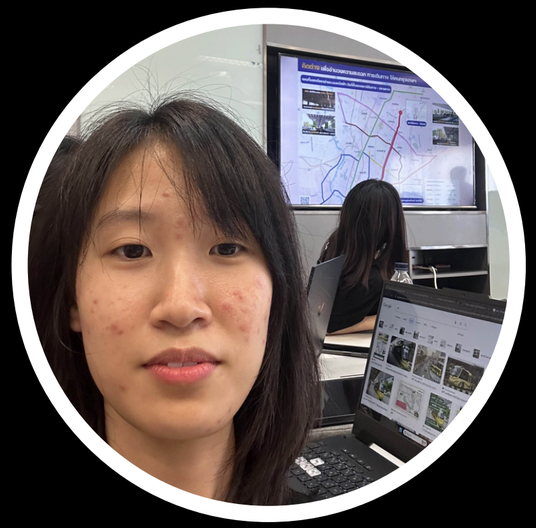
Data Science and Data Engineering Project

2110403 Data Science and Data Engineering (DS-CEDT)

Members

Nawakorn Chamroon
6633063821



Chananut Hussarung
6633045521

Nitchakan Kongkangam
6633076021

Poompat Sritanaanant
6633193721

Content

01 Background

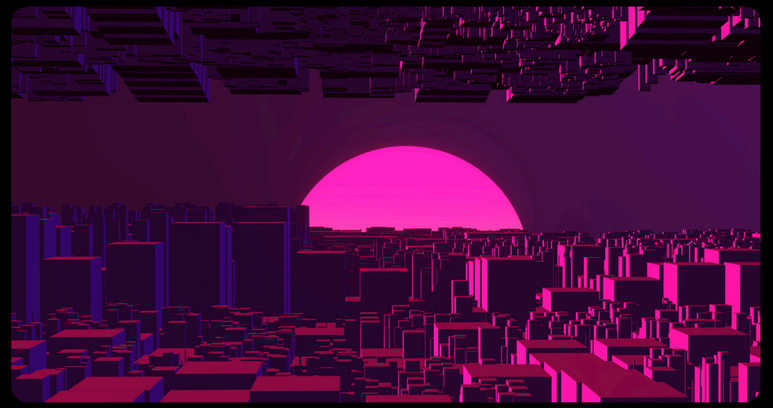
02 Diagram

03 Dataset

04 Data Engineering

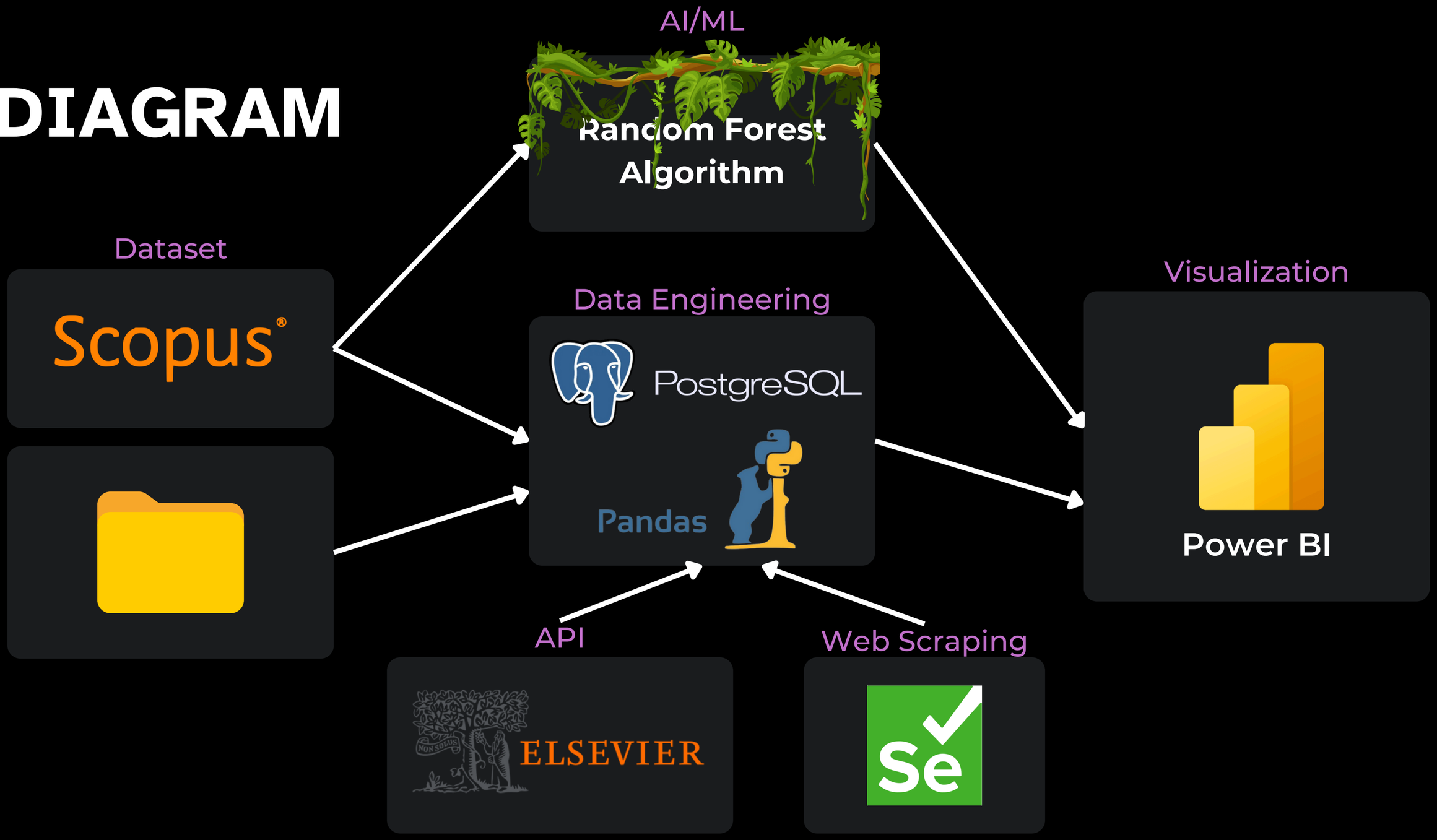
05 AI/ML

06 Visualization



THE TOP SUBJECT AREA FOR EACH COUNTRY IN 2025

DIAGRAM


AI/ML
Random Forest Algorithm
Dataset
Scopus®
Data Engineering
PostgreSQL
Pandas
Visualization
Power BI
API
ELSEVIER
Web Scraping
Se

DATASET

CU Office of Academic Resources

(2018 - 2023)

20216 Papers

Scopus®

Scopus Dataset and
Incorporating additional data sources

(2018 -2024)

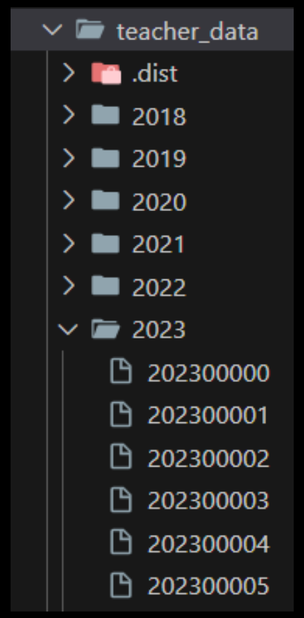
8100 papers and 3805322 Documents

DE

Data Acquisition

1. CU Office of Academic Resources

- dc:title - ชื่อ paper
- affiliation:country - ประเทศ
- prism:coverDate - Published date
- subject-areas - สาขาที่เกี่ยวข้องกับ paper




```
abstracts-retrieval-response": {
  "item": {
      "ait:process-info": {
          "ait:status": {
              "@state": "update",
              "@type": "core",
              "@stage": "S300"
          },
          "ait:date-delivered": {
              "@day": "13",
              "@timestamp": "2023-09-13T08:20:31.000031-04:00",
              "@year": "2023",
              "@month": "09"
          },
          "ait:date-sort": {
              "@day": "31",
              "@year": "2023",
              "@month": "12"
          }
      },
      "xocs:meta": {"xocs:funding-list": {
          "@pui-match": "primary",
          "@has-funding-info": "1",
          "xocs:funding": [
              {
                  "xocs:funding-agency-matched-string": "NSRF",
                  "xocs:funding-id": "B05F640083"
              },
              {
                  "xocs:funding-agency-matched-string": "Chulalongkorn University
                  "xocs:funding-agency-acronym": "CU",
                  "xocs:funding-agency": "Chulalongkorn University",
                  "xocs:funding-agency-id": "http://data.elsevier.com/vocabulary/
                  "xocs:funding-agency-country": "http://sws.geonames.org/1605651
              },
              {
                  "xocs:funding-agency-matched-string": "Thailand Science researc
```

