marks

Study the scenario and complete the questions that follow:

Market Basket Analysis: LL-Stores

LL Stores is a retail company in South Africa that specializes in food products. Aware of the competitive value of data analytics in discovering insight from several aspects of the business, the company would like to invest in including these technologies to inform its processes.

The company wants to gain insights into customer purchase behaviour to advise product placement within their stores and pricing strategies. You have been supplied with data ("**menlyn_customers_transactions.csv**") for their Menlyn branch, where they would like to pilot the project. Your task is thus to perform analytics on the data and provide a matrix of which products customers are likely to buy together. This insight will inform the branch on how to stack its products together and which promotion and pricing strategies to deploy. The company will evaluate the impact of this novelty after a couple of months of deployment.

Source: Matanga, NY (2023)

3.1 Transform your transaction datasets into a data frame with columns, all the existing products in the dataset and rows of boolean values (True/False) that describe whether the product is purchased or not in each given transaction.as described in the figure below.

(5 marks)

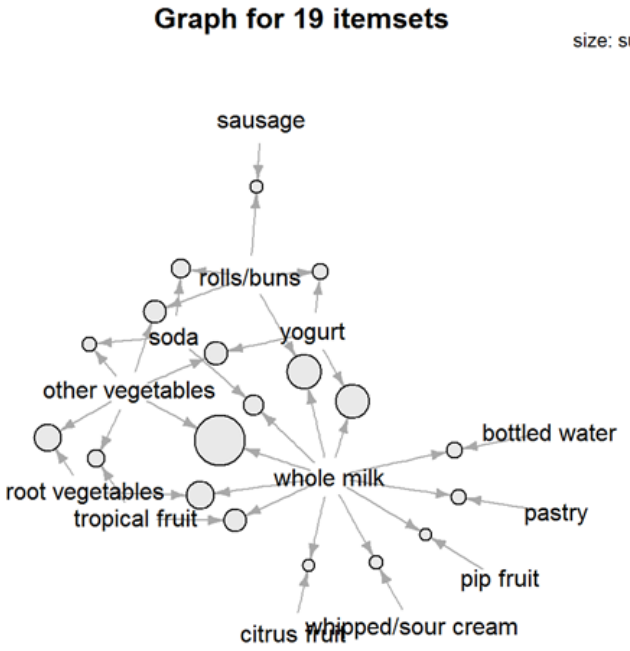
transactionID	Item1	item2	...	itemK
1	1	0	...	1
2	0	1	..	0
3	0	0	...	1
....	..	..	...	0
N	1	0	1	1

3.2. Using the Apriori algorithm, build association rules of that infers the likelihood of purchasing one item if another is purchased and present your results in a detailed graphical table that include the lift, confidence and support of each item set. Sort your table in descending order of the lift of your item sets. You may use a reasonable minimum support of your choice for the pruning process.

(10 marks)

3.3. Using a Network Graph similar to the example given below (e.g. arulezviz package), display graphically, the association rules and interconnections between frequent items for the 20 rules with the highest lift. Comment briefly on your overall findings based on both the association patterns.

(10 marks)



Note: **Submit your Python script, including a Word file containing all the narratives, tables and graphs for each question and sub-question.**

End of Question 3