CourseInsights

***Abstract* — The development of CourseInsights demonstrates the transformative potential of data science in the educational domain by analyzing extensive datasets from leading online learning platforms. Utilizing Python for efficient web scraping and Power BI for interactive visualization, this project processed over 2,300 records from Coursera, FutureLearn, and OpenUniversity. Key insights were derived into course categories, pricing strategies, and the relationship between ratings and enrolments, providing stakeholders with actionable intelligence.**

The project highlights the importance of streamlined automation and user-centric analysis. Trends reveal that Technology, Business, and Data Science courses attract the highest enrolments, particularly free courses, which engage 30% more users than paid ones. Additionally, concise courses with durations of 5–10 hours show the greatest engagement, catering to professionals with limited time. An interactive dashboard facilitates exploration of these insights through dynamic filters, visual graphs, and predictive panels.

The study also addresses challenges encountered, including data variability, platform scraping limitations, and integration complexities, which required robust preprocessing techniques. Broader implications highlight the potential for CourseInsights to guide course design, pricing models, and educational policies, enabling stakeholders to align offerings with user preferences and trends.

This project underscores the value of data-driven decision-making in online education while acknowledging limitations, such as restricted platform inclusion and reliance on static scraping methods. Future enhancements include real-time data integration, advanced predictive modelling, and expanded platform coverage. CourseInsights sets the stage for innovative and scalable solutions in online education analytics.

I. Introduction

The rise of online learning platforms such as Coursera, FutureLearn, and OpenUniversity has revolutionized access to global education, offering a diverse range of courses spanning professional skills, academic knowledge, and personal development. These platforms cater to a wide audience, including working professionals, students, and lifelong learners. However, the rapid proliferation of courses presents challenges for stakeholders in understanding market trends, user preferences, and course performance.

The CourseInsights project addresses these challenges by constructing a robust data pipeline to extract, process, and analyze course information from multiple platforms. By leveraging Python tools for data extraction and Power BI for visualization, the project provides a comprehensive view of critical trends, including popular course categories, pricing structures, and relationships between course ratings and enrollments.

The primary objectives of this study are:

A. *To identify the most popular course categories and pricing strategies.*

B. *To investigate relationships between course features, including ratings, duration, and enrollments.*

C. *To develop predictive models for future trends in online education.*

D. *To present findings through an interactive and user-friendly Power BI dashboard.*

II. Dataset Overview

A. *Platforms and Records:*

1. *Coursera:*

- Total Records: 720.

- Focus: Professional and academic courses with detailed metadata, including ratings, reviews, and pricing information.

2. *FutureLearn:*

- Total Records: 213.

- Focus: Career-oriented programs, offering short courses and micro-credentials.

3. *OpenUniversity:*

- Total Records: 1,441.

- Focus: Free educational resources across diverse subjects, aimed at a broad audience.

B. *Features Extracted:*

The dataset includes several critical features that enable comprehensive analysis:

- Course Title: Identifies and categorizes courses.
- Ratings: Numerical scores reflecting user satisfaction.
- Reviews: Qualitative feedback from users.
- Price: Indicates whether a course is free or paid.
- Duration: Time commitment in hours, weeks and modules.
- Source: The institution or organization offering the course.
- Category: The subject area of the course, such as Technology or Business.
- Type: The level of the course as well as if it is a degree or a specialization.

C. *Data Type Summary:*

Extracted features vary in type, including:

1. *Text (titles, categories)*
2. *Numerical (ratings, duration)*
3. *Categorical (pricing)*

This diversity necessitated robust cleaning and preprocessing techniques to ensure data consistency.

D. *Dataset Size:*

The combined dataset includes 2,374 records. The structured data is saved in CSV format for analysis and visualization, ensuring compatibility with various tools.

III. Methodology

A. *Data Extraction:*

Python's BeautifulSoup library was utilized on Google Colab to extract data from static HTML pages, focusing on course titles, ratings, pricing, and categories. The scraping process followed these steps:

1. *Targeting Course Cards:*

- Data was initially extracted from structured "cards" on webpages, which contained summaries such as course title, institution name, type, and embedded links.

2. *Detail Extraction:*

- Links from the cards (normalized to ensure consistency) were followed to scrape additional details like ratings, number of reviews, and course modules.

3. *HTML Parsing:*

- The requests library sent HTTP requests to fetch the target URL, and BeautifulSoup parsed the HTML content. Relevant elements were identified using specific CSS class names.

4. *Dynamic URL Management:*

- Relative and absolute links were handled to ensure accurate navigation to each course page.

5. *Data Storage:*

- Extracted data was saved in structured formats (CSV and Excel) using pandas, ensuring compatibility with tools like Power BI.