DE

Data Acquisition

2. Scopus Dataset - API

Scopus Search API

https://api.elsevier.com/content/search/scopus


```
_fetch.json > [ ] MULT > {} 0 > [ ] affiliation
"MULT": [
    {
        "link": [
        ],
        "prism:url": "https://api.elsevier.com/content/abstract/scopus_id/85210310976",
        "dc:identifier": "SCOPUS_ID:85210310976",
        "eid": "2-s2.0-85210310976",
        "dc:title": "Emx2 is an essential regulator of ciliated cell development across embryonic tissues",
        "dc:creator": "Nguyen T.K.",
        "prism:publicationName": "iScience",
        "prism:eIssn": "25890042",
        "prism:volume": "27",
        "prism:issueIdentifier": "12",
        "prism:pageRange": null,
        "prism:coverDate": "2024-12-20",
        "prism:coverDisplayDate": "20 December 2024",
        "prism:doi": "10.1016/j.isci.2024.111271",
        "pii": "S2589004224024969",
        "citedby-count": "0",
        "affiliation": [
            {
                "@_fa": "true",
                "affilname": "University of Notre Dame",
                "affiliation-city": "Notre Dame",
                "affiliation-country": "United States"
            }
```

DE

Data Acquisition

2. Scopus Dataset - Web Scraping

ใช้ Selenium และ Pandas

เพื่อดึงข้อมูลจากเว็บไซต์ Scopus และจัดเก็บในรูปแบบ CSV โดยโค้ดมีเป้าหมายเพื่อเก็บข้อมูลของแต่ละประเทศในช่วงปีที่กำหนด รวมถึงจำนวนเอกสารในแต่ละหัวข้อ (Subject Area)

```python
# Create or overwrite the output file with headers
pd.DataFrame(columns=["Country", "Year", "Subject Area", "Number
    output_path, index=False
)

try:
    for country in countries:
        print(f"Processing data for country: {country}")
        # Open the Scopus page
        driver.get("https://www.scopus.com/search/form.uri?displa
        wait_for_element(driver, By.CLASS_NAME, "Select-module__

        # Select "Affiliation country" from the dropdown
        dropdown = Select(driver.find_element(By.CLASS_NAME, "Sel
        dropdown.select_by_visible_text("Affiliation country")

        # Enter the country in the affiliation search bar
        affiliation_search = wait_for_element(
            driver, By.XPATH, '//*[@id="pendo-main-search-bar"]/
        )
        affiliation_search.clear()
        affiliation_search.send_keys(country + Keys.ENTER)
        time.sleep(3)  # Allow results to load

        for year in year_range:
            # Set the year range
            year_from = wait_for_element(
                driver,
                By.XPATH,
                '//*[@id="year-section"]/div/div/div/div[2]/div/
            )
            year_from.clear()
            year_from.send_keys(str(year))

            year_to = wait_for_element(
                driver,
```

```csv
Country,Year,Subject Area,Number of Documents
Afghanistan,2024,Veterinary,9
Afghanistan,2024,Neuroscience,9
Afghanistan,2024,Decision Sciences,7
Afghanistan,2024,Nursing,6
Afghanistan,2024,Health Professions,5
Albania,2018,Computer Science,124
Albania,2018,Medicine,119
Albania,2018,Engineering,99
Albania,2018,Social Sciences,94
Albania,2018,Agricultural and Biological Sciences,71
Albania,2018,Environmental Science,69
Albania,2018,"Economics, Econometrics and Finance",53
Albania,2018,"Business, Management and Accounting",48
Albania,2018,Mathematics,45
Albania,2018,Arts and Humanities,39
```

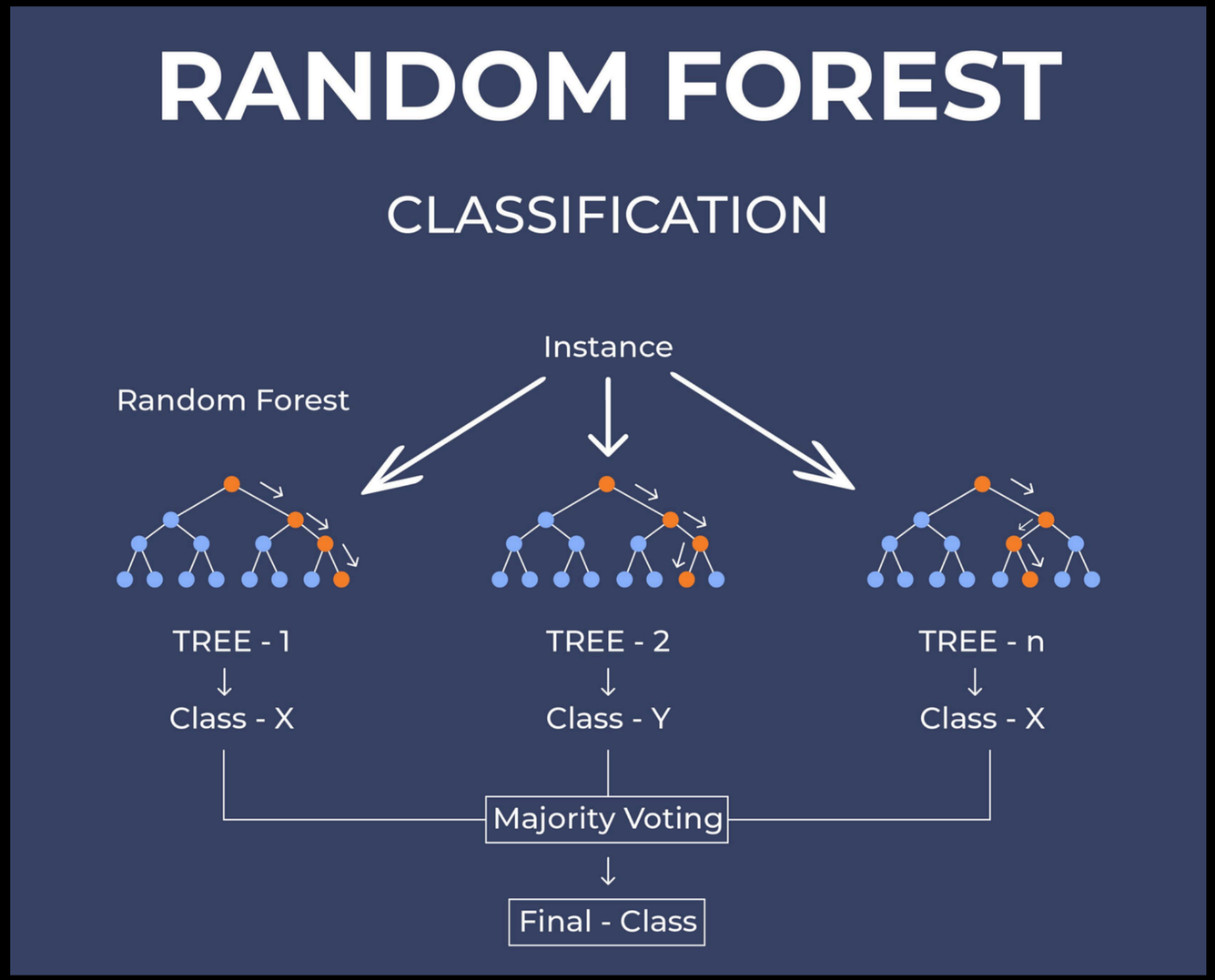
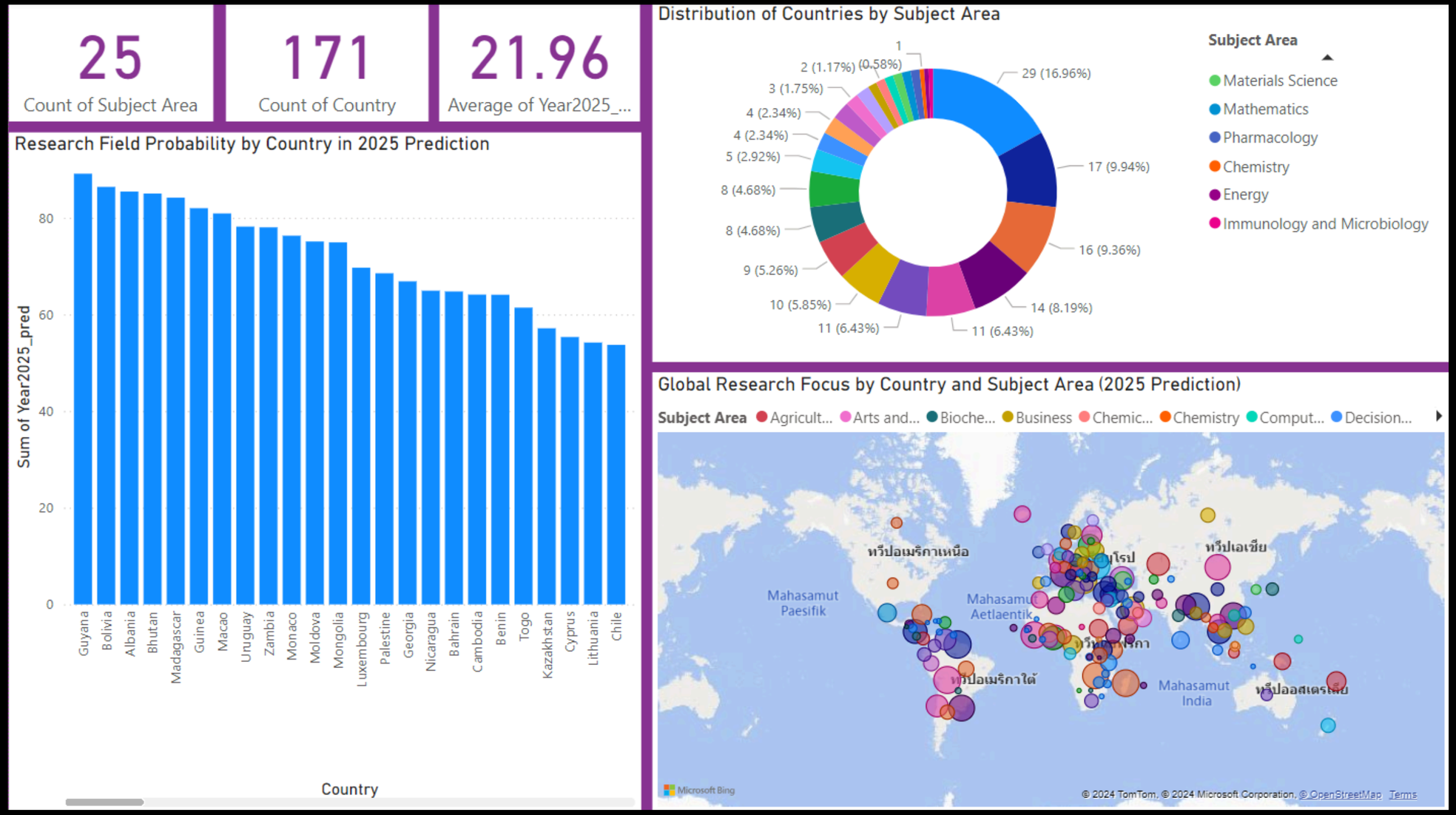
PROBLEMS 21 OUTPUT DEBUG CONSOLE TERMINAL PORTS SERIAL MONITOR

```
Found 27 subject rows.
Data for Albania, 2021 collected successfully!
Close popup Albania, 2021
Found 27 subject rows.
Data for Albania, 2022 collected successfully!
Close popup Albania, 2022
Found 27 subject rows.
Data for Albania, 2023 collected successfully!
Close popup Albania, 2023
Found 27 subject rows.
Data for Albania, 2024 collected successfully!
Close popup Albania, 2024
Data for Albania saved successfully!
Processing data for country: Algeria
Found 27 subject rows.
Data for Algeria, 2018 collected successfully!
Close popup Algeria, 2018
Found 27 subject rows.
Data for Algeria, 2019 collected successfully!
```

DE

Data Tranformation

ML


RANDOM FOREST
CLASSIFICATION
Instance
Random Forest
TREE - 1
Class - X
TREE - 2
Class - Y
TREE - n
Class - X
Majority Voting
Final - Class

VISUALIZATION