# Synthetic Dataset Creation Guide for Indian Student Career Path Recommendation System

### Project Context

**Target Audience**: Indian students completing 12th standard (Age 16-18)
**Purpose**: Recommend optimal career paths based on comprehensive student profiles including academic performance, competitive exam scores, aptitude, interests, and preferences
**Technology Stack**: Kaggle for dataset creation and model training, VS Code (local) for Streamlit interface

## 1. Dataset Overview

### 1.1 Why Synthetic Data?

For career path recommendation targeting Indian 12th standard students, **no comprehensive public dataset exists** that includes:

- Indian competitive exam scores (JEE, NEET, CUET)
- Stream-specific academic profiles (PCM, PCB, Commerce, Arts)
- Aptitude and interest assessments
- Career trajectories spanning 4-5 years post-12th

**Solution**: Generate synthetic dataset that realistically models Indian student profiles and career progressions[67][70][^73].

### 1.2 Dataset Specifications

**Size**: 2,000 student profiles (expandable to 5,000-10,000)
**Features**: 47 features across 6 categories
**Format**: CSV (compatible with Kaggle notebooks)
**Career Paths Covered**: 60+ careers across Engineering, Medical, Commerce, Arts, and Emerging fields

### 1.3 Key Advantages of This Dataset

✓ **Indian Context**: Includes JEE, NEET scores mapped to career outcomes
✓ **Realistic Distributions**: Based on AISHE reports and competitive exam statistics[68][69][72][75]
✓ **Multi-dimensional**: Captures academics, aptitude, interests, and personality
✓ **Temporal**: Tracks 5-year career progression
✓ **Knowledge Graph Ready**: Structured for graph-based RL algorithms

## 2. Dataset Structure

### 2.1 Feature Categories

### Category 1: Demographics (6 features)

| Feature | Type | Values | Purpose |
|---|---|---|---|
| student_id | String | STU_XXXXXX | Unique identifier |
| age | Integer | 16-18 | Student age |
| gender | Categorical | Male/Female/Other | Gender identity |
| state | Categorical | 15 Indian states | Geographic location |
| urban_rural | Categorical | Urban/Semi-Urban/Rural | Area type |
| family_income | Categorical | 5 income brackets | Economic background |

**Rationale**: Demographics influence college access, budget constraints, and location preferences critical for career decisions[69][72].

### Category 2: Academic Background (5 features)

| Feature | Type | Values | Purpose |
|---|---|---|---|
| 12th_stream | Categorical | PCM/PCB/PCMB/Commerce/Arts | Primary determinant of career options |
| 10th_percentage | Float | 60-98% | Academic baseline |
| 12th_percentage | Float | 60-98% | Current academic performance |
| school_board | Categorical | CBSE/ICSE/State | Curriculum standardization |
| school_tier | Categorical | Tier 1/2/3 | School infrastructure quality |

**Distribution**:

- Science (PCM): 35%
- Science (PCB): 25%
- Commerce: 20%
- Science (PCMB): 10%
- Arts: 10%

*Based on AISHE 2021-22 enrollment trends*[72][75].

## Category 3: Competitive Exam Scores (5 features)

| Feature | Type | Values | Purpose |
|---|---|---|---|
| JEE_Main_percentile | Float | 0-100 or null | Engineering college eligibility |
| JEE_Advanced_rank | Integer | 1-250,000 or null | IIT qualification |
| NEET_percentile | Float | 0-100 or null | Medical college eligibility |
| NEET_rank | Integer | 1-2,000,000 or null | MBBS/BDS seat prediction |
| CUET_score | Integer | 200-800 or null | Central university admission |

**Realistic Participation Rates**:

- JEE Main: ~40% of PCM students (12 lakh candidates annually)
- JEE Advanced: ~20% of JEE Main qualifiers (2.5 lakh qualify)[68][71]
- NEET: ~60% of PCB students (20 lakh candidates annually)[77][80]
- CUET: ~30% of all students (expanding adoption)[^75]

**Score Generation Logic**:

```
# JEE Main percentile correlated with 12th marks
base_percentile = 12th_percentage * 0.7 + random.uniform(-10, 15)
JEE_Main_percentile = clip(base_percentile, 45, 99.99)

# JEE Advanced qualification (top 2.5 lakh = ~87+ percentile)
if JEE_Main_percentile &gt; 87:
    JEE_Advanced_rank = int((100 - JEE_Main_percentile) * 2500)
```

## Category 4: Aptitude & Skills (12 features)

**Aptitude Scores** (1-10 scale):

- Logical reasoning
- Quantitative ability
- Verbal ability
- Abstract reasoning
- Spatial reasoning

