Comparative Analysis of Machine Learning Models for Thyroid Cancer Recurrence Prediction

Anay Aggarwal, Ekam Kaur, and Susie Lu

(Mentor: Dr. Marly Gotti)

October 12, 2024

MIT PRIMES October Conference

Outline

Introduction

Understanding Data

Models

Comparative Analysis

Improving Models via Feature Selection

Introduction

Introduction

- Cancer ranks as one of the top causes of death globally.
- Thyroid cancer is among the most prevalent endocrine malignancies worldwide, with differentiated thyroid cancer (DTC) representing the majority of cases.
- Traditional approaches often lack the precision for personalized treatment planning, so there has been growing interest in leveraging machine learning instead (Xi et al, 2022; Bhattacharya et al, 2023; Borzooei et al, 2024).

Machine Learning Algorithms Analyzed

We aim to compare the performance of six machine learning models for predicting DTC recurrence.

Algorithm	A Key Advantage
Support Vector Machines (SVM)	Effective for high-dimensional data
Random Forests (RF)	Ensemble method via bagging
Extreme Gradient Boosting (XGBoost)	Ensemble method via boosting
Artificial Neural Networks (ANN)	Models complex non-linear relationships
K-Nearest Neighbors (KNN)	Efficient, based on space proximity
Logistic Regression (LR)	High interpretability, outputs probabilities

Details: `https://marlycormar.github.io/primes-research-project-2024/`

Reproducibility

- R programming language
- `tidymodels` ecosystem and `workflows` package
- Quarto manuscript structure and `renv` environment

Overview of Dataset

- Differentiated Thyroid Cancer Recurrence dataset (383 patients) from the UCI Machine Learning Repository.
- **16 predictors**
  - **One numerical predictor:** Age
  - **15 categorical predictors:**
    - Gender, Smoking, History of smoking, History of radiotherapy
    - Thyroid function, Physical examination, Adenopathy, Pathology, Focality
    - Risk assessment, Cancer stage, T, N, M, Initial treatment response
- **One outcome variable:** whether DTC recurred.

Understanding Data

Exploratory Data Analysis

- After removing duplicates, the dataset has 364 observations.
- Figure 1: 80.5% of the patients are female, while 19.5% are male. Males are more likely to have DTC recurrence.
- Figure 2: In general, older patients are more likely to have DTC recurrence.

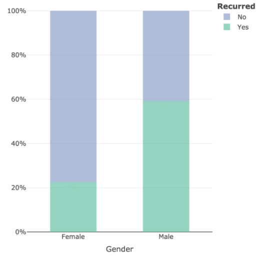


Figure 1: Gender Distribution by Cancer Recurrence

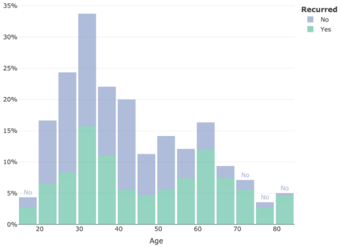
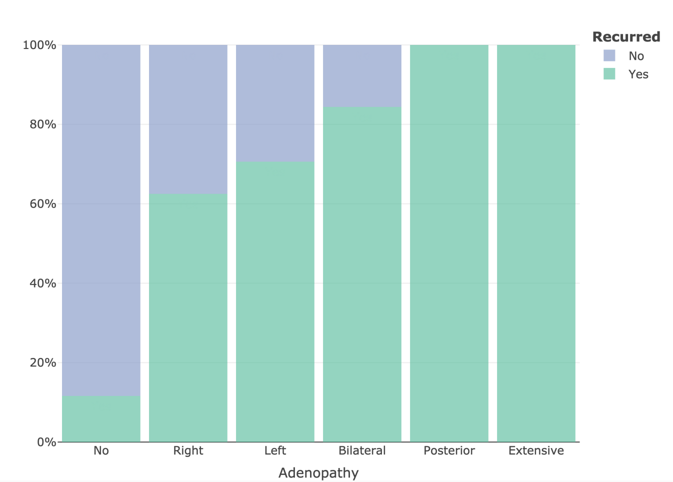


Figure 2: Age Distribution by Cancer Recurrence

Exploratory Data Analysis

- Besides age, the rest of the features are categorical.
- Adenopathy: the presence of swollen lymph nodes during physical examination.



Models

Model Training & Testing

What is the process of creating and testing a machine learning model?

