IT1244 Team 21
Cancer Detection Dataset

Jin Jiarui, Lai Zheyuan, Tan Hui En, Zhuang Yiling

{e1304468, e1297740, e0559373, e1090523}@u.nus.edu

Introduction - Classification Task for Breast Cancer Dataset

The algorithm that we adopt needs to perform well in:

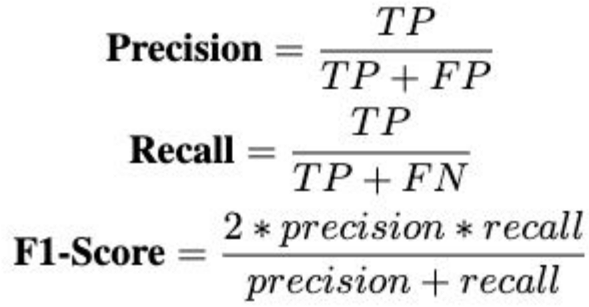
- Healthy vs Cancer
- Healthy vs Screening Stage Cancer
- Healthy vs Early Stage Cancer

The dataset is **highly biased**, so we applied the following techniques:

- **Resampling**
- Adjust the **class_weight** parameter

Appropriate Metrics:

Precision, Recall, and F1-Score:



$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F1-Score} = \frac{2 * precision * recall}{precision + recall}$$

ROC Curve and **ROC-AUC Score**:

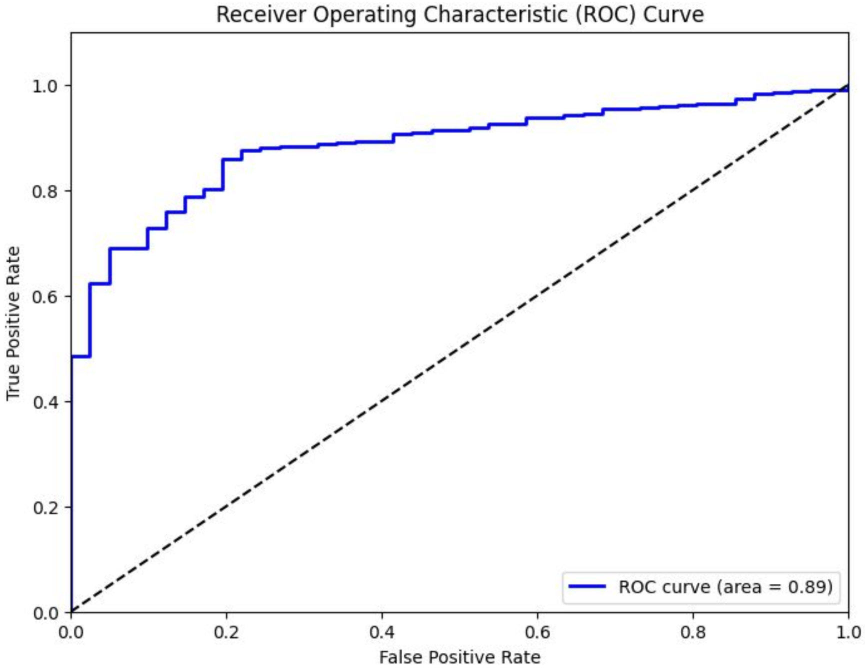
- The Receiver Operating Characteristic (ROC) Curve illustrates the tradeoff between sensitivity and specificity at various thresholds.
- The ROC-AUC (Area Under the Curve) Score provides a single, aggregate measure of performance across all classification thresholds.
- Works well for imbalanced datasets.

Logistic Regression

- **Handling Class Imbalance:**
  - Since the dataset is imbalanced with unequal class distributions (more cancer instances compared to healthy instances), class_weight='balanced' parameter is used in logistic regression to address this issue.
  - Class imbalance occurs when one class (e.g., cancer) has significantly more instances than another class (e.g., healthy) in the dataset.
  - This can lead to biases in the model's learning process, where the model may prioritize accuracy on the majority class at the expense of the minority class.
  - {class_weight='balanced'} is used so the logistic regression algorithm adjusts the weights assigned to each class during training.
  - It assigns higher weights to minority classes and lower weights to majority classes, effectively giving more importance to correctly classifying instances from the minority class.
- **Feature Scaling:**
  - Features are standardized using StandardScaler to ensure that each feature contributes equally to the model fitting process.
- **Model Training:**
  - Logistic regression model is trained using the training data (X_train_scaled and y_train) with class_weight='balanced'.

