

CLASSIFICATION OF THYROID CANCER DATA BY REDUCING DIMENSIONALITY

Laila Yasmin

Supervisor: Ceni Babaoglu
Toronto Metropolitan University

Big Data Project



- Introduction
- Research Goal
- Data Collection
- Distribution of Thyroid Cancer
- Summary Statistics
- Correlation
- Dataset Scaling
- Train and Test Datasets
- K-Nearest Neighbor
- Logistics Regression
- Decision Tree
- Random Forest

- ① Thyroid, an important gland located at the base of neck, produces hormones that regulate heart rate, blood pressure, body temperature and weight.
- ② Thyroid cancer is a growth of cells that starts in the thyroid, and it is the tenth most common cancer in Canada.
- ③ 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.

- ❶ Exploratory study of the distribution of thyroid cancer across geographical regions, race and ethnicity.
- ❷ Predictive analysis to classify thyroid cancer as malignant or benign.
- ❸ Performance comparison of several machine learning algorithms (e.g, Logistic Regression, K-Nearest Neighbour, Random Forest and Decision Tree)
- ❹ Explore Principal component analysis for dimensionality reduction of feature variables.

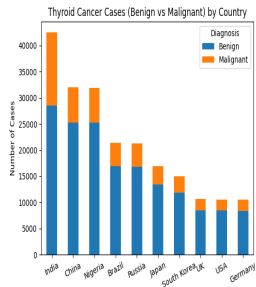
Data Collection and Data Features

- ① Thyroid cancer dataset, simulating real-world thyroid cancer risk, is selected from Kaggle Repository.
- ② The dataset contains many features such as risk factors, demographic variables and thyroid hormone related variables.
- ③ Risk factors are family history of thyroid cancer, exposed to radiation, iodine deficiency, smoking habit, obesity and diabetes with a response yes or no.
- ④ Demographics variables age, gender, country and ethnicity; thyroid hormone related variables are TSH, T3 and T4.
- ⑤ Thyroid cancer risk factor: low, medium or high, and thyroid cancer response: benign or malignant.

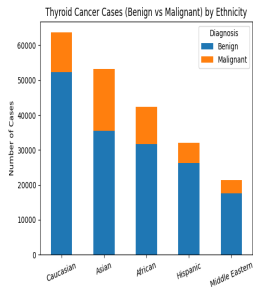
Distribution of Thyroid Cancer

- ❶ Thyroid cancer dataset has 17 features with no missing values for any feature.
- ❷ There are 212,691 observations from different geographical regions such as Russia, Germany, Nigeria, India, UK, South Korea, Brazil, China, Japan and USA.
- ❸ India, China, and Nigeria show higher rates of malignant thyroid cancer, while Asian ethnicity demonstrates a higher prevalence compared to other ethnic groups.
- ❹ Females also experience a higher incidence of malignant thyroid cancer, aligning with existing reports from organizations like the Canadian Cancer Society.

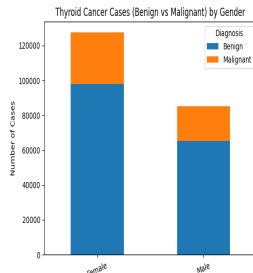
Distribution of Thyroid Cancer



(a)



(b)



(c)

Figure 0.1: Distribution of Thyroid Cancer by (a) Country (b) Ethnicity, and (c) Gender.

Summary Statistics

- ➊ Dataset has 60% female and 40% male patients.
- ➋ The percentages of patients to smoking and diabetes as yes vs. no are 20% vs. 80% for both attributes.
- ➌ There are 30% patients without obesity whereas obese patients are 70% in total.
- ➍ The minimum age of patients in the dataset is 15 years, and maximum is 89 years with a mean 51.92 years and std 21.63 years.
- ➎ The mean nodule size is 2.50 millimetres with std 1.44 millimetres, and the range of nodule size is from 0 to 5 millimetres.

Summary Statistics

| | 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 0.1: Summary statistics of continuous attributes.

