Arab Academy for Science, Technology & Maritime Transport College of Artificial Intelligence

Report

Data Science Project

“**Coursera Course Dataset** “

Its time to learn best

Team Members :

* George Youhana
* Pavli Bahaa

Supervised by :

DR. Essam Seddik

Eng. Aya Mohamed

**Introduction**

**Our Reason &problem:**

Here, we have scrapped data from the Coursera website for the purpose of creating an intelligent course recommendation system for learners who are seeking to find the right course for themselves. This system aims to help learners find the most suitable course by answering a few questions.

**Data Brief:**

The course dataset was scrapped from the Coursera website and contains mainly 6 columns and 890 course data. The detailed description of the columns is as follows:

* course\_title: Contains the title of the course.
* course\_organization: The organization conducting the course.
* Certficate type: The different certifications available for the course.
* course\_rating: The ratings associated with each course.
* course\_difficulty: The level of difficulty of the course.
* course\_students\_enrolled: The number of students enrolled in the course.

**Data Cleaning**

The first step in our data analysis is to clean and preprocess the data. This involves:

Deleting the first Unnamed column

Renaming the column "course\_Certificate\_type" to "Certficate type"

Removing duplicate values

Handling missing data (Nan and None values)

Extracting numerical values in the column "course\_students\_enrolled" and modifying it for further use

Modifying the "course\_difficulty" column for further use

**Exploratory Data Analysis (EDA)**

After cleaning and preprocessing the data, we will now perform Exploratory Data Analysis (EDA) to understand the characteristics and patterns in the data. This includes:

* Calculating the sum, mean, median, min, and max of the "course\_rating" column
* Plotting a histogram of the "course\_rating" column to understand the distribution of the ratings
* Creating a scatter plot of "course\_rating" vs "overall\_rating" to understand the relationship between these two variables
* Creating a heatmap to understand the correlations between the different variables

**Feature Selection**

In this step, we will select the features that we will use to build the recommendation model. We will use the "course\_rating", "course\_students\_enrolled\_modified", and "course\_difficulty\_modified" columns as the features for our model.

**Model Building and Evaluation**

We will now build a linear regression model using the selected features and evaluate its performance. We will use the mean absolute error (MAE) metric to evaluate the model's performance.

**Recommendation**

Based on the results of our machine learning model, we can make the following recommendations for improving the course recommendation system:

1. Use more diverse features: While the current model uses three features (course\_rating, course\_students\_enrolled\_modified, and course\_difficulty\_modified), adding more diverse features such as the duration of the course, the language of the course, and the topic of the course may improve the model's performance.
2. Improve the quality of the data: Ensuring that the data used to train the model is accurate and up-to-date can significantly improve the model's performance. This includes ensuring that there are no missing or incorrect values in the data and that the data is representative of the overall population.
3. By make Coursera Recommendation system recommended that any courses provider that rating over 8.5 is very good to enroll it
4. Overall, by following these recommendations, we can improve the performance of the course recommendation system and provide more accurate recommendations to learners.