Logistic Regression Evaluation

- Model evaluation is performed on the test data (X_test_scaled and y_test) using appropriate evaluation metrics.
- The ROC-AUC score is computed to evaluate the overall performance of the logistic regression model.
- Classification reports are generated, which include metrics such as precision, recall, and F1-score for each class (next slide), providing a more detailed understanding of the model's performance on individual classes.



Logistic Regression Result

- To perform more specific classification tasks within the cancer class, the code filters the dataset to include only instances labeled as 'screening stage cancer' or 'early stage cancer' along with healthy instances.
- Logistic regression is applied similarly to classify screening stage cancer vs. healthy and early stage cancer vs. healthy, allowing the model to focus specifically on distinguishing between these subgroups.

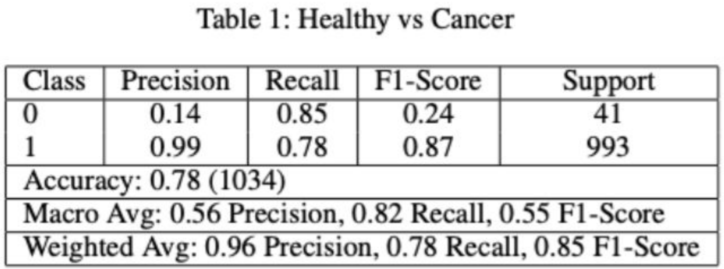


Table 1: Healthy vs Cancer

Class	Precision	Recall	F1-Score	Support
0	0.14	0.85	0.24	41
1	0.99	0.78	0.87	993
Accuracy: 0.78 (1034)				
Macro Avg: 0.56 Precision, 0.82 Recall, 0.55 F1-Score				
Weighted Avg: 0.96 Precision, 0.78 Recall, 0.85 F1-Score				

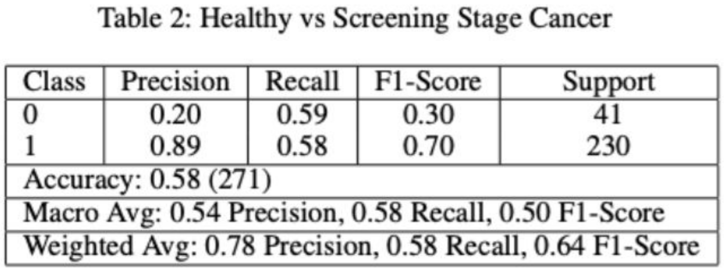


Table 2: Healthy vs Screening Stage Cancer

Class	Precision	Recall	F1-Score	Support
0	0.20	0.59	0.30	41
1	0.89	0.58	0.70	230
Accuracy: 0.58 (271)				
Macro Avg: 0.54 Precision, 0.58 Recall, 0.50 F1-Score				
Weighted Avg: 0.78 Precision, 0.58 Recall, 0.64 F1-Score				