- Training Set: Portion of the dataset used for fitting the model. We use the training set (the predictors **and** response variable) to determine a model by minimizing some error function.
- Test Set: Portion of the dataset used for computing the model's performance.

The goal is to train a model that is not too specific to the training set while maintaining high accuracy.

- We also use k-fold cross-validation in which we split the training data into k parts and sequentially fit the model on $k-1$ parts, leaving the last part as a testing set.

Metrics for Model Performance

	Truth: Positive	Truth: Negative
Prediction: Positive	TP	FP
Prediction: Negative	FN	TN

- **Accuracy**: proportion of correct predictions, or $\frac{TN+TP}{TN+TP+FN+FP}$.
- **Precision**: proportion of positive classified observations that are actually positive, or $\frac{TP}{TP+FP}$.
- **Recall**: proportion of actual positive observations correctly classified as positive, or $\frac{TP}{TP+FN}$.
- **Specificity**: proportion of actual negative observations correctly classified as negative, or $\frac{TN}{TN+FP}$.

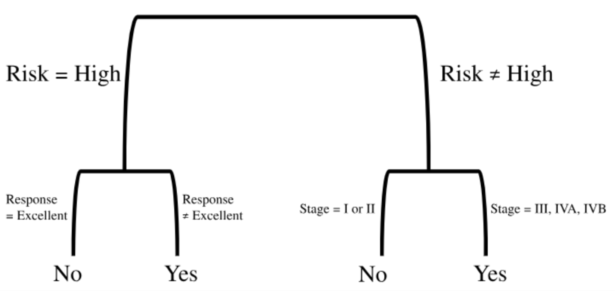
Random Forest (RF)

Random Forest is an ensemble learning method that

- constructs multiple decision trees
- outputs the mode of the classifications given by the individual trees

Each decision tree

- uses recursive binary splits
- minimizes an error criterion, e.g. Gini index



Random Forest (RF): Reduce Variance

Intuition: Given n independent observations, each with variance σ^2, the variance of their average is σ^2/n.

Bagging

- Train multiple trees on different samples of the data.
- The final prediction is the mode of the predictions of all trees.

Random Forest (RF): Reduce Variance

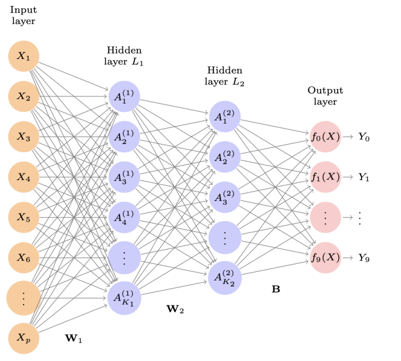
Intuition: Given n independent observations, each with variance σ^2, the variance of their average is σ^2/n.

Random selection of features for each split

- Approximately $\sqrt{p}$ features are used, where p is the total number of features.
- Reduces the correlation between the predictions of different trees.

Artificial Neural Network (ANN)

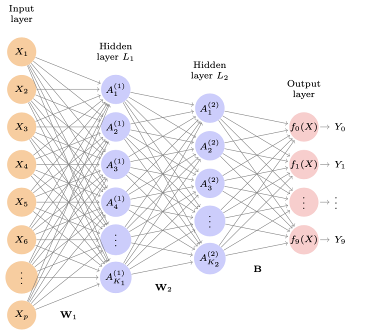
- ANNs are computational models inspired by the brain, composed of interconnected nodes (neurons) in layers.
- An ANN is composed of multiple layers, including an input layer, one or more hidden layers, and an output layer.



Source: Gareth James et al., 2021

Artificial Neural Network (ANN)

- The input layer receives the raw data, the hidden layers process the data, and the output layer produces the final prediction.
- Each connection between neurons has an associated weight, and each neuron has a bias term. These parameters are optimized during training.



Source: Gareth James et al., 2021

Logistic Regression (LR)

- Supervised learning algorithm widely used for classification problems.
- For binary classification, the algorithm estimates the **log odds** of each observation via a linear function.

Odds and Log Odds

The odds of an event X denote the probability of "success" to that of "failure" – that is, the quantity $\frac{p(X)}{1-p(X)}$.

The log odds is the quantity $\log \frac{p(X)}{1-p(X)}$.

