COURSERA'S DATA ANALYSIS COURSES: COMPREHENSIVE ANALYSIS

Prepared By: Marya Asaad - Shahad Alfahad - Alia Alrassan -Najlaa BenDakheel

DATA SOURCE

01. INTRODUCTION

Our project centers on the analysis of **Coursera's data analysis courses,** aiming to extract valuable insights that will **optimize user experience** and **course offerings** on the platform. By leveraging data science methodologies, we seek to **identify** the **top-rated courses,** understand **enrollment patterns** across different difficulty levels, explore **correlations between course attributes,** and recognize **prominent educators.** Through this data-driven approach, we aim to **enhance our understanding of user preferences** and learning needs, ultimately ensuring that Coursera users have access to the most relevant and impactful data analysis courses available.

02. OBJECTIVES

1. Determine **enrollment patterns** across beginner, intermediate, and advanced data analysis courses.
2. Figure out who the **most prominent educators** are for data analysis courses on Coursera.
3. Investigate **correlations between course level** and duration.
4. Find out which data analysis **courses people like** the most on Coursera.

03. DATASET

- **Course Name:** Identifies the subject briefly.
- **Course Educator:** Represents academic leadership guiding the learning experience.
- **Course Rating:** Reflects student satisfaction and course quality.
- **Course Level:** Indicates difficulty or complexity.
- **Course Duration:** Specifies length of each course.
- **Enrolled Students:** Shows popularity and demand, revealing learner preferences.
- **Course Language:** Specifies language of instruction, aiding learners seeking specific languages.

