Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	SWTID1720014456
Project Title	Thyroid Classification
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

The dataset for thyroid classification will be sourced from Kaggle, a platform known for its diverse collection of datasets. This dataset includes comprehensive patient health records and diagnostic data crucial for developing accurate classification models. Data integrity will be rigorously maintained during acquisition from Kaggle, ensuring adherence to ethical guidelines throughout the process. Thorough verification procedures will address missing values and ensure dataset completeness, establishing a robust foundation for the project's predictive modeling efforts.

Data Collection Plan Template

Section	Description
Project Overview	This project aims to develop a machine learning model for early and accurate classification of thyroid conditions, such as hypothyroidism and hyperthyroidism, using clinical and laboratory data.
Data Collection Plan	We are going to collect the dataset from Kaggle website for this project
Raw Data Sources Identified	The dataset for this project was collected from the Kaggle website and includes clinical and laboratory data relevant to thyroid

	conditions, such as T3, T4, and TSH hormone levels, as well as patient demographics and medical history.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Thyroid Disease Data	The datasets featured below were created by reconciling thyroid disease datasets provided by the UCI Machine Learning Repository.	https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-disease-data	CSV	149KB	Public