6. *Error Handling:*

- Invalid pages and missing elements were managed with robust checks to avoid data loss.

Fields Extracted
Title
Source
Duration
Type
Category
Rating
Reviews
Price

Table 1. Data Extracted

B. *Data Cleaning:*

After extracting the raw data into Excel files, we used Python in Google Colab to clean and preprocess the data. The process involved the following steps:

1. *Loading the Data:*

- Imported the Excel file using pandas and inspected the dataset using commands like df.head(), df.info(), and df.describe() to identify null values, data types, and basic statistics.

2. *Data Type Conversion:*

- Converted columns like Title, Source, Type, and Category to object for categorical data, while Price, Rating, and Reviews were converted to numeric types for analysis.

Fields Extracted	Datatype
Title	*Object*
Source	*Object*
Duration	*Object*
Type	*Object*
Category	*Object*
Rating	*Float*
Reviews	*Float*
Price	*Float*

Table 2. Data types

3. *Handling Missing Values:*

- Replaced null values with appropriate substitutes (e.g., "Unknown" for Type and 0 for Reviews). Rows with over 30% missing data were dropped to ensure data quality.

4. *Formatting:*

- Standardized text columns (e.g., Title, Source, Type, and Category) to lowercase and appended "hours" to Duration for consistency.

5. *Duplicates:*

- Identified and removed duplicate rows to eliminate redundancy in the dataset.

6. *Validations:*

- Performed checks for missing values, duplicate rows, and appropriate data types after cleaning to ensure the dataset was ready for analysis.

Finally, the cleaned data was saved in Excel format for further use. Additionally, we performed data cleaning and formatting directly in Power BI for consistency, ensuring the raw data file was ready for visualization and analysis. This combined approach ensured a high-quality, uniform dataset for generating insights.

Field Name	Description
Rating in Stars	*Star-based rating for the course (e.g., 4.5/5).*
Website	*The platform hosting the course (e.g., Coursera, FutureLearn, OpenUniversity).*
Title	*The name of the course.*
Type	*The type of course (e.g., Degree, Specialization, Micro-credential).*
Duration	*Time commitment required to complete the course (e.g., hours or weeks).*
Rating	*Numerical rating of the course (out of 5).*
Reviews	*Number of user reviews provided for the course.*
Category	*Subject area of the course (e.g., Technology, Business).*
Price	*Cost of the course (e.g., Free, $50).*
Source	*The institution or organization offering the course.*
Country	*The country associated with the course or institution.*
Predicted Rating	*The forecasted rating based on historical and analytical trends.*

TABLE 3. FIELDS AFTER CLEANING

C. Data Analysis:

After cleaning the data, we performed an in-depth analysis to determine the most effective comparisons for accurate, efficient, and useful insights tailored to the project's objectives.

Key steps included:

1. *Descriptive Statistics: Summarized key metrics, such as average ratings and enrollments.*
2. *Correlation Analysis: Explored relationships between course features to identify patterns.*
3. *Comparative Analysis: Conducted targeted comparisons, including:*

   a) *Country by Category: To identify regional preferences in course topics.*
   b) *Title by Type: To explore how course formats (e.g., free vs. paid) vary across different course titles.*
   c) *Rating by Category: To understand how different categories perform in terms of user satisfaction.*
   d) *Rating by Reviews: To assess correlations between course ratings and the number of user reviews.*
   e) *Timeline by Type: To examine how course duration aligns with free or paid formats.*

4. *Predictive Modeling: Applied regression techniques to forecast future trends based on historical data. Models included linear regression and polynomial regression to address varying data patterns.*

These comparisons were also analyzed across universities to evaluate which institutions offered the best combination of price, rating, reviews, and duration, helping identify the most suitable options for potential learners.

Additionally, we implemented a regression algorithm to predict future course ratings. The predictive model was based on key fields, including country, reviews, price, and title, leveraging historical data to forecast trends. This predictive analysis added depth to the project, offering insights into how courses might perform in the future, enabling more informed decision-making for learners and course developers.

D. Visualization:

The Power BI dashboard we created showcases all the insights and data analyses from our project in an interactive and user-friendly format. It is designed to offer both detailed and comparative views of the information, enabling efficient decision-making. Key features of the dashboard include:

1. *Navigation and Search:*

- A navigation menu allows users to switch seamlessly between pages.
- Two search bars: One for filtering by website and another for searching specific categories or courses

2. *Filters for Detailed Exploration:*

- Dynamic filters are available for platform, pricing, and ratings, allowing users to refine the displayed data to meet their specific needs.

