

**Master: Data SCIENCE**

**Subject: Data Mining, Machine Learning, Deep Learning, NLP**

**Professor: Carmine Buono**

This assignment delves into the exciting world of Machine Learning (ML), Natural Language Processing (NLP) and Deep Learning. The assignment is divided into **4 levels**, each increasing in complexity:

* Level 1 (Multiple Choice): Test your understanding of basic concepts about ML, DL and NLP.
* Level 2 (Open Ended): Explore data preparation for ML, including cleaning techniques and handling missing data.
* Level 3 (Open Ended): Focus on data preparation for data and text analysis.
* Level 4 (Code Trace): Analyze a script performing cancer classification with an SVM model.

**Level 1**

**Machine Learning, NLP, and Sentiment Analysis (Multiple Choice)**

Instructions: Choose the best answer for each question.

1. Which of the following is NOT a common data cleaning step in NLP tasks?
   1. Removing punctuation marks
   2. Converting text to lowercase
   3. Adding punctuation marks for emphasis
   4. Removing stop words (common words like "the", "a")
2. What is the main difference between a Support Vector Machine (SVM) and K-means clustering?
   1. SVMs work with numerical data, while K-means works with text data.
   2. SVMs are always more accurate than K-means.
   3. SVMs are for classification, while K-means is for unsupervised learning
3. What does NLP stand for?
   1. Numerical Language Processing
   2. Natural Learning Process
   3. Natural Language Processing
   4. None of the above
4. What is the primary goal of sentiment analysis?
   1. To identify the topic of a text document.
   2. To understand the emotional tone of a text document
   3. To summarize a text document.
   4. To translate a text document to another language.
5. What is the primary purpose of backpropagation in a neural network?
   1. To initialize the weights of the network
   2. To calculate the accuracy of the model's predictions
   3. To efficiently adjust the weights of the network based on the error
   4. To store historical training data
6. What does a loss function in a neural network typically measure?
   1. The number of neurons in a layer
   2. The difference between the predicted and actual values
   3. The activation value of a specific neuron
   4. The memory usage of the network

**Level 2:**

**Machine Learning and Exploratory Data Analysis (Open Ended)**

These questions delve into Exploratory Data Analysis (EDA) techniques for machine learning tasks.

1. **Describe two benefits of performing Exploratory Data Analysis (EDA) before training a machine learning model.**
   1. Data Quality Assessment:

Identifies missing values, duplicates, and outliers.

Ensures data consistency and correctness.

* 1. Feature Relationships & Insights:

Helps detect correlations and distributions.

Guides feature selection and engineering.

1. **Imagine you're working with a dataset containing customer demographics and their purchase history. You suspect a correlation between age and the type of products purchased. Describe two EDA techniques you could use to investigate this relationship.**
2. Statistical Analysis:

Correlation Coefficients (Pearson, Spearman, Kendall) to measure the strength of the relationship.

Chi-square test for categorical purchase behavior.

1. Data Visualization:

Scatter Plots to identify trends.

Box Plots to analyze spending habits across age groups.

**Level 3**

**Machine Learning, NLP, and Sentiment Analysis (Open Ended)**

These questions focus on basic data preparation for data and text analysis.

1. **What are two reasons why data preparation is important before applying machine learning algorithms?**
   1. Ensures Consistency & Reduces Noise: Standardizes formats, corrects missing values, and removes duplicates.
   2. Improves Model Performance & Accuracy: Helps algorithms generalize better by removing biases and ensuring clean input.
2. **Describe two common data cleaning techniques used for text data in NLP.**
   1. Stop Word Removal: Eliminates frequent but unimportant words (e.g., "the," "is") to improve efficiency.
   2. Tokenization & Lemmatization/Stemming: Tokenization splits text into meaningful words or phrases. Lemmatization (e.g., "running" → "run") and stemming reduce words to their root form.

**Level 4**

**Trace for Cancer Classification Script**

**Exercise Description: Cancer Classification with Machine Learning**

You are called upon to solve a critical classification problem in the medical field. This assignment simulates a cancer classification system using Support Vector Machines (SVM). We'll be analyzing a dataset containing information about potential cancer patients.