**Interest Areas** (1-5 scale):

- Technology
- Healthcare
- Business
- Creative arts
- Social service

- Research

**Personality Traits** (1-5 scale):

- Leadership
- Teamwork
- Creativity
- Analytical thinking
- Communication

**Rationale**: Modern career counseling incorporates psychometric assessments for better fit[88][94].

### Category 5: Preferences & Constraints (5 features)

| Feature | Type | Values | Purpose |
|---|---|---|---|
| preferred_location | Categorical | Home State/Nearby/Pan India/Abroad | Geographic flexibility |
| budget_constraint_lakhs | Integer | 2-30 | Education budget (INR) |
| career_goal_timeline | Categorical | 4/5/6+ years | Time to career goal |
| work_preference | Categorical | Job/Business/Research/Govt | Career sector |
| risk_tolerance | Categorical | Low/Medium/High | Career risk appetite |

**Rationale**: Aligns recommendations with practical constraints and personal goals[70][88].

### Category 6: Extracurricular Activities (5 features)

| Feature | Type | Values | Purpose |
|---|---|---|---|
| has_sports | Boolean | True/False | Sports participation |
| has_cultural | Boolean | True/False | Cultural activities |
| volunteering_hours | Integer | 0-100 | Community service |
| num_certifications | Integer | 0-5 | Online learning initiatives |
| num_projects | Integer | 0-5 | Hands-on project experience |

**Rationale**: Holistic profile assessment beyond academics[^75].

## 2.2 Career Path Labels (Target Variables)

### Year-wise Career Progression (5 features)

- **career_path_year1**: Initial career choice (60+ options)
- **career_year2**: Position after Year 2
- **career_year3**: Position after Year 3
- **career_year4**: Position after Year 4

- **career_year5**: Position after Year 5

**Example Trajectory**:

```
Student: STU_000003
Stream: Science-PCM, JEE Main: 85th percentile
Year 1: Computer Science Engineering (B.Tech)
Year 2: Software Engineer (Campus placement)
Year 3: Senior Software Engineer
Year 4: Tech Lead
Year 5: Engineering Manager
```

### Career Categories (60+ paths)

**Engineering** (11 paths):

- Computer Science Engineering
- Mechanical Engineering
- Electrical Engineering
- Electronics Engineering
- Civil Engineering
- Chemical Engineering
- Aerospace Engineering
- Biotechnology
- Information Technology
- AI/ML Engineering
- Data Science

**Medical** (10 paths):

- MBBS, BDS, B.Pharm, Nursing, Physiotherapy
- BAMS, BHMS, Veterinary Science
- Medical Lab Technology, Biotechnology

**Commerce/Business** (10 paths):

- B.Com, BBA, CA, CS, CMA
- Economics, Finance, Marketing
- Actuarial Science, B.Com + CA

**Arts/Humanities** (9 paths):

- BA Psychology, Journalism, Mass Communication
- Law (BA LLB), Design, Fine Arts
- Hotel Management, Social Work, Literature

**Science/Research** (10 paths):

- B.Sc Physics, Chemistry, Mathematics, Biology
- Statistics, Environmental Science, Agriculture
- Forestry, Marine Biology, Astronomy

**Emerging Fields** (10 paths):

- Data Science, AI, Cyber Security, Blockchain
- Game Development, Animation & VFX
- Digital Marketing, Content Creation
- Entrepreneurship, Renewable Energy

## 3. Dataset Generation Methodology

### 3.1 Realistic Distribution Modeling

**Academic Performance**:

```
# Normal distribution centered at 75% (realistic for college-bound students)
12th_percentage ~ Normal(μ=78, σ=8.55)
10th_percentage ~ Normal(μ=75, σ=10)
# Clipped to 60-98% range
```

**Competitive Exam Scores**:

```
# Correlation with 12th marks
JEE_Main_percentile = 12th_percentage * 0.7 + noise(-10, 15)
NEET_percentile = 12th_percentage * 0.75 + noise(-12, 12)

# Realistic participation rates
JEE_appeared = 40% of PCM students
NEET_appeared = 60% of PCB students
```

**Career Assignment Logic**:

```
def assign_career(student):
    if stream == 'PCM' and JEE_percentile &gt; 90:
        # Top performers prefer CSE, AI/ML
        weights = [CSE: 0.35, Data Science: 0.20, AI/ML: 0.15, ...]
    elif stream == 'PCB' and NEET_percentile &gt; 95:
        # High scorers target MBBS
        weights = [MBBS: 0.60, BDS: 0.10, ...]
    else:
        # Balanced distribution
```

### 3.2 Career Progression Rules

**Progression Mapping**:

```
progression_map = {
    'Computer Science Engineering': [
        'Software Engineer' → 'Data Scientist' →
        'Senior Software Engineer' → 'Tech Lead' →
        'Engineering Manager'
    ],
    'MBBS': [
        'Junior Doctor' → 'Resident Doctor' →
        'Senior Resident' → 'Specialist' →
        'Consultant'
    ],
    'CA': [
        'Articleship' → 'Qualified CA' →
        'Finance Manager' → 'CFO'
    ]
}
```

