Notebook 1: Case2.ipynb

Objective:

The primary goal of this notebook is to build a machine learning model that predicts whether a student will complete all five recommended subjects based on their demographic, engagement, and academic data. By combining historical performance, engagement with educational content, and personal information, the model provides actionable predictions for student success.

---

1. Data Preparation

The first part of this notebook focuses on loading and merging multiple datasets:

- **Student Demographics**: Contains attributes like `age`, `gender`, `major`, `region`, and `year`. These attributes help provide context about the student's background.
- **Engagement Data**: Tracks behavioral metrics such as `logins per week`, `videos watched`, `time spent on platform`, and `quiz scores`.
- **Performance Data**: Includes course-specific data, such as `courses completed`, `courses started`, and `average score across courses`.

Why these steps matter:

- **Comprehensive Data**: By merging demographic, engagement, and performance data, you create a richer dataset that captures various dimensions of a student's academic life. This multidimensional view allows the model to learn from a wide array of factors influencing student success.
- **Data Merging**: The merging of these datasets is crucial because it combines different aspects of a student's academic behavior into one dataset, making it possible to train a single, unified machine learning model.

Innovative Aspect:

- **Holistic Student View**: By combining data from different sources, this approach enables more robust predictions. It moves beyond traditional models that might rely solely on academic performance and taps into engagement data to capture early warning signs of success or failure.

---

2. Synthetic Data Generation

This step is particularly important due to the relatively small size of the real-world dataset. In this section, 50,000 synthetic student records are generated with random values for attributes like `age`, `gender`, `major`, and academic performance data.

Why this step matters:

- **Data Augmentation**: The goal here is to significantly increase the dataset size, allowing for better generalization of the machine learning model. When the original dataset is small, a model might overfit (learn the noise in the data), so generating additional synthetic data reduces this risk.

Innovation:

- **Synthetic Data to Enhance Model Training**: Using synthetic data to augment a dataset is an innovative strategy when real-world data is limited. In educational systems, gathering large datasets with comprehensive student records can be difficult, and synthetic data offers a solution to bridge this gap. By randomly generating realistic student profiles, the model is trained on a wider variety of scenarios, making it more resilient and able to generalize better to new data.

---

3. Feature Engineering

Feature engineering involves creating new features or transforming existing ones to improve the model's performance. In this case, several key steps include:

- **One-Hot Encoding for Subjects**: The subjects (e.g., AI, Data Science, Cybersecurity, etc.) are one-hot encoded. This means each subject is turned into a binary feature where a "1" indicates that a subject was recommended for that student, and a "0" means it wasn't.
- **Completion Score Calculation**: A custom function is designed to calculate a "completion score" for each student based on their subject preferences (one-hot encoded features), academic performance, and engagement data.

Why this step matters:

- **Subject-Specific Modeling**: The use of one-hot encoding for subjects allows the model to capture how a student's interest and performance in certain subjects impact their overall likelihood of completing the recommended courses.

Innovation:

- **Threshold-Based Completion Scoring**: The innovative aspect here is the introduction of a custom scoring system that computes a "completion score" for each student based on various features, such as engagement and academic history. Setting a threshold (e.g., 0.4) ensures that the model distinguishes between students who are likely to complete the recommended subjects and those who aren't. This scoring system mimics real-world scenarios where multiple factors influence academic success.

---

4. Model Building

The model pipeline includes:

- **Data Preprocessing**: Numeric features (e.g., `courses completed`, `avg quiz score`, etc.) are standardized, and categorical features (subjects) are passed through without any modification. This ensures that the model treats all types of data appropriately.
- **XGBoost Classifier**: The model uses XGBoost, a powerful gradient boosting algorithm, known for its ability to handle tabular data efficiently and deal with missing or imbalanced data.

Why XGBoost matters:

- **Handling Complex Data**: XGBoost is ideal for this task because it can effectively handle datasets with mixed types of features (numeric and categorical). It also performs well with imbalanced datasets, which is likely in educational contexts where the number of students completing all subjects might be much smaller than those who do not.

Innovation:

- **Efficient Model with XGBoost**: Choosing XGBoost offers efficiency and high performance, particularly for structured data like this. The use of a gradient boosting model also ensures that the model learns the most important interactions between features, leading to better predictions.

---

5. Feature Importance Analysis

Once the model is trained, feature importance is computed and visualized to understand which features have the most influence on student completion.

Why this step matters:

- **Actionable Insights**: By analyzing feature importance, educators can determine which factors (e.g., quiz scores, engagement data, or demographic information) most strongly influence student success. This insight can help tailor educational interventions.

Innovation:

- **Data-Driven Educational Strategies**: The ability to identify key predictors of student success (e.g., logins per week or time spent on the platform) can help institutions focus on improving these specific areas. For instance, if the model shows that quiz scores are the most important predictor, educators might prioritize providing better feedback or extra quizzes to struggling students.