**✅ Main Goal:**

Predict the best A/L stream for an O/L student, based on long-term success outcomes like university program, career, and performance.

So your model needs data that connects:

* O/L subject marks & preferences → A/L stream chosen → University course → Career or degree satisfaction.

**🔍 🔢 Alternative Data Collection & Dataset Sources**

**1. 🧠 Kaggle (and other ML Repositories)**

* ✅ **Search Terms to Use:**
  + student career prediction
  + education dataset
  + student performance + university + career
  + academic performance
  + personality + career choice
* 🔗 Example Datasets:
  + Student Performance Dataset
  + Career Prediction Dataset
  + Personality-Career Match Dataset
  + Job Prediction Dataset

⚠️ Most Kaggle datasets may not directly link *O/L → A/L → Degree*, but you can combine them or use for **feature mapping, pretraining, or benchmarking**.

**2. 📊 Open Government Datasets (Sri Lanka / Global)**

* **Sri Lanka:**
  + [Department of Examinations SL](https://www.doenets.lk/)
  + [Ministry of Education](https://moe.gov.lk/)
  + [University Grants Commission SL](https://ugc.ac.lk/)
* **What to Look For:**
  + National results statistics by subject
  + Enrollment by A/L stream
  + University admission cutoffs by A/L stream and district
  + Employment trends by degree field
* **Use case:** Helps with **labeling trends**, ranking universities, and **macro-level demand mapping**.

**3. 🎓 University Student Forums / Alumni Networks**

* Platforms like:
  + **LinkedIn (public profiles with education & career path)**
  + **Reddit: r/SLUni, r/SriLanka**
  + **Facebook Groups of SL Universities**
  + **Quora discussions**
* You can use **web scraping** (carefully respecting privacy rules) or **manual review** to extract:
  + O/L → A/L → Degree → Job transitions
  + Common regrets or satisfaction insights

Great for identifying **patterns from real life**, not just from scores.

**4. 🏫 Partnerships with Schools, Tuition Centers, & Universities**

* You already plan to use Google Forms — that’s great.
* Go further:
  + Ask schools to share **anonymized historical performance data** (past students).
  + Partner with tuition centers: they often track student streams and university outcomes.
  + Reach out to university departments to access **enrollment + degree outcome trends.**

These can become your **main training datasets** if structured well.

**5. 👥 Career Counseling Platforms / LMS Integrations**

If you're open to integrating:

* **Platforms like Coursera, FutureLearn, edX**, and Sri Lankan LMS portals (if they share APIs).
* You can **analyze learning preferences**, career-oriented course selections, and map back to academic stream.

**6. 🧩 Web Scraped Data (If Ethical & Legal)**

* Public student success stories from university websites, job boards, or success blogs.
* Examples:
  + University rank lists by A/L stream
  + Top performer spotlights in each stream
  + Career outcomes of alumni featured online

**🔄 Summary Table**

| **Source** | **Type of Data** | **Usage** |
| --- | --- | --- |
| **Kaggle** | Student performance, career paths | Pretraining, feature alignment |
| **Sri Lankan Government Data** | National trends (O/L, A/L, UGC) | Macro-labeling, stream performance |
| **LinkedIn / Forums** | Individual real-life journeys | Pattern mining |
| **Schools / Tuition Centers** | O/L + A/L + uni entry records | Supervised training data |
| **Online Courses / LMS** | Learning behavior → career interest | Behavioral labeling |
| **Scraping + Surveys** | Public outcome records | Deep but manual insights |

**Key Citations**

* [Faker Library Documentation on PyPI](https://pypi.org/project/Faker/)
* [Synthetic Data Generation by Niklas Heidloff](https://heidloff.net/article/generating-synthetic-data/)
* [CareerVillage Dataset on Kaggle](https://www.kaggle.com/datasets/ryanxjhan/careervillage)
* [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/320/student+performance)
* [O\*NET Resource Center](https://www.onetcenter.org/)
* [KARRIEREWEGE Paper on ResearchGate](https://www.researchgate.net/publication/387263408_KARRIEREWEGE_A_Large_Scale_Career_Path_Prediction_Dataset)
* [Skill & Career Recommendation Dataset on Kaggle](https://www.kaggle.com/datasets/tea340yashjoshi/skill-and-career-recommendation-dataset)
* [Beautiful Soup Documentation](https://www.crummy.com/software/BeautifulSoup/bs4/doc/)
* [Selenium Documentation](https://www.selenium.dev/documentation/)
* [MyMajors Career Quiz](https://www.mymajors.com/)
* [LinkedIn Careers](https://www.linkedin.com/careers)
* [Google Forms](https://www.google.com/forms/about/)
* [Typeform](https://www.typeform.com/)
* [Google Forms Help](https://support.google.com/docs/answer/6281888?hl=en)
* [Reddit r/careerguidance](https://www.reddit.com/r/careerguidance/)
* [Wikipedia API](https://www.mediawiki.org/wiki/API:Main_page)
* [OpenAI API](https://platform.openai.com/docs/api-reference)
* [Hugging Face Models](https://huggingface.co/models)
* [OpenAI Pricing](https://openai.com/pricing)
* [Scrapy Documentation](https://docs.scrapy.org/en/latest/)
* [Kaggle API](https://www.kaggle.com/docs/api)