**Progression Probability**:

- 70% chance of advancement each year
- 30% chance of staying in current role (realistic stagnation)

### 3.3 Data Quality Checks

✓ **Consistency**: JEE scores only for PCM/PCMB students
✓ **Realism**: Percentiles clipped to 60-98% (college-bound population)
✓ **Diversity**: Balanced representation across states, genders, streams
✓ **Completeness**: No missing values in core features (nulls only in optional exams)

## 4. Using the Dataset

### 4.1 Dataset Files Generated

**1. indian_students_career_dataset_synthetic_2000.csv**

- **Size**: 2,000 rows × 47 columns
- **Format**: CSV with headers
- **Encoding**: UTF-8
- **Missing Values**: Represented as empty cells (for competitive exams not appeared)

**2. data_dictionary.csv**

- **Purpose**: Complete codebook
- **Contents**: Feature name, type, range, description

- **Use**: Reference for understanding each column

### 4.2 Loading in Kaggle Notebook

```
import pandas as pd
import numpy as np

# Load dataset
df = pd.read_csv('/kaggle/input/your-dataset-name/indian_students_career_dataset_syntheti

# Quick overview
print(f"Dataset Shape: {df.shape}")
print(f"Features: {df.columns.tolist()}")
print(df.head())

# Check missing values
print(df.isnull().sum())

# Statistical summary
print(df.describe())
```

### 4.3 Data Preprocessing for RL Model

#### Step 1: Handle Missing Values

```
# Competitive exams: Fill null with -1 (not appeared)
df['JEE_Main_percentile'].fillna(-1, inplace=True)
df['JEE_Advanced_rank'].fillna(-1, inplace=True)
df['NEET_percentile'].fillna(-1, inplace=True)
df['NEET_rank'].fillna(-1, inplace=True)
df['CUET_score'].fillna(-1, inplace=True)
```

#### Step 2: Encode Categorical Variables

```
from sklearn.preprocessing import LabelEncoder

categorical_cols = ['gender', 'state', 'urban_rural', 'family_income',
                    '12th_stream', 'school_board', 'school_tier',
                    'preferred_location', 'career_goal_timeline',
                    'work_preference', 'risk_tolerance']

label_encoders = {}
for col in categorical_cols:
    le = LabelEncoder()
    df[col + '_encoded'] = le.fit_transform(df[col])
    label_encoders[col] = le
```

#### Step 3: Normalize Numerical Features

```
from sklearn.preprocessing import MinMaxScaler

numerical_cols = ['10th_percentage', '12th_percentage',
                  'logical_reasoning', 'quantitative_ability', 'verbal_ability',
                  'interest_technology', 'interest_healthcare',
                  'leadership', 'teamwork', 'creativity']

scaler = MinMaxScaler()
df[numerical_cols] = scaler.fit_transform(df[numerical_cols])
```

**Step 4: Create Feature Vector**

```
def create_feature_vector(student_row):
    features = []

    # Demographics (encoded)
    features.extend([
        student_row['gender_encoded'],
        student_row['urban_rural_encoded'],
        student_row['family_income_encoded']
    ])

    # Academic (normalized)
    features.extend([
        student_row['10th_percentage'],
        student_row['12th_percentage'],
        student_row['12th_stream_encoded']
    ])

    # Competitive exams (normalized)
    features.extend([
        student_row['JEE_Main_percentile'] / 100 if student_row['JEE_Main_percentile'] &g
        student_row['NEET_percentile'] / 100 if student_row['NEET_percentile'] &gt; 0 els
    ])

    # Aptitude scores (already 1-10, normalize to 0-1)
    features.extend([
        student_row['logical_reasoning'],
        student_row['quantitative_ability'],
        student_row['verbal_ability']
    ])

    # Interest areas (already 1-5, normalize to 0-1)
    features.extend([
        student_row['interest_technology'],
        student_row['interest_healthcare'],
        student_row['interest_business']
    ])

    # Personality traits
    features.extend([
        student_row['leadership'],
        student_row['teamwork'],
        student_row['analytical_thinking']
```

```
    ])

    return np.array(features)

# Apply to all students
X = df.apply(create_feature_vector, axis=1)
X = np.vstack(X.values)

print(f"Feature matrix shape: {X.shape}")  # Should be (2000, ~20-30 features)
```

