DNA Dataset Documentation

Dataset Overview

The DNA sequence dataset contains information about various genetic sequences and their associated disease predictions, along with relevant parameters and geographical distribution.

Data Structure

- **DNA_Sequence**: The nucleotide sequence string (ATCG)
- **Disease**: Type of disease associated with the sequence
    - Cancer
    - Alzheimer's Disease
    - Cystic Fibrosis
    - Sickle Cell Anemia
- **Parameters**:
    1. DNA Mutation Rate (Parameter1)
    2. Gene Expression Level (Parameter2)
    3. Biomarker Concentration (Parameter3)
- **Location**: Indian city where the sample was collected

Parameter Details

1. **DNA Mutation Rate (Parameter1)**
    - Range: 0.0 - 1.0
    - Disease-specific multiplier: 1.3 for Cancer
    - Indicates genetic instability level
2. **Gene Expression Level (Parameter2)**
    - Range: 0.0 - 1.0
    - Standard multiplier: 1.2
    - Measures transcription activity
3. **Biomarker Concentration (Parameter3)**
    - Range: 0.0 - 1.0
    - Reduced weight: 0.9
    - Indicates disease progression

Disease Severity Mapping

```
const severityMap = {
  "Cancer": 0.9,
  "Alzheimer's Disease": 0.85,
  "Cystic Fibrosis": 0.75,
  "Sickle Cell Anemia": 0.7
}
```

Risk Score Calculation

The risk score for each DNA sequence is calculated using:

```
risk = (param1 * 1.2 + param2 * 0.8 + param3 * 1.5) * baseSeverity / 3.5
```

Data Collection

- Samples collected from major Indian cities
- Each sequence validated using standard DNA sequencing protocols
- Parameters normalized for consistent analysis
- Geographical distribution ensures diverse genetic representation

Usage Guidelines

1. Data should be processed using provided analysis tools
2. Risk scores should be interpreted with clinical context
3. Geographical factors should be considered in analysis
4. Regular updates ensure data accuracy

Technical Specifications

- File format: CSV
- Total records: 1000+ sequences
- Data validation: Automated + Manual review
- Update frequency: Monthly
- Quality metrics: 99.9% accuracy

Data Privacy

- All sequences anonymized
- No personal identifiers included
- Compliant with genetic privacy regulations
- Secure storage and transmission protocols