Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	SWTID1720437019
Project Title	Thyroid Classification
Maximum Marks	2 Marks

Data Collection Plan :

- **Objective:**
  - To collect high-quality and relevant data for the classification of thyroid disorders.
- **Data Sources:**
  - **Healthcare Databases:** Access to large healthcare datasets such as the UCI Machine Learning Repository, National Health and Nutrition Examination Survey (NHANES), or hospital records.
  - **Medical Records:** Patient records from hospitals or clinics with details on thyroid tests and diagnoses.
- **Data Types:**
  - **Demographic Data:** Age, gender, ethnicity.
  - **Clinical Data:** Thyroid function tests (TSH, T3, T4), ultrasound reports, biopsy results.

Raw Data Sources Identification :

Thyroid classification:

- **Dataset Name:** Thyroid Disease Data Set
- **Description:** Contains records of patients with and without thyroid disease, including various test results.
- **Link** -https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-disease-data/code

Data Collection Plan Template

Section	Description
Project Overview	The Thyroid Classification project aims to develop a robust machine learning model to accurately classify various thyroid disorders such as hyperthyroidism, hypothyroidism, and thyroid cancer. By leveraging diverse and meticulously curated data sources, including clinical databases, public health datasets, and electronic health records, the project ensures high data quality and integrity. The ultimate goal is to provide healthcare professionals with a reliable diagnostic tool, enhancing patient outcomes through precise and timely diagnosis. Ethical considerations and data protection compliance are integral to the project's data strategy
Data Collection Plan	**Thyroid Disease Dataset:** • **Description:** A publicly available dataset containing records of patients with various thyroid conditions. • **Data Types:** Numerical (T3, T4, TSH levels), Categorical (gender, diagnosis). • **Access:**https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-disease-data/code
Raw Data Sources Identified	**Thyroid Disease Dataset**

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Dataset 1	This dataset likely includes anonymized medical records of patients diagnosed with thyroid diseases. It typically includes features such as patient demographics (age, gender), clinical measurements (T3, T4, TSH levels), thyroid function test results, and possibly other relevant medical history data	https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-disease-data/code	CSV	732 KB	Public