Correlation of Continuous Attributes

- ❶ A weak negative correlation (-0.000925) exists between age and TSH levels. This suggests that as age increases, TSH levels may tend to slightly decrease as well.
- ❷ Weak negative correlation (-0.000795) is observed between TSH and T4 levels, indicating as TSH levels increase, T4 levels tend to decrease, and vice versa.
- ❸ A weak negative (-0.004069) correlation is found between T3 and T4 levels.
- ❹ The nodule size shows a weak positive correlation with TSH (0.000416), indicating a slight tendency for larger nodules to be associated with slightly higher levels of this hormone.

Dataset Scaling of Continuous Attributes

- ❶ The continuous attributes have different scales or units; for example, age has unit years while thyroid related hormones have unit milliliter (ml), and the nodule size has a unit of millimetre (mm).
- ❷ Before we fit any machine learning algorithm, it is important to scale the attributes to have the same scale across all the attributes.
- ❸ The MinMax scaling function is chosen from Python library **sklearn.preprocessing**, which scales the continuous variables to have mean 0 with standard deviation 1.

Train and Test Datasets

- ➊ First using a random state for reproducibility, the thyroid cancer dataset is divided as train and test data.
- ➋ 80% of the data is kept a train data on which the machine learning algorithms will be fitted, and the rest of the 20% data is kept as test data for the evaluation of model performance and accuracy of prediction.
- ➌ The train dataset has 170,152 observation while the test dataset has 42,539 observations.
- ➍ As some of the attributes are categorical, this categorical attributes are converted as factor attributes, giving a total of 28 attributes in the dataset.

K-Nearest Neighbor

- ❶ The optimal value of K (i.e., number of nearest neighbors) is determined using **GridSearchCV**; K=5 is the optimal.
- ❷ The performance of the KNN model is evaluated on a test dataset, which is 20% of the original dataset.
- ❸ KNN classifier achieved an accuracy of 72.14% on the test dataset.
- ❹ The model showed reasonable performance in identifying benign thyroid cancer patients with a precision of 77%.
- ❺ In contrast, its performance is less satisfactory for identifying malignant cases, with a precision of only 24%.

Logistics Regression

- 1 The logistic regression classifier has achieved approximately 82.50% accuracy on the test dataset.
- 2 Logistic regression has achieved a precision of 85% for benign cases and 69% for malignant cases.
- 3 The model shows a higher precision in identifying benign cases compared to malignant cases.

| | Predicted Benign | Predicted Malignant |
|------------------|------------------|---------------------|
| Actual Benign | 30657 | 1958 |
| Actual Malignant | 5484 | 4440 |

Table 0.2: Confusion Matrix for Logistic Regression

Decision Tree

- ❶ Decision Tree model, utilizing entropy and a maximum depth of 3 for pre-pruning, achieved an accuracy of 82.50% on the test dataset.
- ❷ The precision is 85% for benign patients and 69% for malignant patients.
- ❸ Sensitivity of decision tree is 44.74%. Specificity is 94.00% with false positive rate of 6.00% and false negative rate of 55.26%.

| Metric for Decision Tree | Percentage |
|----------------------------------|------------|
| Recall | 44.74% |
| Precision | 69.40% |
| True Positive Rate (Sensitivity) | 44.74% |
| True Negative Rate (Specificity) | 94.00% |
| False Positive Rate | 6.00% |
| False Negative Rate | 55.26% |

Table 0.3: Evaluation Metrics for Decision Tree

- ➊ Random Forest model achieved an accuracy of 82.46% on the test data, indicating that the model correctly predicted the diagnosis (benign or malignant) for approximately 82.46% of the patients in the test dataset.
- ➋ The model has obtained a precision of 85% for benign cases and 70% for malignant cases.
- ➌ This means that out of all the patients predicted as benign, approximately 85% were truly benign. Similarly, out of all the patients predicted as malignant, approximately 70% were truly malignant.

Principal Component Analysis

- ❶ The `FAMD()` (Python's `Prince` library) function implements the principal component method dedicated to explore data with both continuous and categorical variables.
- ❷ Two eigenvectors (principal components) that explains the majority of variances are found.
- ❸ Train and test data are transformed onto the principal component space. Then in transformed space, the machine learning algorithms are fitted to evaluate model accuracy on test data.