### 4.4 Creating Career Knowledge Graph

```
import networkx as nx
import matplotlib.pyplot as plt

# Build graph from career trajectories
G = nx.DiGraph()

# Add nodes (unique career positions)
all_careers = set()
for col in ['career_path_year1', 'career_year2', 'career_year3',
            'career_year4', 'career_year5']:
    all_careers.update(df[col].unique())

for career in all_careers:
    G.add_node(career)

# Add edges (career transitions)
for idx, row in df.iterrows():
    trajectory = [
        row['career_path_year1'],
        row['career_year2'],
        row['career_year3'],
        row['career_year4'],
        row['career_year5']
    ]

    for i in range(len(trajectory) - 1):
        if G.has_edge(trajectory[i], trajectory[i+1]):
            G[trajectory[i]][trajectory[i+1]]['weight'] += 1
        else:
            G.add_edge(trajectory[i], trajectory[i+1], weight=1)

print(f"Knowledge Graph: {len(G.nodes())} nodes, {len(G.edges())} edges")

# Visualize subset
plt.figure(figsize=(12, 8))
pos = nx.spring_layout(G, k=0.5)
nx.draw(G, pos, with_labels=True, node_size=100, font_size=6,
        arrows=True, edge_color='gray', alpha=0.6)
plt.title("Career Path Knowledge Graph")
plt.savefig('knowledge_graph.png', dpi=300, bbox_inches='tight')
plt.show()
```

## 5. Student Input Questionnaire Design

### 5.1 Questionnaire Structure

The student input module should collect **exactly the same features** as the dataset to ensure compatibility with the trained model.

**Total Questions**: 35-40 questions
**Time to Complete**: 10-15 minutes
**Format**: Web form (Streamlit)

### 5.2 Questionnaire Categories

#### Section 1: Personal Information (5 questions)

```
Q1. Age: [^16] [^17] [^18]
Q2. Gender: ○ Male ○ Female ○ Other ○ Prefer not to say
Q3. State: [Dropdown: 29 states + 8 UTs]
Q4. Area Type: ○ Urban ○ Semi-Urban ○ Rural
Q5. Family Annual Income:
    ○ Less than 5 Lakhs
    ○ 5-10 Lakhs
    ○ 10-20 Lakhs
    ○ 20-50 Lakhs
    ○ Above 50 Lakhs
```

#### Section 2: Academic Background (7 questions)

```
Q6. 12th Standard Stream:
    ○ Science - PCM (Physics, Chemistry, Mathematics)
    ○ Science - PCB (Physics, Chemistry, Biology)
    ○ Science - PCMB (All four)
    ○ Commerce
    ○ Arts/Humanities

Q7. 10th Standard Percentage: [Slider: 60-100%] or [Number input]

Q8. 12th Standard Percentage (Expected/Actual): [Slider: 60-100%]

Q9. School Board:
    ○ CBSE
    ○ ICSE
    ○ State Board
    ○ Other

Q10. School Type:
     ○ Tier 1 (Well-equipped, high resources)
     ○ Tier 2 (Adequate facilities)
     ○ Tier 3 (Basic facilities)

Q11. Have you appeared for JEE Main?
```

```
     ○ Yes → Q12: Percentile score: [___]
     ○ No

Q12. Have you appeared for NEET?
     ○ Yes → Q13: Percentile score: [___]
     ○ No
```

### Section 3: Aptitude Assessment (5 questions)

*Use sliders (1-10 scale) with descriptive anchors*

```
Q14. Logical Reasoning: How well can you solve puzzles and identify patterns?
     [1 - Poor] -------- [5 - Average] -------- [10 - Excellent]

Q15. Quantitative Ability: How comfortable are you with numbers and calculations?
     [1 - Poor] -------- [5 - Average] -------- [10 - Excellent]

Q16. Verbal Ability: How strong are your reading, writing, and communication skills?
     [1 - Poor] -------- [5 - Average] -------- [10 - Excellent]

Q17. Abstract Reasoning: Can you understand complex concepts and think creatively?
     [1 - Poor] -------- [5 - Average] -------- [10 - Excellent]

Q18. Spatial Reasoning: Can you visualize shapes and understand 3D objects?
     [1 - Poor] -------- [5 - Average] -------- [10 - Excellent]
```

### Section 4: Interest Areas (6 questions)

*Use sliders (1-5 scale)*

```
Q19. How interested are you in TECHNOLOGY (computers, coding, engineering)?
     [1 - Not at all] --- [3 - Somewhat] --- [5 - Very interested]

Q20. How interested are you in HEALTHCARE (medicine, nursing, helping patients)?
     [1 - Not at all] --- [3 - Somewhat] --- [5 - Very interested]

Q21. How interested are you in BUSINESS (finance, marketing, entrepreneurship)?
     [1 - Not at all] --- [3 - Somewhat] --- [5 - Very interested]

Q22. How interested are you in CREATIVE ARTS (design, media, performing arts)?
     [1 - Not at all] --- [3 - Somewhat] --- [5 - Very interested]

Q23. How interested are you in SOCIAL SERVICE (NGOs, teaching, public service)?
     [1 - Not at all] --- [3 - Somewhat] --- [5 - Very interested]

Q24. How interested are you in RESEARCH (science, innovation, discovery)?
     [1 - Not at all] --- [3 - Somewhat] --- [5 - Very interested]
```

