Date: Friday, 24 May 2024, 10:14 PM

ITBDA4-14 Assessments (2024)

Project 1

1. Project 1

This project is an individual project assessing the following:

Exploratory Data Analysis, Hypothesis Testing, Regression analysis, Classification and Clustering, R programming

All datasets are available within the questions

Faculty Name:	Information Technology
Module Name:	Big Data Analytics
Module Code:	ITBDA4-14
Date:	Block 2 (Week 6)
Total Marks:	100
Content Writer:	Dr Yves Matanga

Section	Marks Subtotal
Section A: Application-based	30 Marks
Section B: Scenario-based	70 Marks

2. Question 1

Study the scenario and complete the question(s) that follow:

School Performance – Northwest High School

In recent times, there has been a notable surge in the demand for STEM (Science, Technology, Engineering, and Mathematics) skills. Governments worldwide recognise the pivotal role played by STEM in driving technological innovation, supporting research and development, and facilitating ongoing, field-specific analyses. This increased emphasis on quantitative thinking has led to a concerted effort to bolster education in science, mathematics, engineering, and technology, aligning with national goals. Recognizing the significance of STEM skills, a high school situated in the province of Northwest has undertaken a proactive initiative. The school is committed to conducting a comprehensive analysis of the data pertaining to its last term high school students ('**northwest_highschool_grades.csv**') with the objective of evaluating the various factors that may influence student grades.

Source: MATANGA NY (2024)

1.1. Load the datasets into R, and provide the following graphs:

a. A density function of the overall final grade of students

(2 marks)

b. A box plot of students' final grade performance per home type (urban/rural)

(2 marks)

c. A box plot of students' final grade performance per parents' education. Average both parents' education level

(3 marks)

d. A box plot of students' final grade performance per study time

(3 marks)

e. A box plot of students' final grade performance per romantic status

(2 marks)

f. A box plot of students' final grade performance per quality of family relationships

(3 marks)

All graphs must be fully labelled, and the findings must be well described. Include your R scripts in the narrative.

[Sub Total 15 Marks]

3. Question 2

Based on the scenario in question 1, make use of the ANOVA and Tukey's HSD test, to determine whether the following attributes have an impact of the student's performance:

a. Student study time

(3 marks)

b. The student's home address type

(3 marks)

c. Student's romantic status

(3 marks)

d. Quality of family relationship.

(3 marks)

e. The parents' education level

(3 marks)

The statistical significance value of 5% is used to reject the null hypothesis.

All graphs must be fully labelled, and the findings must be well described. Include your R scripts in the narrative.

[Sub Total 15 Marks]

4. Question 3

Study the scenario and complete the question(s) that follow:

Second-Hand Car Pricing Model

A well-established car reseller company, with a decade of experience in the market, specializes in the sale of pre-owned vehicles acquired from previous owners. Over the years, the company has relied on a manual valuation process, accumulating a wealth of knowledge in its database records. In response to increasing demand and to streamline operations, the company has reached out to the IT team for the development of a machine learning model (**'car_pricing_datasets'**). The objective is to create a first-degree price estimation tool for potential sellers, automating the valuation process. This pricing engine will be integrated into the company's online platform via API, enabling a global reach and expediting negotiations with incoming sellers. As a data science consultant, you have been tasked with designing the machine learning model to enhance the efficiency of the pricing process. The ultimate goal is to provide a reliable and automated solution that aligns with the company's growth and service expansion.

Source: MATANGA, NY (2024)

3.1 Load the datasets into R, list all datasets and transform the CarName to extract the actual car name. Remove the car_ID and symbolling columns as part of the datasets

(4 marks)

3.2 Perform One Hot encoding to transform categorical variables into binary variables, normalise all feature variables and split the datasets into Training/Test with an 80/20 proportion.

(7 marks)

3.3 Build the following models

a. Multiple Linear Regression

(4 Marks)

b. Decision Tress Regression

(4 Marks)

3.4 Compute the coefficient of determination and root mean square error for the two models on both the training and test sets.

(8 Marks)

3.5. Draw the goodness of fit scatter plot on the test set for both models and comment on the model performances.

(8 marks)

3.6. Comment on the model performances, explain why one model performs better than the other

(5 marks)

[Sub Total 40 Marks]

5. Question 4

Study the scenario and complete the question(s) that follow:

Machine Learning and Spatial Analysis – Forest-type Mapping

A Japanese agency is embarking on a project to enhance its understanding of the nation's forest landscapes through the application of advanced technology. Leveraging remote sensing data, the agency aims to classify various forest types using a combination of satellite imagery and spectral analysis. The goal is to extract crucial map features that can significantly contribute to spatial analysis and, more specifically, the classification of different forest types prevalent in Japan ('forest_datasets').

In the current scenario, the agency has conducted a thorough analysis, resulting in the identification of 27 pertinent features derived from the spectral data. These features are now readily available for comprehensive data analytics. As a machine learning engineer, your pivotal role is to design robust models capable of accurately classifying the diverse types of forests found within the unique Japanese landscape

Source: MATANGA, NY (2024)

4.1 Load the datasets in R, extract the features and target variable and split the data in training/test sets with an 80/20 proportion.

(4 marks)

4.2 Build a logistic regression, a Naïve Bayes model, and a decision tree model to classify the types of forest in the Japanese landscape.

(9 marks)

4.3 Compute the confusion matrices and classification accuracy for the three models on both the training and test sets. Comment on the performance of the models.

(10 marks)

4.4 Using similarity information only (exclude the class variable) via optimal k-means clustering distinguish the different types of forest with the Japanese landscape. Based on your analysis of the training set only. Make use of the elbow method, and provide a within-cluster sum of squares line graph

(7 marks)

[Sub Total 30 Marks]