LR Model

For each point X with predictors $X_1, \ldots, X_p$, the algorithm fits the equation

$$\text{log odds}_X = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

where the $\beta_0, \beta_1, \ldots, \beta_p$ are parameters to be estimated.

Logistic Regression (LR)

The model is fitted by maximizing the probability of the observed data.

LR Parameter Estimation

For a given training set of points $x_1, \ldots, x_n$ with response variables $y_1, \ldots, y_n$, we are maximizing the likelihood function

$$L(\beta) = \prod_{y_i=1} P(x_i) \prod_{y_i=0} (1 - P(x_i)).$$

We can write this as a single product

$$L(\beta) = \prod_{i=1}^{n} P(x_1)^{y_i} \cdot (1 - P(x_i))^{(1-y_i)}.$$

This is equivalent to maximizing the log-likelihood function

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log P(x_i) + (1 - y_i) \log(1 - P(x_i))].$$

Comparative Analysis

All Models' Metrics

	SVM	RF	XGBoost	ANN	KNN	LR
Accuracy	78.0%	94.5%	91.2%	92.3%	90.1%	93.4%
Precision	23.1%	84.6%	73.1%	84.6%	76.9%	84.6%
Recall	100.0%	95.7%	95.0%	88.0%	87.0%	91.7%
Specificity	76.5%	94.1%	90.1%	93.9%	91.2%	94.0%

Best Models

- RF is the most robust and balanced classifier for predicting DTC recurrence.
- **RF achieved the highest accuracy and specificity rates**, both at 94%, demonstrating its reliability in correctly identifying both positive and negative cases.
- **ANN, LR, and RF all achieve** 85% **precision**, so they are equally competent at correctly predicting positive cases.

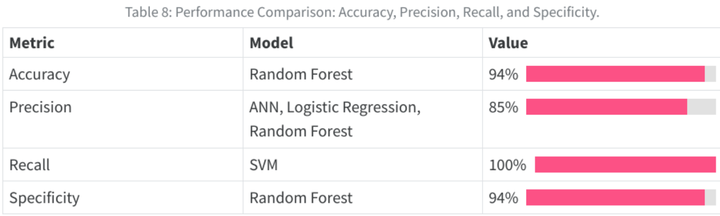
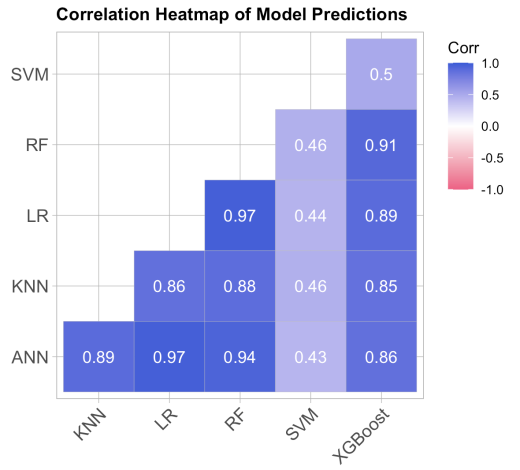


Table 8: Performance Comparison: Accuracy, Precision, Recall, and Specificity.

Metric	Model	Value
Accuracy	Random Forest	94%
Precision	ANN, Logistic Regression, Random Forest	85%
Recall	SVM	100%
Specificity	Random Forest	94%

Correlation Matrix of Model Predictions

Here is a correlation heatmap pairwise comparing the predictions.



Correlation Matrix of Model Predictions

- The pairwise correlations between the non-SVM models are all **at least** 0.85, while those between SVM and the other models are all **at most** 0.5.
- In fact, SVM has a really high false positive rate.
- There is not a single test case in which SVM predicts negative, but the other model predicts positive.
- For each model, there are **between 14 and 20 test cases** (among the 91 total) for which SVM predicts positive but the other model predicts negative.

Bayesian Model Comparison

- Bayesian Model Comparison leverages the resampling results obtained during model tuning to approximate model performance.
- While this is not the same as test performance, we expect it to be a reasonable approximation.

Choose some base model and specify its metric β_0. A **standard ANOVA model** predicts how different each model is in that metric via the equation

$$y = \beta_0 + \beta_1 m_1 + \cdots + \beta_k m_k,$$

where the y denotes the metric and the m_j serve as indicator variables for the rest of the models.

- **In the Bayesian case**, each of the parameters β_i represent a distribution, rather than a single value.