## Section 5: Personality Traits (5 questions)

Q25. Leadership: I can lead teams and take initiative
[1 - Disagree] --- [3 - Neutral] --- [5 - Strongly Agree]

Q26. Teamwork: I work well with others
[1 - Disagree] --- [3 - Neutral] --- [5 - Strongly Agree]

Q27. Creativity: I enjoy thinking of new ideas and solutions
[1 - Disagree] --- [3 - Neutral] --- [5 - Strongly Agree]

Q28. Analytical Thinking: I like analyzing data and solving problems logically
[1 - Disagree] --- [3 - Neutral] --- [5 - Strongly Agree]

Q29. Communication: I can express my ideas clearly
[1 - Disagree] --- [3 - Neutral] --- [5 - Strongly Agree]

## Section 6: Career Preferences (5 questions)

Q30. Where do you prefer to study?
- ○ In my home state
- ○ Nearby states
- ○ Anywhere in India
- ○ Abroad

Q31. What is your annual budget for education (in Lakhs)?
- ○ 2-5 Lakhs
- ○ 5-10 Lakhs
- ○ 10-20 Lakhs
- ○ 20-30 Lakhs
- ○ Above 30 Lakhs

Q32. Timeline to achieve career goal:
- ○ 4 years (Bachelor's degree)
- ○ 5 years (Bachelor's + initial job)
- ○ 6+ years (Master's/Professional course)

Q33. Preferred work type:
- ○ Private Job
- ○ Business/Entrepreneurship
- ○ Research & Academia
- ○ Government Job

Q34. Risk tolerance in career choices:
- ○ Low (Prefer stable, traditional careers)
- ○ Medium (Open to some experimentation)
- ○ High (Ready for unconventional paths)

### Section 7: Extracurricular Activities (5 questions)

```
Q35. Do you participate in sports?
     ○ Yes (Specify: [___]) ○ No

Q36. Do you participate in cultural activities (music, dance, drama)?
     ○ Yes ○ No

Q37. Hours spent on volunteering/community service per year:
     [Slider: 0-100 hours]

Q38. Number of online certifications completed:
     [Dropdown: 0, 1, 2, 3, 4, 5+]

Q39. Number of hands-on projects completed:
     [Dropdown: 0, 1, 2, 3, 4, 5+]
```