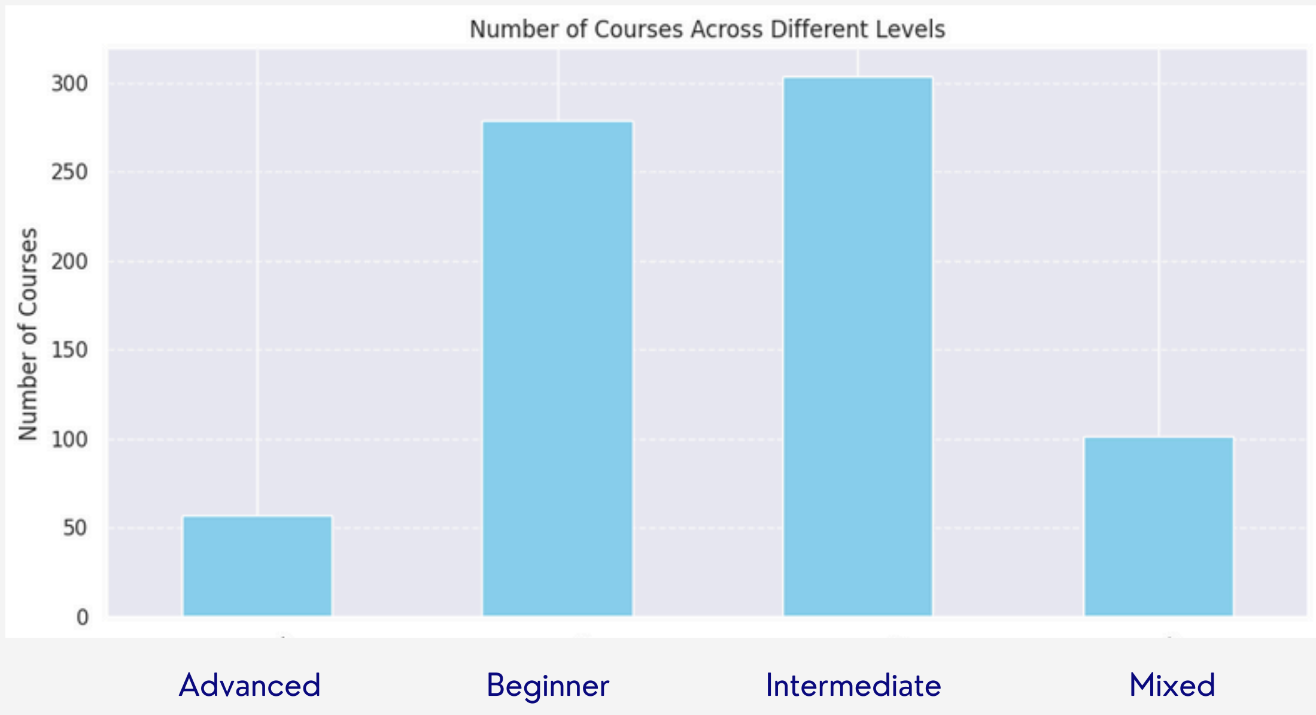
04. DATA COLLECTION

Data was collected by **web scraping** Coursera using **Python's requests library** and **BeautifulSoup** for HTML parsing. Requests were made to Data Science course pages, **filtering** by parameters like **difficulty** and **duration**. Details such as course name, organization, rating, level, and duration were extracted from **course cards.** Adapting to changes in Coursera's HTML structure was a challenge, affecting data extraction accuracy.

05. ANALYSIS

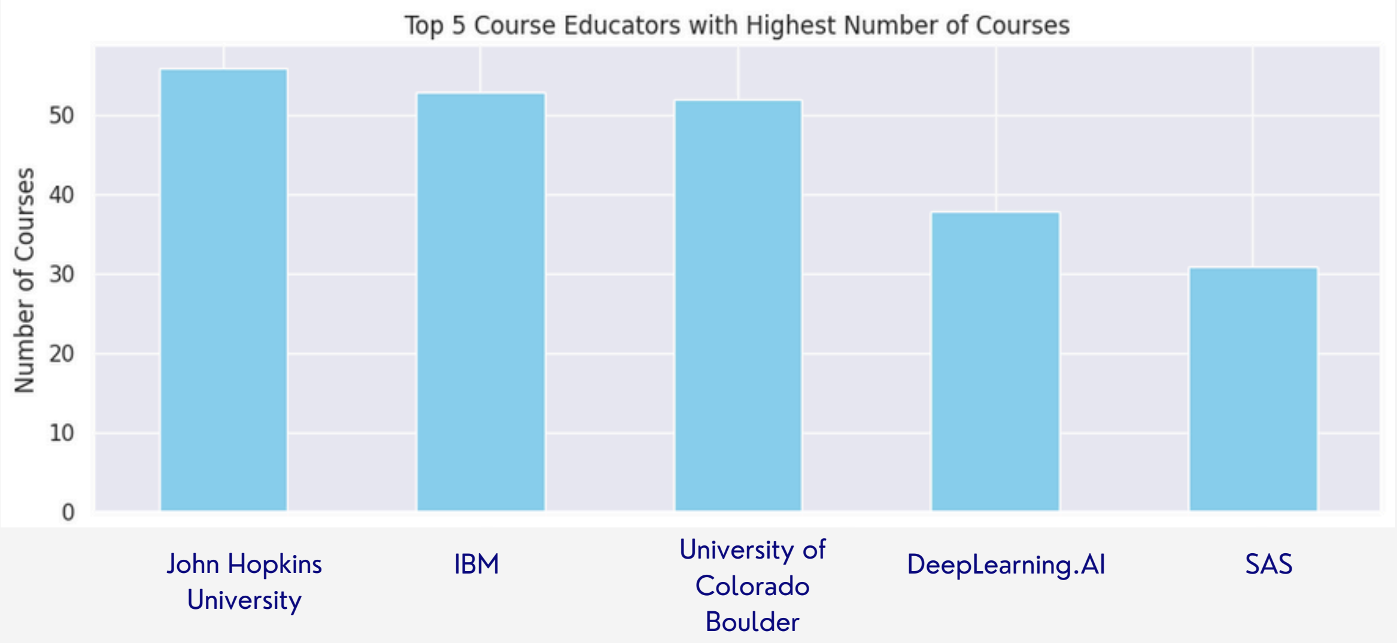
5.1

This chart displays the **enrollment patterns** across various levels, with intermediate having the highest count, followed by beginner, mixed, and advanced, in descending order.