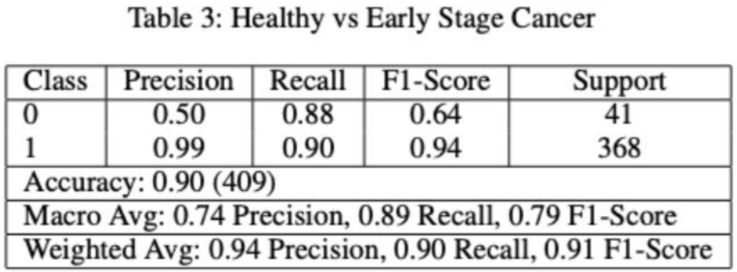


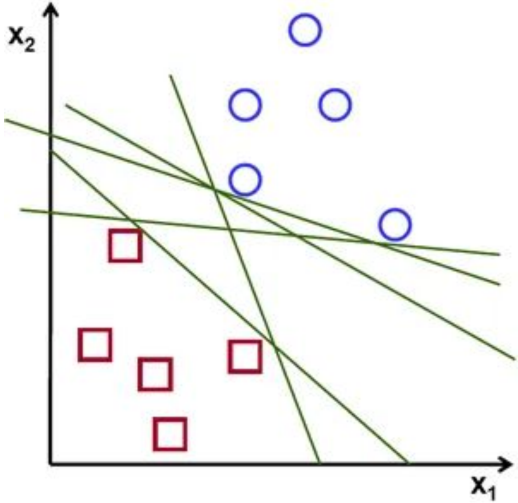
Table 3: Healthy vs Early Stage Cancer

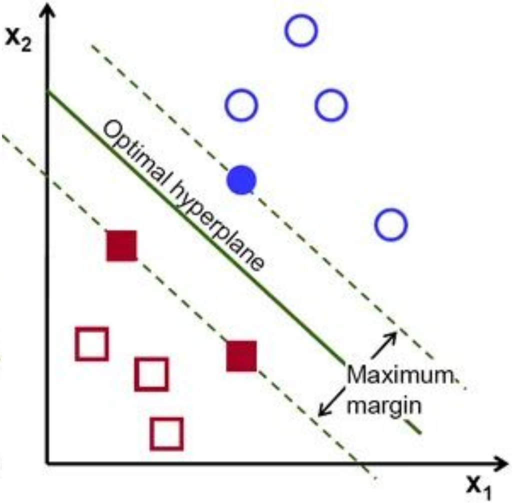
Class	Precision	Recall	F1-Score	Support
0	0.50	0.88	0.64	41
1	0.99	0.90	0.94	368
Accuracy: 0.90 (409)				
Macro Avg: 0.74 Precision, 0.89 Recall, 0.79 F1-Score				
Weighted Avg: 0.94 Precision, 0.90 Recall, 0.91 F1-Score				

Support Vector Machine

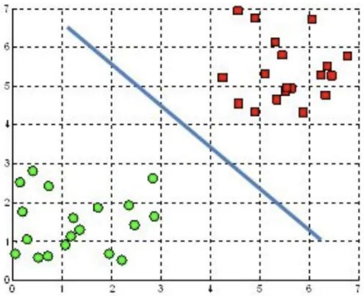
In classification tasks, Support Vector Machine (SVM) algorithm aims to find the optimal **hyperplane** that best separates the data into its classes.

In two-dimensional space, this hyperplane is a line dividing a plane in two parts where in each class lay on either side.

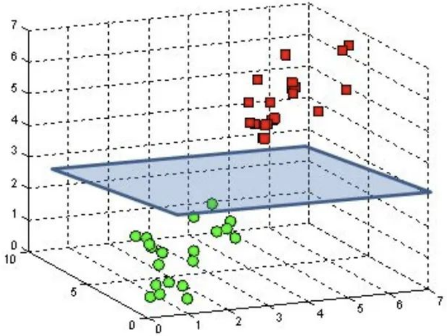




A hyperplane in $\mathbb{R}^2$ is a line



A hyperplane in $\mathbb{R}^3$ is a plane



Mathematical Optimization of SVM Algorithm (Basic)

This **hyperplane** can be written as $\mathbf{w}^\mathsf{T}\mathbf{x} - b = 0$,

For our nonlinear case, the aim of this optimization is to minimize:

$$\|\mathbf{w}\|^2 + C\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i - b)\right)\right],$$

where $\max\left(0, 1 - y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i - b)\right)$ is defined as **hinge loss** function.

By deconstructing the hinge loss, the optimization goal can be expressed as:

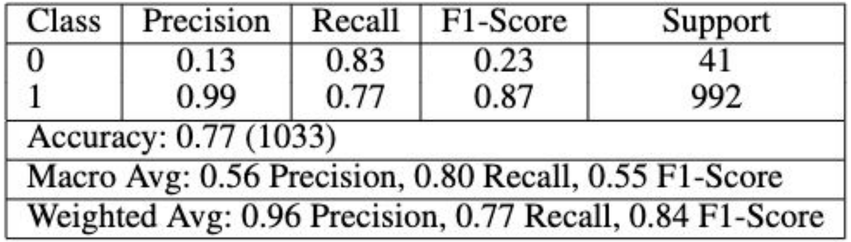
$$\begin{aligned} &\underset{\mathbf{w},\ b,\ \zeta}{\text{minimize}} && \|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n}\zeta_i \\ &\text{subject to} && y_i(\mathbf{w}^\top\mathbf{x}_i - b) \geq 1 - \zeta_i, \quad \zeta_i \geq 0 \quad \forall i \in \{1, \dots, n\} \end{aligned}$$

Parameters for Support Vector Classifier

- The hyperparameter **kernel** is set to **linear** given our assumption that the data can be linearly separated, a situation where SVM excels.
- The **class_weight=’balanced’** parameter is used to address class imbalance during model training.
- The **probability** setting is enabled, which allows us to generate the ROC Curve and the ROC-AUC Score.

