CLASSIFICATION OF THYROID CANCER DATA BY REDUCING DIMENSIONALITY

Laila Yasmin, ID: 501345101

Supervisor Name: Ceni Babaoglu

Toronto Metropoliton University

March 2, 2025

1 Introduction

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, simulating real-world thyroid cancer risk, is selected from Kaggle Repository (Kaggle, 2024). The dataset contains many features such as risk factors, demographic variables and thyroid hormone related variables. The risk factors are family history of thyroid cancer, exposed to radiation, iodine deficiency, smoking habit, obesity and diabetes with a response yes or no. The dataset has demographics variables age, gender, country and ethnicity with thyroid hormone related variables thyroid simulating hormone (TSH), thyroxine (T4) and triiodothyronine (T3). The dataset also contains thyroid nodule size and thyroid cancer risk factor as low, medium or high along with the response of thyroid cancer as benign or malignant.

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. The Principal Component Analysis (Hastie et al., 2013) will be explored to transform the original thyroid cancer dataset to the selected principal component space and then perform the machine learning algorithms on transformed data. 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) or Python library **Prince**. The **FAMD ()** function implements the principal component method dedicated to explore data with both continuous and categorical

variables, and (d) 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.

2 Literature Review

The research goals, in section 1, have two parts - the first part aims is to employ popular classification algorithms to classify thyroid cancer as benign or malignant, and the second part aims to explore principal component analysis algorithm for dimensionality reduction. A comparative analysis of performance based on prediction accuracy in full feature space vs reduced number of features in the transformed space will also be explored thereafter.

According to Sheta et al. (2022), the most popular and commonly used classification algorithms in machine learning are Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes and K-Nearest Neighbours. However, in this project only a subset of the popular algorithms is planned to implement on the thyroid cancer dataset. The article review is limited to the classification algorithms proposed in section 1. A number of relevant articles have been analysed to explore the applicability of the classification algorithms in real-life datasets, with the goal is to gain knowledge about how other authors have implemented the algorithms in classification problems.

Khalilia et al. (2011) presents a method for predicting disease risks from highly imbalanced healthcare data using random forest classification. The authors used the Nationwide Inpatient Sample dataset from the Healthcare Cost and Utilization Project to train Random Forest, Support Vector Machine, Bagging, and Boosting classifiers to predict the risk of chronic diseases. To address the

class imbalance problem in the dataset, the authors employed a repeated random sub-sampling approach, where the training data is divided into balanced sub-samples. The results show that the Random Forest ensemble learning method outperformed the other classifiers in terms of the area under the receiver operating characteristic (ROC) curve. The authors also discussed the application of disease prediction in areas like risk management, health communication, and decision support systems.

Jackins et al. (2021) presents Naive Bayes and Random Forest classification algorithms to diagnose and predict the risk of diabetes, coronary heart disease, and cancer using patient data. The authors compared the performance of the two algorithms on the three disease datasets. The results show that the Random Forest algorithm outperforms the Naive Bayes algorithm in terms of accuracy for all three diseases. The paper also compares the performance of the proposed algorithms with K-means clustering, and the Random Forest algorithm is found to be more effective. The paper concludes that the proposed model works well for both training and test data, and can be used for real-time disease diagnosis. The authors also discussed the limitations of the proposed model; the limitations are: processing time - the model uses a large amount of data to estimate the performance of the training data, and accuracy testing with different datasets - the proposed model needs to be tested with different datasets, beyond the ones used in this study, and the potential to explore other AI algorithms beyond the ones used in the study.

Sruthi et al. (2024) has studied the main purpose of the Random Forest algorithm, that is, to combine the outputs of multiple decision trees to make a single, more accurate prediction. It is an ensemble learning technique that helps overcome the overfitting problem associated with individual decision trees. The key steps involved in the Bagging (Bootstrap Aggregation) technique used in Random Forest are: select a random subset of data points and features to construct each decision tree, train individual decision trees

on these bootstrap samples and combine the outputs of all the decision trees through majority voting for classification or averaging for regression. The author has demonstrated that Random Forest is better at handling overfitting compared to a single decision tree. By combining the predictions of multiple trees, the algorithm reduces the variance and improves the robustness of the model. The diversity of the trees also helps prevent overfitting. As per the article, the real-life applications of the Random Forest algorithm include customer churn prediction, fraud detection, stock price prediction, medical diagnosis and image recognition.