Bayesian Model Comparison Results

Model Comparison: Specify a difference in metrics we consider negligible – usually 0.02 – and look at the difference distributions of the model metrics. If a large proportion of area lies within this region, then the models' performances in the metric are not practically different.

Results: RF had the best distributions in all the metrics tested, though the metric distributions did lie in the practical equivalence region for LR and ANN, with high probability.

Improving Models via Feature Selection

Why Feature Selection?

- Dimensionality reduction
- Identification of critical predictors of DTC recurrence

Factor Analysis for Mixed Data (FAMD)

- **Approach:** Find a low-dimensional representation of the data that captures most of the variance.
- Numerical predictors $p_1, \dots, p_P$
- Categorical predictors $q_1, \dots, q_Q$
- **First principal component:** linear combination Z_1 of predictors with maximal

  $$\sum_{i=1}^{P} \underbrace{r^2(Z_1, p_i)}_{\text{Correlation coef.}} + \sum_{i=1}^{Q} \underbrace{\eta^2(Z_1, q_i)}_{\text{Correlation ratio}} .$$

  The vector of coefficients of predictors in Z_1 is denoted by ϕ_1.
- **Second principal component:** linear combination Z_2 such that ϕ_2 is orthogonal to ϕ_1.

Results from Factor Analysis for Mixed Data

Top three contributors to the first three principal components

PC1	**PC2**	**PC3**
Risk (12.7%)	Risk (18.7%)	T (33.8%)
T (11.1%)	T (17.1%)	Pathology (27.7%)
Response (10.8%)	Stage (11.3%)	Physical Examination (7.7%)

Results from Factor Analysis for Mixed Data

Top three contributors to the first three principal components

PC1	PC2	PC3
Risk (12.7%)	Risk (18.7%)	T (33.8%)
T (11.1%)	T (17.1%)	Pathology (27.7%)
Response (10.8%)	Stage (11.3%)	Physical Examination (7.7%)

Feature Selection via FIA

It is important to understand **why** the models produced the predictions that it did. This can be done with a process called feature importance analysis (FIA). FIA is also useful for

- **Model Improvement**: By identifying the most impactful features, FIA helps one focus efforts on the data that truly matters.
- **Overfitting Detection**: Features with surprisingly high importance might indicate overfitting. FIA helps you identify such features and potentially adjust the model to reduce overfitting.

We can perform two types of FIA: **Impurity Importance** and **Permutation Importance**.

FIA: Impurity Importance

Method: Impurity importance measures the importance of a feature based on the total reduction of the criterion (impurity) brought by that feature. Features with higher impurity reduction are considered more important.

Advantages: It is relatively fast compared to other feature importance methods and provides a global view of feature importance across the entire model.

Limitations: Impurity importance can be biased towards numerical features or those with many categories, and is sensitive to overfitting.

FIA: Permutation Importance

Method: Permutation importance measures the importance of a feature by fixing a feature and shuffling the other features. The change in a performance metric (accuracy here) gives information about the influence of the given feature.

Advantages: It directly measures the impact on model accuracy.

Limitations: Permutation importance can overestimate the importance of correlated features. Additionally, it can be computationally expensive, especially for large datasets or complex models.

Combined Results of FAMD and FIA

Important features: Risk, T, Response, Pathology, Thyroid Function, Adenopathy, Age

Improved model performance: Most pronounced for KNN.

Metric	Original Value	New Value
Accuracy	90.1%	94.5%
Precision	76.9%	84.6%
Recall	87.0%	95.7%
Specificity	91.2%	94.1%

Acknowledgements

- We would like to thank our PRIMES mentor, Dr. Marly Gotti, for introducing us to this fascinating research area and guiding us throughout the project.
- We are also grateful to the PRIMES organizers for making this program possible.

References

[1] Shiva Borzooei and Aidin Tarokhian. **Differentiated Thyroid Cancer Recurrence.** 2023. DOI: 10.24432/C5632J.

[2] Gareth James et al. **An Introduction to Statistical Learning.** Springer Texts in Statistics, 2021.

[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. **“Why Should I Trust You?”: Explaining the Predictions of Any Classifier.** New York, NY, USA, 2016. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

[4] Nan Miles Xi, Lin Wang, and Chuanjia Yang. **Improving the diagnosis of thyroid cancer by machine learning and clinical data.** 2022. DOI: 10.1038/s41598-022-15342-z.

Thank you!