## 5.3 Streamlit Implementation

**File**: `student_input_form.py`

```
import streamlit as st
import pandas as pd
import numpy as np

st.set_page_config(page_title="Career Path Recommender", layout="wide")

st.title(" AI-Powered Career Path Recommendation")
st.markdown("*For Indian students completing 12th standard*")

# Initialize session state
if 'student_data' not in st.session_state:
    st.session_state.student_data = {}

# Create form
with st.form("student_profile_form"):
    st.header(" Student Profile Questionnaire")

    # Section 1: Personal Information
    st.subheader("1. Personal Information")
    col1, col2 = st.columns(2)

    with col1:
        age = st.selectbox("Age", [16, 17, 18])
        gender = st.selectbox("Gender", ["Male", "Female", "Other", "Prefer not to say"])
        state = st.selectbox("State", [
            "Maharashtra", "Karnataka", "Tamil Nadu", "Delhi", "Uttar Pradesh",
            "West Bengal", "Gujarat", "Rajasthan", "Kerala", "Punjab"
            # ... add all 29 states + 8 UTs
        ])

    with col2:
        urban_rural = st.selectbox("Area Type", ["Urban", "Semi-Urban", "Rural"])
        family_income = st.selectbox("Family Annual Income", [
```

```
            "Less than 5 Lakhs", "5-10 Lakhs", "10-20 Lakhs",
            "20-50 Lakhs", "Above 50 Lakhs"
        ])

    # Section 2: Academic Background
    st.subheader("2. Academic Background")
    col1, col2 = st.columns(2)

    with col1:
        stream_12th = st.selectbox("12th Standard Stream", [
            "Science - PCM", "Science - PCB", "Science - PCMB",
            "Commerce", "Arts/Humanities"
        ])
        percentage_10th = st.slider("10th Standard Percentage", 60.0, 100.0, 75.0, 0.5)
        percentage_12th = st.slider("12th Standard Percentage", 60.0, 100.0, 75.0, 0.5)

    with col2:
        school_board = st.selectbox("School Board", ["CBSE", "ICSE", "State Board", "Othe
        school_tier = st.selectbox("School Type", [
            "Tier 1 (Well-equipped)",
            "Tier 2 (Adequate)",
            "Tier 3 (Basic)"
        ])

    # Competitive Exams
    st.markdown("**Competitive Exam Scores**")
    col1, col2 = st.columns(2)

    with col1:
        jee_appeared = st.checkbox("Appeared for JEE Main")
        if jee_appeared:
            jee_percentile = st.number_input("JEE Main Percentile", 0.0, 100.0, 75.0, 0.0
        else:
            jee_percentile = None

    with col2:
        neet_appeared = st.checkbox("Appeared for NEET")
        if neet_appeared:
            neet_percentile = st.number_input("NEET Percentile", 0.0, 100.0, 75.0, 0.01)
        else:
            neet_percentile = None

    # Section 3: Aptitude Assessment
    st.subheader("3. Aptitude Self-Assessment (1-10 scale)")
    col1, col2 = st.columns(2)

    with col1:
        logical = st.slider("Logical Reasoning", 1, 10, 5)
        quant = st.slider("Quantitative Ability", 1, 10, 5)
        verbal = st.slider("Verbal Ability", 1, 10, 5)

    with col2:
        abstract = st.slider("Abstract Reasoning", 1, 10, 5)
        spatial = st.slider("Spatial Reasoning", 1, 10, 5)

    # Section 4: Interest Areas
```

```
    st.subheader("4. Interest Areas (1-5 scale)")
    col1, col2, col3 = st.columns(3)

    with col1:
        interest_tech = st.slider("Technology &amp;amp; Engineering", 1, 5, 3)
        interest_health = st.slider("Healthcare &amp;amp; Medicine", 1, 5, 3)

    with col2:
        interest_business = st.slider("Business &amp;amp; Finance", 1, 5, 3)
        interest_arts = st.slider("Creative Arts &amp;amp; Design", 1, 5, 3)

    with col3:
        interest_social = st.slider("Social Service &amp;amp; Education", 1, 5, 3)
        interest_research = st.slider("Research &amp;amp; Science", 1, 5, 3)

    # Section 5: Personality Traits
    st.subheader("5. Personality Traits (1-5 scale)")
    col1, col2 = st.columns(2)

    with col1:
        leadership = st.slider("Leadership", 1, 5, 3)
        teamwork = st.slider("Teamwork", 1, 5, 3)
        creativity = st.slider("Creativity", 1, 5, 3)

    with col2:
        analytical = st.slider("Analytical Thinking", 1, 5, 3)
        communication = st.slider("Communication", 1, 5, 3)

    # Section 6: Career Preferences
    st.subheader("6. Career Preferences")
    col1, col2 = st.columns(2)

    with col1:
        location_pref = st.selectbox("Preferred Study Location", [
            "Home State", "Nearby States", "Anywhere in India", "Abroad"
        ])
        budget = st.selectbox("Annual Education Budget (Lakhs)", [
            "2-5", "5-10", "10-20", "20-30", "30+"
        ])

    with col2:
        timeline = st.selectbox("Career Goal Timeline", [
            "4 years", "5 years", "6+ years"
        ])
        work_pref = st.selectbox("Preferred Work Type", [
            "Private Job", "Business/Entrepreneurship",
            "Research &amp;amp; Academia", "Government Job"
        ])
        risk_tolerance = st.selectbox("Risk Tolerance", [
            "Low (Stable careers)", "Medium", "High (Unconventional)"
        ])

    # Section 7: Extracurricular
    st.subheader("7. Extracurricular Activities")
    col1, col2 = st.columns(2)
```

```
    with col1:
        has_sports = st.checkbox("Participate in sports")
        has_cultural = st.checkbox("Participate in cultural activities")
        volunteering = st.slider("Volunteering hours per year", 0, 100, 0, 5)

    with col2:
        certifications = st.selectbox("Online certifications completed", [0, 1, 2, 3, 4,
        projects = st.selectbox("Projects completed", [0, 1, 2, 3, 4, 5])

    # Submit button
    submitted = st.form_submit_button("🔮 Get Career Recommendations", use_container_width

    if submitted:
        # Collect all data
        student_data = {
            'age': age,
            'gender': gender,
            'state': state,
            'urban_rural': urban_rural,
            'family_income': family_income,
            '12th_stream': stream_12th,
            '10th_percentage': percentage_10th,
            '12th_percentage': percentage_12th,
            'school_board': school_board,
            'school_tier': school_tier,
            'JEE_Main_percentile': jee_percentile,
            'NEET_percentile': neet_percentile,
            'logical_reasoning': logical,
            'quantitative_ability': quant,
            'verbal_ability': verbal,
            'abstract_reasoning': abstract,
            'spatial_reasoning': spatial,
            'interest_technology': interest_tech,
            'interest_healthcare': interest_health,
            'interest_business': interest_business,
            'interest_creative_arts': interest_arts,
            'interest_social_service': interest_social,
            'interest_research': interest_research,
            'leadership': leadership,
            'teamwork': teamwork,
            'creativity': creativity,
            'analytical_thinking': analytical,
            'communication': communication,
            'preferred_location': location_pref,
            'budget_constraint_lakhs': int(budget.split('-')[^0]),
            'career_goal_timeline': timeline,
            'work_preference': work_pref,
            'risk_tolerance': risk_tolerance,
            'has_sports': has_sports,
            'has_cultural': has_cultural,
            'volunteering_hours': volunteering,
            'num_certifications': certifications,
            'num_projects': projects
        }

        st.session_state.student_data = student_data
```

```
        # TODO: Call RL model for recommendations
        st.success("✅ Profile saved! Generating recommendations...")

        # Display collected data (for debugging)
        with st.expander("View Submitted Profile"):
            st.json(student_data)
```

