CLASSIFICATION OF THYROID CANCER DATA BY REDUCING DIMENSIONALITY

Laila Yasmin, ID: 501345101

Supervisor Name: Ceni Babaoglu

Toronto Metropoliton University

February 4, 2025

Abstract

Thyroid, an important gland located at the base of neck, produces hormones that regulate heart rate, blood pressure, body temperature and weight. Thyroid cancer is a growth of cells that starts in the thyroid, and it is the tenth most common cancer in Canada (Canadian Cancer Society Website, 2024). The prevalence of thyroid cancer in Canada was estimated to 6,600 new cases in 2024 with a rapidly increasing rate than any other cancer in Canada (Canadian Cancer Society Website, 2024).

The thyroid cancer dataset is selected from Kaggle Repository (Kaggle, 2024); the dataset has 212,691 records with 23 features, simulating real-world thyroid cancer risk factors. The risk factors are family history of thyroid cancer, exposed to radiation, iodine deficiency, smoking habit, obesity and diabetes with response as yes or no. The dataset has demographics variables age, race, country and ethnicity to study the distribution of thyroid cancer across geographical regions. In addition, the dataset contains thyroid simulating hormone (TSH), thyroxine (T4), triiodothyronine (T3), nodule size and thyroid cancer risk factor (low, medium or high).

In this project, the goals are: (a) exploratory study of the distribution of thyroid cancer across geographical regions, race, ethnicity and country, (b) classify thyroid cancer as malignant or benign; several machine learning algorithms like Logistic Regression, K-Nearest Neighbour, Random Forest and Support Vector Machine (Hastie et al., 2013) will be used for this classification. The performance of the algorithms on test data will be compared in order to find a classification algorithm with the best prediction accuracy. The goal here is to present a machine learning algorithm with the best prediction accuracy that can be feed into test features and the model classify it as malignant or benign,

(c) The third goal of the project is to explore machine learning algorithms for dimensionality reduction of feature variables. Some techniques such as Principal Component Analysis (PCA), and Forward and Backward feature selection under Logistic Regression (Hastie et al., 2013) will be explored to narrow down few important features. Since the dataset has both the continuous and categorical feature variables; the principal component analysis will be based on the FAMD() function from R-package FactoMineR (R-

Package Repository, 2025). The FAMD function implements the principal component method dedicated to explore data with both continuous and categorical variables.

In the dimensionality reduction, the aim is to explore if a few feature variables can be selected as the dominating feature in classifying thyroid cancer, (d) The fourth goal is to apply the above mentioned machine learning algorithms on the thyroid cancer dataset with the selected features (found from dimensionality reduction) to classify thyroid cancer as malignant or benign, and (e) finally the project aims to explore how effective the dimensionality reduction than using the full set of feature variables in terms of prediction accuracy in the classification of thyroid cancer.

References

Canadian Cancer Society Website. (2024).

Retrieved from https://cancer.ca/en/cancer-information/cancer-types/thyroid/statistics

Hastie et al. (2013). The elements of statistical learning: Data mining, inference and prediction. In The elements of statistical learning (Second ed.). Springer, New York.

Kaggle. (2024).

Retrieved from https://www.kaggle.com/datasets/ankushpanday1/thyroid-cancer-risk-prediction-dataset/data

R-Package Repository. (2025).

Retrieved from https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf