LEVEL: 5

MODULE TITLE: Big Data Analytics

MODULE CODE: MOD008909
Assessment Element: 010

Word Count: 6000 words (or equivalent)

Introduction

This assessment is based on the production of a data warehouse using data from the Office for National Statistics. You are to demonstrate knowledge of the creation process for a data warehouse that will link together data on Gross Disposable Household Income (GDHI) per head to Life Expectancy. The assignment is split into 2 elements which evaluate the module learning outcomes as follows:

1. Knowledge and Understanding. Understand and critically appraise the underlying theories, statistical and mathematical concepts associated with big data analytics.

2. Knowledge and Understanding Demonstrate knowledge of different types of tools for data collection, data cleaning, data integration and data visualisation.

3. Intellectual, practical, affective and transferrable skills. Analyse large datasets and extract statistics and features.

4. Intellectual, practical, affective and transferrable skills. Critically discuss the challenges with big data e.g., privacy, security, storage, and scaling.

5. Intellectual, practical, affective and transferrable skills. Train and validate a range of models for machine learning.

E010-1 Demonstration Task (30%)

Demonstration

Prepare a 5-10 minute demonstration discussing your choice of models, the training and validation process, and the results obtained. Include visual aids such as graphs or charts to support your findings.

Learning Outcome 5:

You are to apply machine learning models to predict life expectancy based on GDHI per head and other relevant features. The task involves the following:

Data Preparation: Prepare the dataset for machine learning, which may include feature selection, and splitting the data into training and test sets.

Model Selection: Choose two machine learning models suitable for the task. This could include regression models, decision trees, or neural networks, among others.

Training and Validation: Train the selected models on the training set and validate them using appropriate metrics such as RMSE, accuracy, or F1-score.

Model Comparison: Critically evaluate the performance of different models and discuss the trade-offs involved.

You **MUST** apply **machine learning models** to predict **life expectancy** based on **GDHI per head** and other relevant features. This task involves the following steps, with marks awarded according to the level of implementation:

For Marks Between 0-49% (MUST Criteria):

1. **Data Preparation:**
   - You **MUST** prepare the dataset for machine learning.
   - This **MUST** include **feature selection** and **splitting** the data into training and test sets.
2. **Model Selection:**
   - You **MUST** choose **two machine learning models** that are suitable for this task.
   - These models **MAY** include **regression models, decision trees, neural networks**, among others.
3. **Training and Validation:**
   - You **MUST** train the selected models on the training set.
   - You **MUST** validate the models using appropriate metrics such as **RMSE, accuracy, or F1-score**.
4. **Model Comparison:**
   - You **MUST** evaluate the performance of the different models based on standard metrics.
5. **Presentation:**
   - You **MUST** prepare a **5-10 minute demonstration** discussing the choice of models, the training and validation process, and the results obtained.
   - You **MUST** have a clear, basic presentation structure.

Marks in this range are awarded based on the correct and complete implementation of the **MUST** requirements. Failure to meet all **MUST** criteria would result in marks below 40%.

For Marks Between 50-69% (SHOULD Criteria):

In addition to fulfilling the **MUST** requirements:

1. **Model Selection and Justification:**
   - You **SHOULD** clearly explain the **rationale** behind the choice of machine learning models.
2. **Training and Validation Process:**
   - You **SHOULD** demonstrate the process of training and validation with evidence.
3. **Model Comparison:**
   - You **SHOULD** critically evaluate the performance of the models, discussing the **trade-offs** between them.
4. **Presentation:**
   - You **SHOULD** use **visual aids** such as graphs or charts to support your findings.
   - The structure **SHOULD** be well-defined and logically presented.

Marks in this range are awarded for meeting all **MUST** criteria and the correct implementation of **SHOULD** criteria. More detailed analysis, explanation, and a higher level of technical competence are expected.

For Marks of 70% or Above (MAY Criteria):

In addition to fulfilling the **MUST** and **SHOULD** requirements:

1. **Advanced Model Selection:**
   - You **MAY** apply **innovative or original approaches** to the selection of machine learning models.
2. **Training and Validation:**
   - You **MAY** demonstrate advanced techniques for training and validation, such as **cross-validation** or **hyperparameter tuning**.
3. **Model Comparison:**
   - You **MAY** provide an in-depth, critical comparison of models, considering alternative approaches and providing **unique insights**.
4. **Presentation:**
   - You **MAY** provide an **engaging and fluent** presentation, demonstrating a strong understanding of the task.
   - You **MAY** include **sophisticated visual aids** and other multimedia elements to enhance the clarity of your demonstration.

Marks of 70% and above are awarded for implementing **MAY** criteria, where students demonstrate creativity, deep analysis, and superior understanding of the models and techniques involved.

Marking Criteria

- **0-49%:** Awarded for implementing **MUST** criteria, including basic data preparation, model selection, training, validation, and presentation structure.
- **50-69%:** Awarded for implementing **SHOULD** criteria, which includes clear explanations, rationale, critical evaluations, and the use of visual aids.
- **70% and above:** Awarded for implementing **MAY** criteria, which demonstrate original thinking, in-depth analysis, advanced techniques, and sophisticated presentation.