## 6. Next Steps for Implementation

### 6.1 Immediate Actions (Week 1)

**Day 1-2**: Dataset Expansion

- [ ] Increase dataset size to 5,000 students
- [ ] Add more career paths (target 80-100 total)
- [ ] Validate distributions against AISHE 2021-22 data

**Day 3-4**: Kaggle Setup

- [ ] Upload dataset to Kaggle
- [ ] Create Kaggle notebook for data exploration
- [ ] Publish dataset as public/private dataset

**Day 5-7**: Questionnaire Development

- [ ] Build complete Streamlit form (40 questions)
- [ ] Test locally with sample inputs
- [ ] Create input validation logic

### 6.2 Model Training (Week 2-3)

**Kaggle Notebook Tasks**:

1. Load and preprocess dataset
2. Build knowledge graph from trajectories
3. Implement CareerEnvironment (RL environment)
4. Adapt existing DDQN code for career domain
5. Train model for 500-1000 episodes
6. Evaluate with metrics (path validity, skill alignment)
7. Save trained model weights

### 6.3 Integration (Week 4)

- Connect Streamlit form to trained model
- Implement recommendation generation
- Create result visualization
- Test end-to-end workflow

## 7. Dataset Quality Assurance

### 7.1 Validation Checks

✓ **Statistical Validation**:

```
# Check distributions
assert df['12th_percentage'].mean() &gt; 75 and df['12th_percentage'].mean() &lt; 80
assert df['JEE_Main_percentile'].count() / len(df) &lt; 0.25  # ~20% appearance rate
assert df['NEET_percentile'].count() / len(df) &lt; 0.30  # ~25% appearance rate
```

✓ **Logical Validation**:

```
# JEE scores only for science students
assert df[df['JEE_Main_percentile'].notna()]['12th_stream'].isin(['Science-PCM', 'Science

# NEET scores only for biology students
assert df[df['NEET_percentile'].notna()]['12th_stream'].isin(['Science-PCB', 'Science-PCM
```

✓ **Career Path Validation**:

```
# Engineering careers only for PCM students
engineering_careers = ['Computer Science Engineering', 'Mechanical Engineering', ...]
assert df[df['career_path_year1'].isin(engineering_careers)]['12th_stream'].isin(['Scienc
```

### 7.2 Expansion Strategies

**To 5,000 students**:

```
# Simple approach: Generate 3,000 more students
for i in range(2000, 5000):
    students.append(generate_student_profile(i))
```

**To 10,000 students (with data augmentation)**:

```
# Add noise to existing students
def augment_student(student):
    augmented = student.copy()
    augmented['12th_percentage'] += np.random.normal(0, 2)
```

```
    augmented['logical_reasoning'] += np.random.normal(0, 0.5)
    # ... apply slight variations
    return augmented

# Generate 5,000 new + 5,000 augmented
```

## 8. Conclusion

This synthetic dataset provides a **realistic, comprehensive foundation** for building an AI-powered career recommendation system tailored to Indian students completing 12th standard.