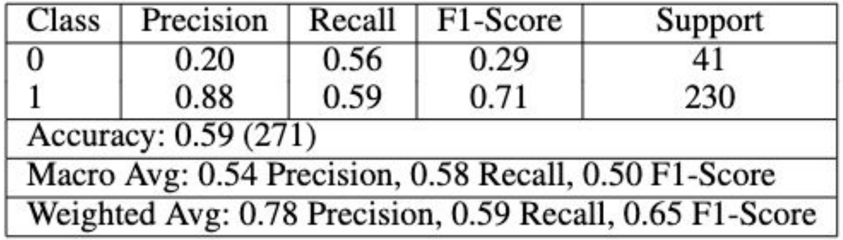
Result of SVM

Table 4: Healthy vs Cancer



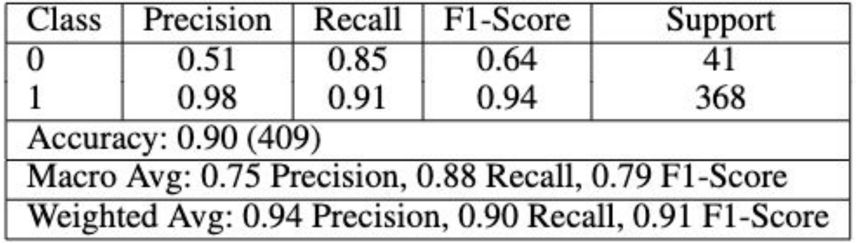
Class	Precision	Recall	F1-Score	Support
0	0.13	0.83	0.23	41
1	0.99	0.77	0.87	992
Accuracy: 0.77 (1033)				
Macro Avg: 0.56 Precision, 0.80 Recall, 0.55 F1-Score				
Weighted Avg: 0.96 Precision, 0.77 Recall, 0.84 F1-Score				

Table 5: Healthy vs Screening Stage Cancer

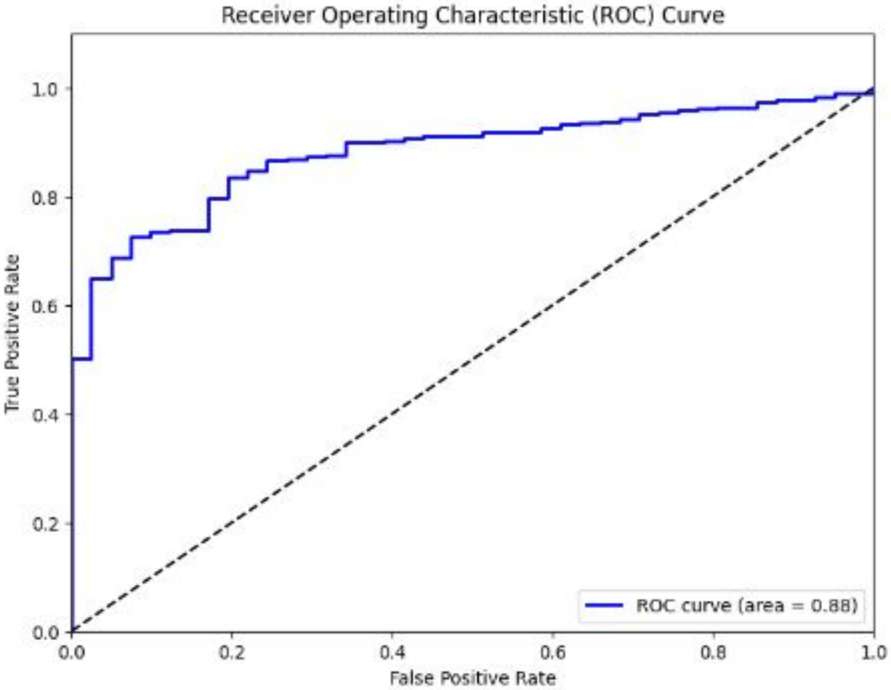


Class	Precision	Recall	F1-Score	Support
0	0.20	0.56	0.29	41
1	0.88	0.59	0.71	230
Accuracy: 0.59 (271)				
Macro Avg: 0.54 Precision, 0.58 Recall, 0.50 F1-Score				
Weighted Avg: 0.78 Precision, 0.59 Recall, 0.65 F1-Score				