Boateng et al. (2019) provides a comprehensive review of the Logistic Regression model, a widely used statistical technique for modelling the relationship between multiple independent variables and a categorical dependent variable, with a focus on medical research. The review covers the concepts such as odds, odds ratio, logit transformation, logistic curve, assumptions, selecting dependent and independent variables, model fitting, including overall model evaluation, statistical significance of individual predictors, and measures of predictive accuracy and discrimination. Then, the authors highlighted the importance of adhering to recommended guidelines and best practices in the use and reporting of logistic regression, as many studies have been found to have deficiencies in these areas. The authors also present a good example of the application of the logistic regression model using data on factors influencing the decision of expectant mothers to opt for caesarean delivery or vaginal birth.

Robust and sparse logistic regression estimator that addresses the limitations of the standard maximum likelihood estimator for logistic regression has been studied by Cornilly et al. (2024). Their proposed method uses an elastic net penalty to ensure sparsity in the regression coefficients and robustness against outlying observations. The authors show that the influence function of the proposed estimator is bounded, demonstrating its robustness properties. They also evaluate the performance of the estimator in

simulations and an empirical application involving the classification of car fuel types.

Dey et al. (2025) has highlighted the methodological issues regarding the application of Logistic Regression models to complex survey data. The review emphasizes the need for greater emphasis on evaluating how well the logistic regression models fit the data. Many studies have neglected this important aspect, and the review calls for more efforts to raise awareness on proper model evaluation and performance assessment. The review underscores the importance of accounting for complex survey design features, such as: sampling weights, clusters and strata variables. The authors found that while many studies did consider these survey design factors, a significant subset did not, highlighting the need for further investigation into the implications of not properly accounting for the complex survey structure. The main approaches discussed in the review for handling survey design effects in the estimation of binary outcomes or regression coefficients are weighted logistic regression and survey-weighted logistic regression, resampling techniques (e.g., jackknife, balanced repeated replication, bootstrap) and multilevel or mixed-effects logistic regression. Finally, the review highlights that while many studies used these advanced statistical methods to account for complex survey designs, there is still room for improvement in terms of consistent and transparent reporting of the methodological approaches employed.

Syriopoulos et al. (2023) provides a comprehensive review of the k-nearest neighbors (KNN) algorithm, a popular non-parametric classifier. It covers the strengths and weaknesses of KNN, its applications in various data science tasks, available benchmarks and software, and the latest developments in the field. The review aims to serve as a valuable resource for researchers and practitioners to understand and apply KNN effectively. It discusses topics such as anomaly detection, dimensionality reduction, and missing value imputation using KNN. The document also includes a detailed analysis of the KNN algorithm, including its theoretical foundations,

distance metrics, and optimization techniques.

The theoretical foundation of nearest neighbour is covered in the article by Cover and Hart (1967). The authors focused on the analysis of the nearest neighbour pattern classification rule, which assigns an unclassified sample to the category of its nearest neighbour from a set of previously classified samples. The key results of the article are presented as: the nearest neighbour probability of error R is bounded above by 2R*(1-R*), where R* is the Bayes probability of error (the minimum error over all decision rules). This upper bound is shown to be tight. A lower bound on R in terms of R* has also been provided. The paper compares the nearest neighbour rule to the Bayes probability of error R*. It shows that the nearest neighbour rule has a probability of error that is at most twice the Bayes error, indicating that a significant portion of the classification information in the sample set is contained in the nearest neighbour. The authors also analysed the convergence properties of the nearest neighbour, showing that as the sample size increases, the nearest neighbour to a given point converges to that point with probability 1. This convergence result is key to establishing the bounds on the nearest neighbour probability of error. For the multi-category case (more than 2 classes), the paper extends the analysis and provides analogous upper and lower bounds on the nearest neighbour probability of error in terms of the Bayes error rate.

Halder et al. (2024) reviews and analyzes 43 modifications to the k-Nearest Neighbors (KNN) algorithm, focusing on high-dimensional data, and provides a comprehensive overview of KNN search and join methods, including source code and future research directions. This study applies Support Vector Machine (SVM) algorithm to predict MERS-CoV risk distribution in East Java, Indonesia, utilizing 2023 data from the East Java Health Department, achieving highest accuracy of 90% with 75:25 data

proportion.

