Lab Assignment 2

Pradeep Kumar Gupta, Keerthana Ajith and Nayantara Varadharajan

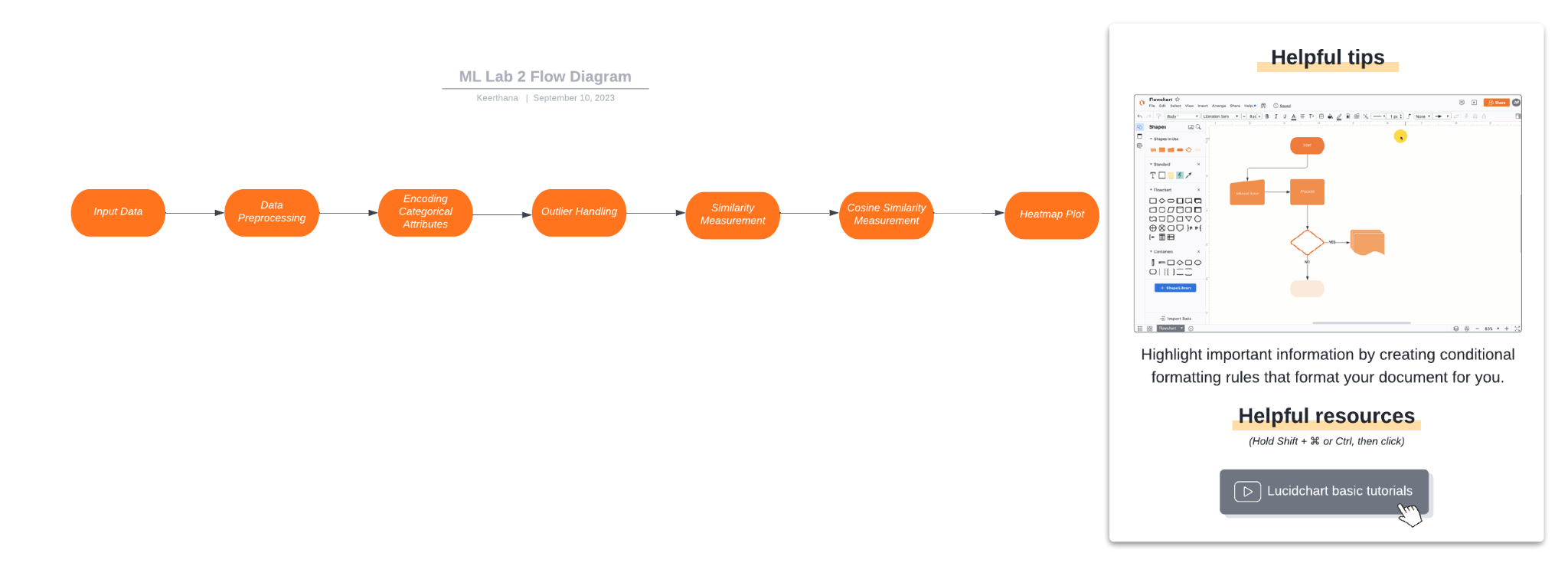
Computer Science Engineering

Amrita School of Engineering

Bangalore, Karnataka, India

1. INTRODUCTION

This project consists of the thorough analysis of a dataset containing information related to thyroid conditions. Data loading and preprocessing are the initial steps, involving encoding of categorical variables and the addressing of missing values. Numeric attributes are normalised for consistency and outliers are detected. The project then probes into calculating the Jaccard Coefficient, Simple Matching Coefficient, and Cosine Similarity to determine how similar the data points are. The associations between data points are visually represented by heatmaps. The importance of this research rests in its prospective applications for the analysis of healthcare data, particularly in the diagnosis of thyroid problems and the enhancement of data quality for medical research and decision making.



Input Data:

Load the dataset from the "thyroid0387\_UCI" worksheet in the "Data2.xlsx" file into a Pandas DataFrame.

1. Data Preprocessing

* Replace '?' and blank spaces with NaN values to clean up the data.
* Remove columns ('TSH', 'T3', 'TT4', 'FTI', and 'TBG') with missing values from rows.
* The remaining columns' empty spaces should be filled up with the corresponding means.
* 'TSH', 'T3', 'TT4', 'FTI', and 'TBG' should be converted to integers.
* Use Min-Max scaling to normalise the numerical attributes "age" and "TT4".

1. Encoding Categorical Attributes

Select categorical attributes, one-hot encode them, and produce binary columns for every category.

1. Outlier Handling

* Using the IQR approach, locate outliers and replace them with median values for numerical attributes.
* Replace values that deviate from the mode with the mode value for categorical attributes.

1. Similarity Measurement

* For binary attributes, determine the Jaccard and simple matching coefficients between the first two observation vectors.
* By comparing their values, decide which coefficient is more suited to measuring similarity.

1. Cosine Similarity Measurement

Implement a comparison of the first two observation vectors' entire vectors using the cosine similarity method.

1. Heatmap Plot

* Select the first 20 observation vectors.
* Determine the Jaccard, simple matching, and cosine similarity coefficients for each pair of vectors.
* To see the commonalities for each coefficient, create heatmaps.

1. METHODOLOGY