Table 6: Healthy vs Early Stage Cancer



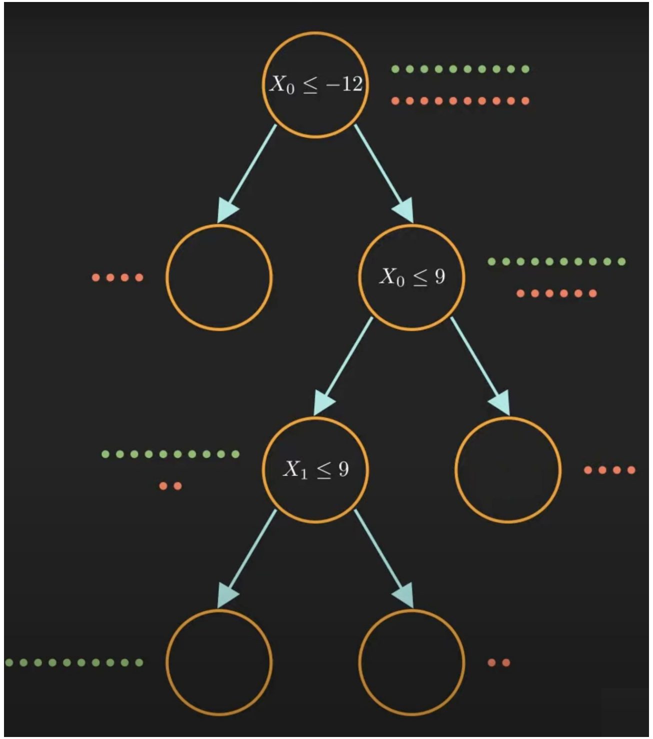
Class	Precision	Recall	F1-Score	Support
0	0.51	0.85	0.64	41
1	0.98	0.91	0.94	368
Accuracy: 0.90 (409)				
Macro Avg: 0.75 Precision, 0.88 Recall, 0.79 F1-Score				
Weighted Avg: 0.94 Precision, 0.90 Recall, 0.91 F1-Score				



Decision Tree

- Aims at finding a series of conditions that maximally separates the data into classes.

- Tree data structure
  - Decision node
  - Leaf node (end node)
  - Purity



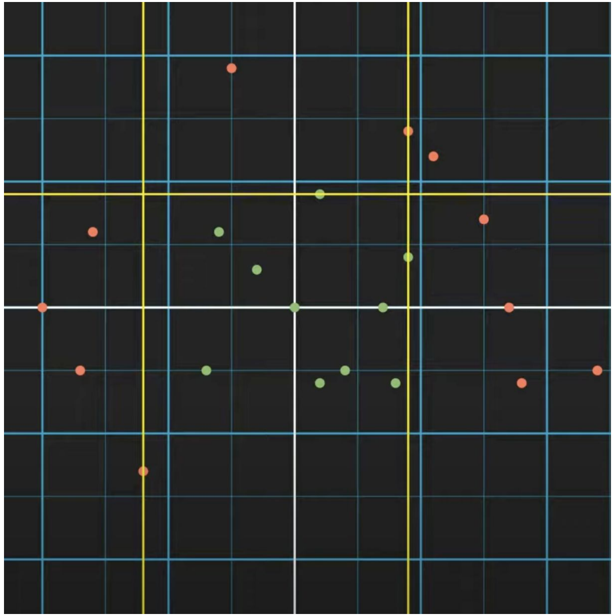


Image credit: (*"Decision Tree Classification Clearly Explained!"*, by Normalized Nerd, YouTube)

Parameters

- Classification criterion: **Gini impurity**

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk})$$

The function that outputs the *impurity* of node **m** given the set of observations $\mathbf{Q_m}$.

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k)$$

Proportion of class **k** observations in mode **m**.

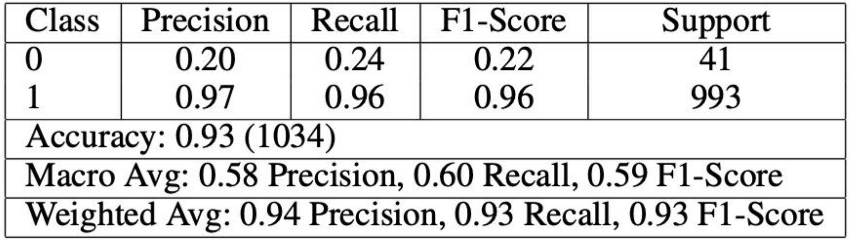
- $\mathbf{n_m}$ is the total number of observations, and **I(y=k)** is the indicator function.
- I(y=k) = 1 if y=k, and I(y=k) = 0 otherwise.