**Key Strengths**:
✓ **Authentic Indian Context**: JEE, NEET, stream-based pathways
✓ **Multi-dimensional**: Academics + aptitude + interests + preferences
✓ **Temporal**: 5-year career progression for RL training
✓ **Ready-to-Use**: Compatible with existing DDQN implementation
✓ **Scalable**: Easily expandable to 10,000+ students

**Alignment with Project Goals**:
✓ Questionnaire matches dataset features → seamless integration
✓ Knowledge graph structure → compatible with RL algorithms
✓ Kaggle-ready → smooth training workflow
✓ Streamlit-compatible → clean local deployment

**Next**: Proceed to model training in Kaggle notebook using the generated dataset and knowledge graph structure.

## Appendix: Dataset Statistics Summary

**Demographics**:

- Gender: Male (52%), Female (47%), Other (1%)
- Urban (40%), Semi-Urban (35%), Rural (25%)
- 15 states represented

**Academic**:

- Streams: Science-PCM (35%), Science-PCB (25%), Commerce (20%), Science-PCMB (10%), Arts (10%)
- Mean 12th %: 78.0 ± 8.55
- Boards: CBSE (45%), State (40%), ICSE (15%)

**Competitive Exams**:

- JEE Main appeared: 18.7%
- NEET appeared: 21.2%
- CUET appeared: 28.5%

**Top Career Paths**:

1. Computer Science Engineering (138 students)
2. B.Pharm (124)
3. Mechanical Engineering (119)
4. B.Com (98)
5. Data Science (97)

**Files Generated**:

1. `indian_students_career_dataset_synthetic_2000.csv` - Main dataset
2. `data_dictionary.csv` - Feature codebook
3. `student_input_form.py` - Streamlit questionnaire (to be created)

*Document Version: 1.0*

*Last Updated: October 2025*

*For questions or modifications, refer to dataset generation code*
[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36]

⁂


1. https://articles.cbseguess.com/best-career-options-in-india/
2. https://amityonline.com/blog/list-of-best-career-options-after-12th
3. https://www.indeduconsultancy.com/neet-2025-ultimate-predictor-the-most-accurate-rank-percentile-college/neet-marks-to-rank-and-percentile-calculator
4. https://www.dataforindia.com/higher-education/
5. https://www.sdseed.in/docs/Career Handbook.pdf
6. https://predictor.v4edu.in
7. https://www.kaggle.com/datasets/adilshamim8/education-and-career-success
8. https://www.upgrad.com/blog/career-options-after-pcmb/
9. https://www.shiksha.com/engineering/jee-main-college-predictor
10. https://www.kaggle.com/datasets/mrutyunjaybiswal/iitjee-neet-aims-students-questions-data
11. https://jims.in/blogs/career-opportunities-after-data-science-course-india
12. https://www.dheya.com/report-on-career-counselling/
13. https://www.aakash.ac.in/blog/top-iits-offering-b-tech-in-artificial-intelligence-and-data-science-in-2024/
14. https://www.onlinemanipal.com/blogs/data-science-career-path
15. https://www.slideshare.net/slideshow/career-guidance-career-options-for-class-xii-students/251647724
16. https://www.aaaedu.in/knowledge-desk/neet-and-jee-mastery-in-the-digital-age-the-data-science-advantage
17. https://nareshit.com/blogs/job-demand-for-data-scientists-in-india-for-freshers
18. https://www.ijcrt.org/papers/IJCRT2105009.pdf
19. https://testbook.com/question-answer/what-type-of-ai-model-did-the-indian-institute-of--68795d71d6b070de495ca798

20. https://media.collegedekho.com/media/django-summernote/2024-06-12/dd28e137-0673-49a9-b5ee-88824ca0af0a.pdf
21. https://in.indeed.com/q-synthetic-data-jobs.html
22. https://setmycareer.com/career-counselling-after-12th.php
23. https://jeemain.nta.ac.in/images/information-bulletin-for-jee-main-2024.pdf
24. https://yourstory.com/2025/08/ai-startups-indika-onix-kroop-synthetic-data-platforms-solutions
25. https://labour.py.gov.in/sites/default/files/lrdeestudentguide2016.pdf
26. https://legiit.com/HammadImtiaz/synthetic-data-generation
27. https://www.timesjobs.com/job-detail/master-thesis-synthetic-data-generation-for-camera-based-perception-in-amrs-volvo-group-india-pvt-ltd-sweden-1-to-2-yrs-jobid-zZx0Ll9obQJzpSvf__PLUS__uAgZw==
28. https://www.kaggle.com/datasets/breejeshdhar/career-recommendation-dataset
29. https://www.ietlucknow.ac.in/content/5322
30. https://www.simplilearn.com/tutorials/data-science-tutorial/synthetic-data-generation
31. https://www.mindgroom.in/blogs/exploring-career-paths-after-12th-your-guide-to-success-in-2025
32. https://www.reddit.com/r/JEENEETards/comments/185qgt4/massive_database_i_have_analysed_where_every/
33. https://www.onlinemanipal.com/blogs/state-of-indian-higher-education
34. https://www.foundit.in/career-advice/career-options-after-12th-list-of-all-courses-after-12th-in-all-streams/
35. https://www.mycareer.matrixedu.in/jee-main-2025-college-predictor/
36. https://ciieducation.in/pdf/ASHE Report 2024 _14-11-2024_v3.pdf