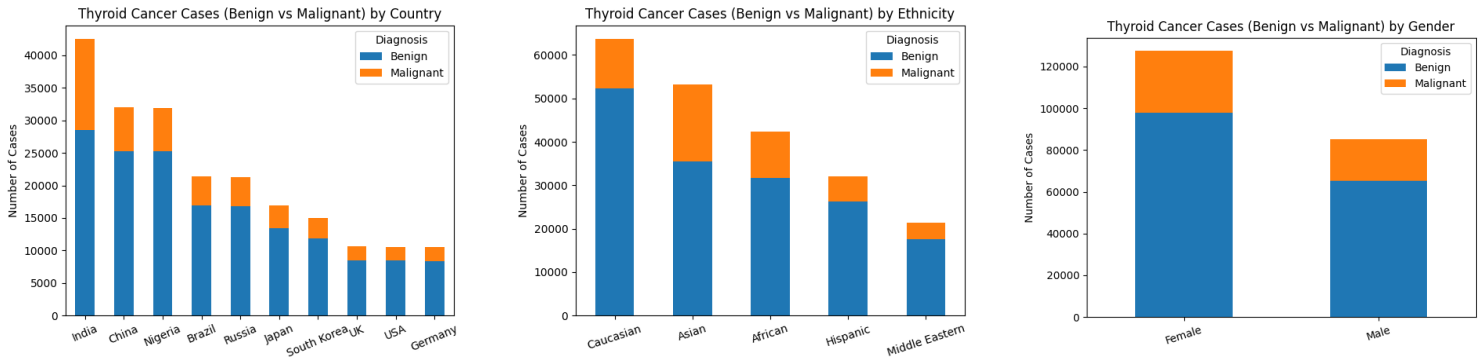
Principal component analysis (PCA) as discussed by Kalantan et al. (2019) is a multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent variables, and its goal is to extract the important information from the table, to represent it as a set of new orthogonal variables called principal components, and display the pattern of similarity of the observations and of the variables as points in maps.

In a summary of the literature review, the aim is now to explore few articles that has comparatively analysed the performance of the above machine learning algorithms. With this in mind, the article by Sheta et al. (2022) is an excellent resource; in this article, the authors have compared the accuracy of four classification algorithms such as Decision Tree, Support Vector Machine, Naive Bayesian, and K-nearest Neighbour on five different datasets. The Naive Bayesian algorithm is proven to be the most effective among other algorithms. Classification of data is crucial for risk management, compliance, and data security and each data mining model has a distinct level of information; the authors have concluded that the success of a model is solely determined by the datasets being used, as there is no such thing as an excellent or a poor model.

3 Data Preprocessing

The thyroid cancer dataset has 17 features with no missing value for any feature and 212,691 observation from different geographical regions such as Russia, Germany, Nigeria, India, UK, South Korea, Brazil, China, Japan and USA. Among the feature variables age, and thyroid hormone related features TSH, T3, T4 and nodule size are continuous variables; all the remaining variables are

categorical with a response yes or no. The distribution of thyroid cancer across the geographical regions and ethnicity is presented in Figure 1. The highest number of malignant thyroid cancer patients are identified in India followed by China and Nigeria. Asian ethnic patients are more identified as having malignant thyroid cancer followed by African ethnic patients than Caucasian. Female patients are more identified as having malignant thyroid cancer than the male patients; this supports the Canadian cancer society reporting that in 2024 that more female was diagnosed with thyroid cancer than male (Canadian Cancer Society Website, 2024).



(a) (b) (c)

Figure 1: Distribution of Thyroid Cancer by: (a) Country (b) Ethnicity (c) Gender

Variable	Category	% Patient
Gender	Male	60
	Female	40
Family History	No	70
	Yes	30
Radiation Exposure	No	85
	Yes	15
Iodine Deficiency	No	75
	Yes	25
Smoking	No	80
	Yes	20
Obesity	No	70
	Yes	30
Diabetes	No	80
	Yes	20

	Age	TSH	T3	T4	Nodule Size
Mean	51.92	5.05	2.00	8.25	2.50
Std	21.63	2.86	0.87	2.16	1.44
Min	15.00	0.10	0.50	4.50	0.00
Q1	33.00	2.57	1.25	6.37	1.25
Median	52.00	5.04	2.00	8.24	2.51
Q3	71.00	7.52	2.75	10.12	3.76
Max	89.00	10.00	3.50	12.00	5.00

Table 1: Left panel presents the demographic characteristic of thyroid cancer patients, and the summary statistics of age and thyroid hormone related variables are shown in right panel.