3. *Data Visualization:*

- Visual representations, including bar charts, heatmaps, and line graphs, highlight key insights across different dimensions.
- The dashboard can display detailed information for an individual course or category.
- Comparative analysis is supported, allowing users to compare data from up to three platforms simultaneously.

4. *Key Metrics and KPIs:*

- Star Rating Card: Highlights the percentage distribution of star ratings for each course.
- Map Integration: A visual map highlights the country of origin for each course, providing a geographic perspective.
- KPIs (Key Performance Indicators) provide a quick overview of high-performing courses, categories, and platforms, offering an at-a-glance understanding of trends.

5. *Predictive Analysis:*

- The final row of the dashboard includes predictive analytics, displaying future trends in course ratings using regression models.

6. *Custom Insights:*

- Visualizations such as charts and graphs analyze fields like title by category, timeline by modules, and rating by reviews.
- Comprehensive information on pricing, reviews, ratings, and more is readily available.

IV. Results And Discussion

A. *Key Insights:*

The analysis provided several valuable insights into user preferences and platform dynamics, which can inform decision-making for various stakeholders:

1. *Popular Categories:*

- Technology, Business, and Data Science emerged as the most popular and highly enrolled categories, reflecting current global trends in skills demand and workforce needs.
- Free courses in these categories significantly outperformed paid ones in terms of enrollments, highlighting the preference for accessible and low-cost learning options in high-demand fields.

2. *Pricing Trends:*

- Free courses attracted approximately 30% more users compared to paid courses, demonstrating that cost remains a major factor influencing user decisions.
- Despite the popularity of free courses, premium courses with strong ratings and reviews maintained steady enrollments, suggesting that users are willing to pay for high-quality content when it adds clear value.

3. *Duration Impact:*

- Courses lasting between 5 to 10 hours achieved the highest engagement rates, indicating a preference for concise yet comprehensive learning experiences.
- Shorter courses (<5 hours) were especially appealing to working professionals with limited time, making them an effective option for upskilling.

B. *Challenges:*

The project faced several challenges during data collection and processing, which were addressed with targeted solutions:

1. *Data Variability:*

- Discrepancies in data formatting across platforms required extensive cleaning and standardization to ensure consistency in analysis.
- For example, variations in how ratings, prices, and durations were recorded across platforms needed to be reconciled for accurate comparisons.

2. *Integration Complexity:*

- Combining datasets from multiple sources posed challenges in maintaining uniformity and avoiding duplication.
- Each platform used its own structure for presenting course information, necessitating tailored extraction and transformation processes.

3. *Scalability:*

- The scraping process needed optimization to handle the growing volume of data efficiently, particularly for platforms with extensive course offerings.
- Iterative improvements in the scraping and processing pipeline helped address these issues, making the workflow more robust and scalable for future use.

C. *Use Case for Stakeholders:*

The insights derived from this project offer practical applications for various stakeholders in the online education ecosystem:

1. *Educators:*

- Tailor course offerings to align with popular categories, engagement trends, and user preferences.

2. *Educational Institutions:*

- Evaluate the popularity of courses to guide syllabus design and develop effective marketing strategies.

3. *Administrators:*

- Optimize pricing strategies to balance free and paid offerings, attracting a broader range of users.

4. *Policy Makers and Government Agencies:*

- Use insights to develop policies that promote accessible education, particularly in high-demand fields like Technology and Business.

- Assess trends in online education to allocate funding or introduce supportive initiatives that improve access and quality.

5. *Students:*

- Identify high-rated free courses to maximize learning opportunities while staying within budget constraints.

D. *Broader Implications:*

The findings from this study underscore several key themes in the online education landscape:

1. *Accessibility:*

- The success of free courses highlights the importance of affordability in expanding access to education, particularly in developing regions or for financially constrained learners.

2. *Customization:*

- Offering diverse course options across categories and tailoring course durations to user preferences can significantly enhance learner engagement and satisfaction.

3. *Quality over Quantity:*

- While free courses attract a larger audience, the consistent performance of premium courses with strong ratings underscores the value of quality content. Institutions and platforms should focus on maintaining high standards in course design and delivery.

4. *Focus on High-Demand Skills:*

- Technology, Business, and Data Science remain critical areas of interest, reflecting evolving market demands. Institutions should prioritize developing programs in these categories to address workforce needs and global trends.

V. DASHBOARD FEATURES

The Power BI dashboard serves as an interactive and versatile tool tailored for educators, administrators, and policymakers. Its key features include:

A. *Dynamic Filters:*

- Easily refine data by platform, category, price, or other key metrics to customize the analysis based on user needs.

B. *Graphical Insights:*

- Provides clear and impactful visualizations, including line charts, bar graphs, and pie charts, to highlight trends in enrollments, ratings, pricing, and other critical metrics.

C. *Key Performance Indicators (KPIs):*

- Quick-access indicators display essential insights, such as top-rated courses, most-enrolled categories, and high-performing platforms, ensuring a snapshot view of key data.

D. *Interactive Drill-Downs:*

- Explore detailed metrics by diving deeper into specific courses, categories, or platforms, offering a granular view of the data for deeper analysis.

E. *Predictive Panels:*

- Utilizes regression analysis to display forecasts for enrollments, ratings, and other trends, enabling data-driven predictions and planning.

VI. LIMITATIONS AND FUTURE WORK

A. *Limitations:*

1. *Data Source Variability: Limited to three platforms, which may not represent the entire online education market.*
2. *Dynamic Content Challenges: Certain dynamic data was not fully accessible using static scraping techniques.*
3. *Predictive Model Constraints: Forecasts are limited by the quality and scope of historical data.*

B. *Future Work:*

1. *Real-Time Data Integration: Incorporate APIs to fetch live data for real-time insights.*
2. *Enhanced Predictive Models: Use machine learning algorithms for more accurate forecasting.*

3. *Expanding Platforms: Include data from additional platforms like edX and Udemy.*
4. *User Behavior Analysis: Study interaction data, such as click rates and completion rates, to provide deeper insights.*
5. *Mobile App Development: Create a mobile-friendly version of the dashboard to expand accessibility.*

C. Websites Excluded:

1. *Life Global: Excluded due to very limited course categories and insufficient data fields available for scraping.*
2. *Cursa: Excluded because the platform provided limited data fields, which restricted comprehensive analysis.*
3. *Khan Academy: Excluded as its courses are primarily designed for children, making it unsuitable for broader adult learning trends.*
4. *Udemy: Excluded because the website had a high security protocol, preventing data extraction.*
5. *Skillshare: Excluded because the platform focuses on creative and hobby-based courses, which did not align with the study's focus on academic and professional learning.*
6. *edX: Excluded due to restricted access and limited availability of metadata for scraping without advanced permissions.*

D. Data Excluded:

1. *Images: Images are not directly useful for identifying trends or generating insights in a data-driven analysis of courses.*
2. *Teacher/Instructor Names: Teacher names do not significantly contribute to understanding broader trends or patterns in online education.*

VII. CONCLUSION

CourseInsights exemplifies the transformative potential of data science in the education sector. By automating data extraction and creating actionable visualizations, this project provides valuable insights into online learning trends. The interactive dashboard ensures accessibility, enabling stakeholders to make informed decisions. Future enhancements will focus on integrating real-time APIs, expanding the dataset scope, and developing advanced predictive analytics to enrich the framework further.

VIII. APPENDIX

A. Detailed Dataset Breakdown:

- FutureLearn: 213 records focusing on career progression.

- Coursera: 720 records with comprehensive metadata.

- OpenUniversity: 1,441 records emphasizing free education.

B. Tools and Technologies Used:

1. *Python Libraries: BeautifulSoup, Pandas, Matplotlib, Scikit-learn.*
2. *Visualization: Power BI with dynamic filtering and KPI integration.*
3. *Regression Models: Linear and Polynomial regression for predictive analysis.*

C. User Guide for Dashboard Interaction:

1. *Select filters by platform or pricing to narrow down trends.*
2. *Hover over graphs for detailed insights on ratings and enrollments.*
3. *Use predictive panels to view forecasts for specific categories.*
4. *Export data views directly from Power BI for offline analysis.*

ACKNOWLEDGMENTS

We would like to express our heartfelt gratitude to our mentors, instructors, and peers who provided invaluable guidance and support throughout this project. Their encouragement and feedback were instrumental in shaping the direction of our work. We also extend our thanks to the creators and maintainers of the tools and platforms we utilized, such as Python, BeautifulSoup, and Power BI, which made this project possible. Finally, we acknowledge the efforts of the online learning platforms that served as the foundation of our analysis, enabling us to explore and derive meaningful insights.

REFERENCES

[1] BeautifulSoup Documentation. Retrieved from BeautifulSoup.

[2] McKinney, W. Python for Data Analysis. O'Reilly Media.

[3] Microsoft Power BI Documentation. Retrieved from Power BI.

[4] Han, J., Kamber, M., & Pei, J. Data Mining: Concepts and Techniques. Morgan Kaufmann.

[5] VanderPlas, J. Python Data Science Handbook. O'Reilly Media.

[6] OpenUniversity Course Data. Accessed through project scraping methods.

[7] Coursera Course Data. Accessed through project scraping methods.

[8] FutureLearn Course Data. Accessed through project scraping methods.