Model Performance Comparison

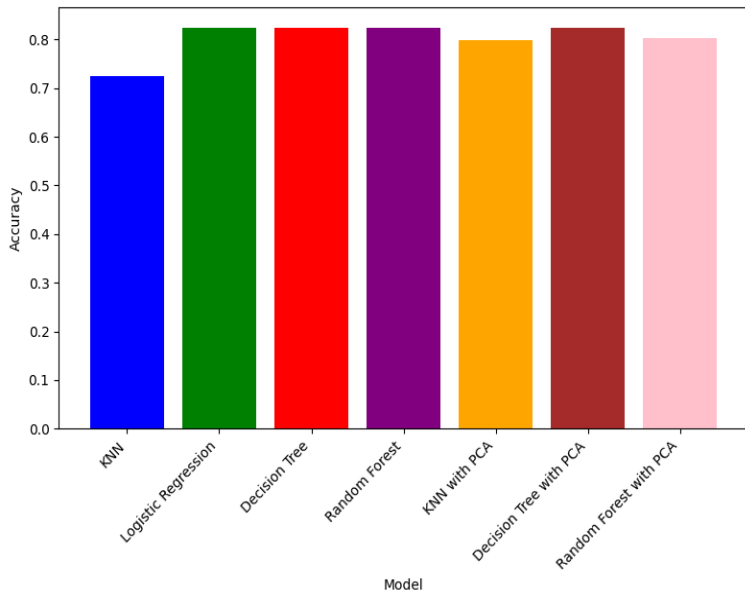
- ❶ KNN classifier has the lowest prediction accuracy among all the machine learning algorithms that have been used in this analysis.
- ❷ Logistic regression, Decision Tree and Random Forest have almost the same prediction accuracy on test data but these prediction accuracies decrease on PCA transformed data except KNN algorithm.
- ❸ KNN classifier performs better prediction accuracy on PCA transformed dataset than the original dataset.

Model Performance Comparison

| Model | Accuracy |
|------------------------|----------|
| KNN | 72.14% |
| Logistic Regression | 82.50% |
| Decision Tree | 82.50% |
| Random Forest | 82.46% |
| KNN with PCA | 79.80% |
| Decision Tree with PCA | 82.50% |
| Random Forest with PCA | 80.30% |






Table 0.4: Model performance comparison of different machine learning algorithms.

Model Performance Comparison








- ➊ Logistic regression, Decision Tree, and Random Forest demonstrated generally good performance on predicting thyroid cancer diagnoses, with accuracy exceeding 82%.
- ➋ The KNN initially suffered from lower accuracy, but PCA helped it perform better.
- ➌ The effect of PCA on model performance varied, highlighting the importance of considering different techniques and their potential impact on model results.
- ➍ The choice of a "best" model might depend on specific needs, such as the trade-off between complexity and accuracy, as well as the relative importance of correctly identifying benign and malignant cases.

References

-  E.Y. Boateng et al. (2019), 'A Review of the Logistic Regression Model with Emphasis on Medical Research', Journal of Data Analysis and Information Processing, 7(4)
-  Canadian Cancer Society Website, (2024), <https://cancer.ca/en/cancer-information/cancer-types/thyroid/statistics>
-  T.M. Cover et al. (1967). 'Nearest neighbor pattern classification', IEEE Transactions on Information Theory, 13(1), 21–27.
-  R.K. Halder et al., (2024), 'Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications', Journal of Big Data, 11:113
-  T. Hastie et al. (2013), 'The elements of statistical learning: Data mining, inference and prediction', In The elements of statistical learning, Springer, New York

References

-  V. Jackins et al. (2021), 'AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes', Journal of Supercomputing 77, 5198–5219
-  Kaggle Website (2024),
<https://www.kaggle.com/datasets/ankushpanday1/thyroid-cancer-risk-prediction-dataset/data>
-  M. Khalilia et al. (2011), 'Predicting disease risks from highly imbalanced data using random forest', BMC Medical Informatics and Decision Making 11 (51)
-  Sruthi, (2024), 'Understanding Random Forest Algorithm With Examples', <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>
-  P.K. Syriopoulos et al. (2023), 'KNN Classification: a review', Ann Math Artif Intell