The demographic characteristics of thyroid cancer patients and the summary of continuous variables are presented in Table 1. The left panel of Table 1 presents the demographic characteristic of thyroid cancer patients, and the summary statistics of age and thyroid hormone related variables are shown in the right panel. The correlation matrix and the heat map (Figure 2) show that the correlation among the variables are negative with a small magnitude except TSH and T3 level; the correlation between TSH and T3 is positive.

	Age	TSH Level	T3 Level	T4 Level	Nodule Size
Age	1.000000	-0.000925	-0.001013	-0.002373	-0.001489
TSH Level	-0.000925	1.000000	0.000335	-0.000795	0.000416
T3 Level	-0.001013	0.000335	1.000000	-0.004069	-0.001799
T4 Level	-0.002373	-0.000795	-0.004069	1.000000	-0.001860
Nodule Size	-0.001489	0.000416	-0.001799	-0.001860	1.000000

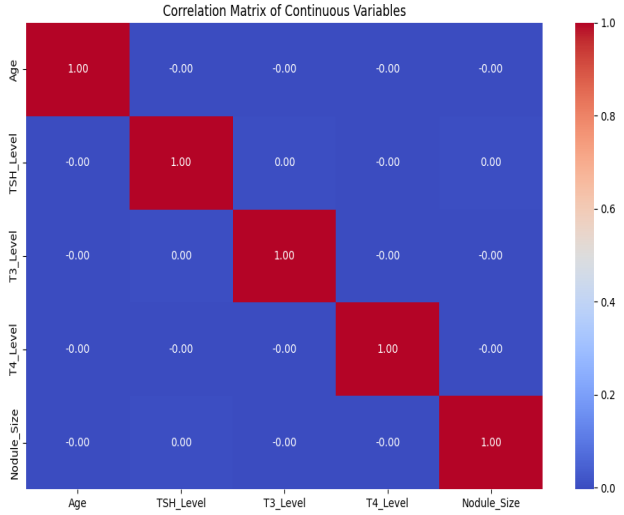


Figure 2: Correlation of continuous variables; the correlation among the variables are negative with a small magnitude except TSH and T3 level. The correlation between TSH and T3 is positive with a small magnitude as well.

References

[1] E.Y. Boateng & D.A. Abaye, (2019), 'A Review of the Logistic Regression Model with Emphasis on Medical Research', Journal of Data Analysis and Information Processing, 7(4)

[2] Canadian Cancer Society Website, (2024), https://cancer.ca/en/cancer-information/cancer-types/thyroid/statistics

[3] D. Cornilly, L. Tubex, S. Van Aelst et al., (2024), 'Robust and sparse logistic regression', Adv Data Anal Classif 18, 663–679

[4] T.M. Cover, & P.E. Hart, (1967). 'Nearest neighbor pattern classification', IEEETransactionsonInformationTheory, 13(1), 21–27.

[5] D. Dey, M.S. Haque, M.M. Islam et al., (2025), 'The proper application of logistic regression model in complex survey data: a systematic review', BMC Med Res Methodol, 25 (15), https://doi.org/10.1186/s12874-024-02454-5

[6] R.K. Halder, M. N. Uddin , A. Uddin , S. Aryal & A. Khraisat, (2024),'Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications', Journal of Big Data, 11:113

[7] T. Hastie, J.H. Friedman & R.Tibshirani, (2013), 'The elements of statistical learning: Data mining, inference and prediction', In The elements of statistical learning (Second ed.), Springer, New York

[8] V. Jackins, S. Vimal, M. Kaliappan et al., (2021), 'AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes', Journal of Supercomputing 77, 5198–5219, https://doi.org/10.1007/s11227-020-03481-x

[9] Kaggle Website, (2024), https://www.kaggle.com/datasets/ankushpanday1/thyroid-cancer-risk-prediction-dataset/data

[10] M. Khalilia, S. Chakraborty & M. Popescu, (2011), 'Predicting disease risks from highly imbalanced data using random forest', BMC Medical Informatics and Decision Making 11 (51), https://doi.org/10.1186/1472-6947-11-51

[11] R-package Repository, (2025), https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf

[12] V. Sheta, U. Tripathi, A. Sharma, (2022), 'A Comparative Analysis of Machine Learning Algorithms for Classification Purpose', Procedia Computer Science, 422-431

[13] Sruthi, (2024), 'Understanding Random Forest Algorithm With Examples', https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[14] P.K. Syriopoulos, N.G. Kalampalikis, S.B. Kotsiantis et al. (2023), 'KNN Classification: a review', Ann Math Artif Intell, https://doi.org/10.1007/s10472-023-09882-x