The provided trace will guide you through the key steps involved in building and evaluating this system. You'll gain insights into how the script:

* Loads and explores the patient data.
* Prepares the data for the model.
* Trains an SVM model to classify cancer types.
* Evaluates the model's performance.

Predicts diagnoses for new patients.

**Data Description:**

* The dataset is in CSV format, named "The\_Cancer\_data\_1500\_V2.csv"
* It contains information about **1500 patients** and likely includes the following features:
  + **Age:** Patient's age (numerical value).
  + **Gender:** Patient's gender (categorical, male/female – 0/1).
  + **BMI:** Body Mass Index (numerical value).
  + **Smoking:** Smoking habits (categorical, yes/no – 0/1).
  + **Genetic Risk:** Genetic predisposition to cancer (categorical, possibly high/low/unknown, 0/1/2).
  + **Physical Activity:** Level of physical activity (numerical).
  + **Alcohol Intake:** Alcohol consumption habits (numerical).
  + **Cancer History:** Prior history of cancer (categorical, possibly yes/no – 0/1).
* The target variable is **Diagnosis**, which indicates the type of cancer diagnosed for each patient (categorical, yes/no - 0/1).

**Coding steps:**

1. Data Loading and Exploration

* Imports necessary libraries: pandas (data manipulation), scikit-learn (model training and evaluation), and matplotlib (plotting).
* Reads the CSV data file ("sample\_data/The\_Cancer\_data\_1500\_V2.csv") into a pandas DataFrame (df).

2. Target Variable and Feature Selection

* Extracts the target variable ("Diagnosis") representing the cancer diagnosis labels.
* Creates a new DataFrame (X) containing all features except "Diagnosis" by dropping that column.

3. Data Splitting for Training and Testing

* Splits the data into training and testing sets using train\_test\_split.
  + X\_train: Training data for model fitting.
  + X\_test: Testing data for evaluating the model's performance.
  + y\_train: Training labels corresponding to the training data.
  + y\_test: Testing labels corresponding to the testing data.
* test\_size=0.3 specifies that 30% of the data will be used for testing.
* Prints the lengths of the training and testing sets to confirm the split.

4. Model Training

* Creates a Support Vector Machine (SVM) model with a linear kernel (model = SVC(kernel='linear')).
* Trains the model on the training data (model.fit(X\_train, y\_train)). This involves learning the decision boundary between different cancer types based on the training features.

5. Model Evaluation and Prediction

* Makes predictions on the testing data (predictions = model.predict(X\_test)).
* Evaluates the model's accuracy using the score method (percentage = model.score(X\_test, y\_test)). This calculates the percentage of correctly predicted labels in the testing set.
* Prints the predicted labels (predictions) for the testing data.
* Calculates and prints the accuracy as a percentage (f"Accuracy = {percentage\*100} %").

6. Confusion Matrix

* Imports NumPy library (import numpy as np).
* Creates a confusion matrix (res = confusion\_matrix(y\_test, predictions)) to visualize the distribution of correct and incorrect predictions across different classes.

7. Prediction on New Data

* Defines a new data point (nr) representing a new patient's features.
* Prints the predicted diagnosis ("The Prediction about new patience is:", predictions") for the new data point.

**Grading Criteria**

* **Level 1**
  + 18 points out of 30
* **Level 2**
  + 4 points out of 30
* **Level 3**
  + 4 points out of 30
* **Level 4**
  + 4 points out of 30
* **Honors ("30 cum laude"): 1 extra point to a work which is excellent in all its components.**

**IMPORTANT:**

* **Be concise, NOT verbose; be schematic and conceptual, NOT discursive nor vague.**
* **When you draw up a technical document, a good practice is to use** Courier New font **for** **the code and the results**
* **About the level 4 you can use your preferred editor/notebook (Google Colab, Jupyter, VScode, ecc)**
* **If possible, please attach the entire notebook file and all the results of the script (.ipynb file)**
* **Use bullet points to list deliverables and objectives.**