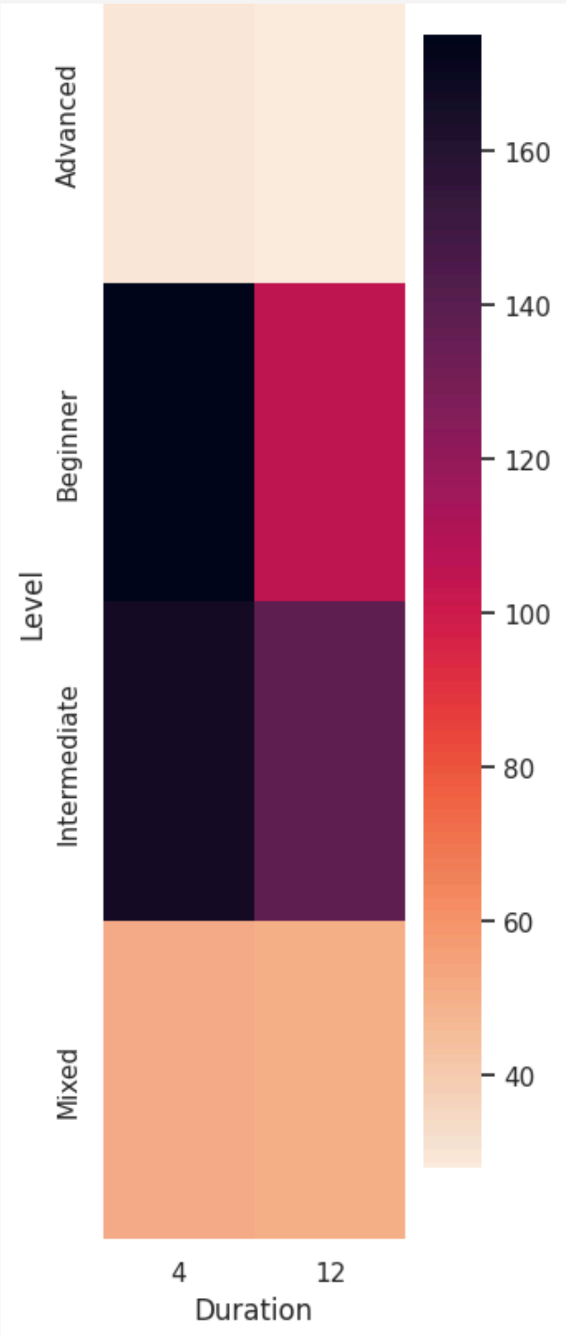
5.2

This chart identifies the **most prominent educators** for data analysis courses on Coursera.



5.3

This chart illustrates the **correlations between course level and duration,** with the beginner level being the most highly correlated level with the duration.



5.4

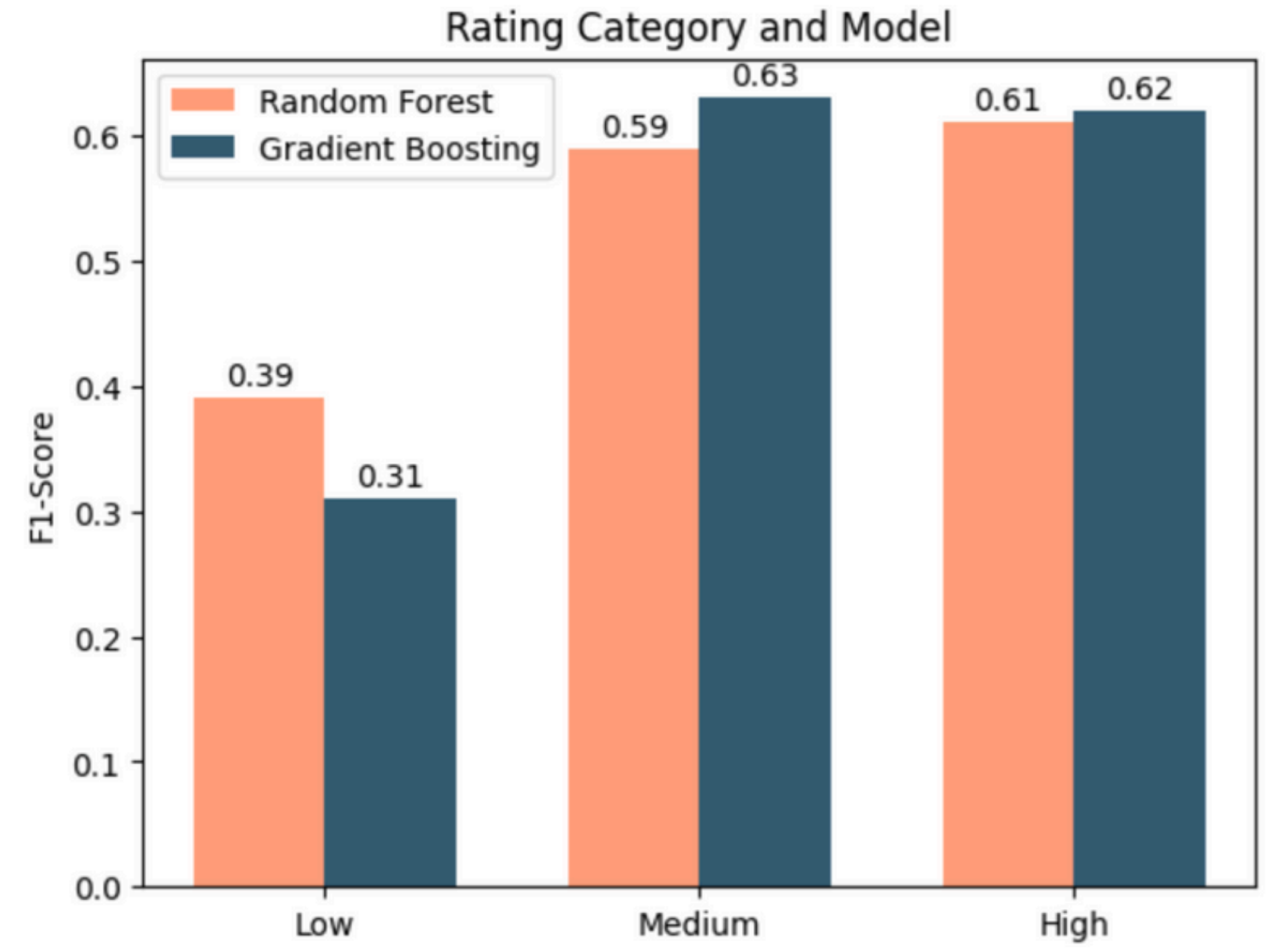
The **top-rated** data analysis courses on Coursera, all boasting a perfect 5-star rating, include:

1. Human Decision Making and its Biases - **495** ratings
2. Foundations of Machine Learning - **518** ratings
3. Setting a Generative AI Strategy - **305** ratings

06. RESULTS/FINDINGS

After analyzing the Coursera courses dataset using Python tools like pandas, matplotlib, and seaborn, we found the data exceptionally clean and well-maintained, showing few missing values or duplicates. Our exploration revealed a pattern in course ratings, indicating a consistent quality across the platform with most courses clustered within certain rating brackets.

In the subsequent classification phase, we preprocessed the data by categorizing the 'Rating' and encoding features such as 'Level' and 'Organization'. We then divided the data into training and testing sets for performance evaluation. Using a Decision Tree as a baseline model, we assessed its performance against more complex models like Random Forest and Gradient Boosting. The Gradient Boosting model proved most effective, equalling the baseline in accuracy and achieving the highest overall F1-score.



The bar chart demonstrates that Gradient Boosting generally outperforms Random Forest, particularly in the 'Medium' and 'High' categories, making it more suitable for complex classifications. However, Random Forest slightly excels in the 'Low' category, indicating its effectiveness in simpler scenarios.

07. CONCLUSION

TThroughout our comprehensive project, we conducted an in-depth analysis of the Coursera courses dataset using Python's powerful libraries,discovering high data quality and consistent course ratings. We preprocessed the data for classification, comparing various models. The Gradient Boosting model excelled, matching our baseline in accuracy and outperforming in F1-score, proving its effectiveness for complex classification tasks. This study highlights the importance of careful data preparation and model evaluation in deriving meaningful insights and achieving precise classification outcomes.