- Class_weight: **balanced**
- Max_depth: **None** (guarantees all leaf nodes are pure)
- Random_state: **0** (guarantees deterministic outputs)

Reference: SciKit Learn Documentation. 1.10.7.1
Classification Criteria (Decision Tree)

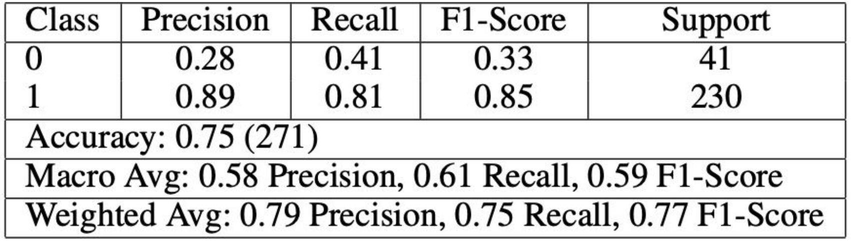
Results

Table 7: Healthy vs Cancer (Decision Tree)



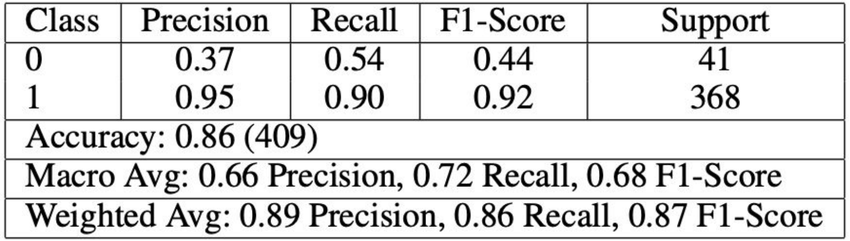
Class	Precision	Recall	F1-Score	Support
0	0.20	0.24	0.22	41
1	0.97	0.96	0.96	993
Accuracy: 0.93 (1034)				
Macro Avg: 0.58 Precision, 0.60 Recall, 0.59 F1-Score				
Weighted Avg: 0.94 Precision, 0.93 Recall, 0.93 F1-Score				

Table 8: Healthy vs Screening Stage Cancer (Decision Tree)

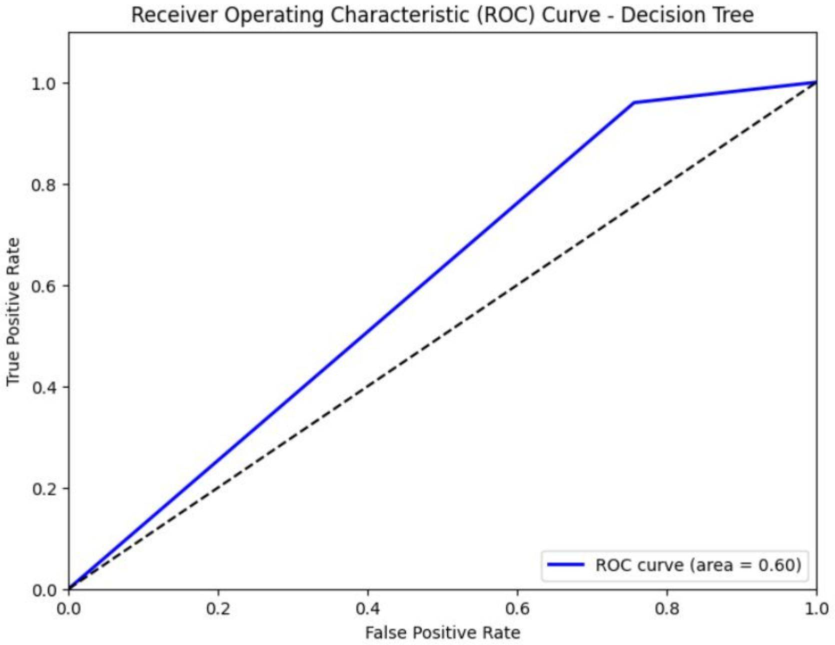


Class	Precision	Recall	F1-Score	Support
0	0.28	0.41	0.33	41
1	0.89	0.81	0.85	230
Accuracy: 0.75 (271)				
Macro Avg: 0.58 Precision, 0.61 Recall, 0.59 F1-Score				
Weighted Avg: 0.79 Precision, 0.75 Recall, 0.77 F1-Score				

Table 9: Healthy vs Early Stage Cancer (Decision Tree)



Class	Precision	Recall	F1-Score	Support
0	0.37	0.54	0.44	41
1	0.95	0.90	0.92	368
Accuracy: 0.86 (409)				
Macro Avg: 0.66 Precision, 0.72 Recall, 0.68 F1-Score				
Weighted Avg: 0.89 Precision, 0.86 Recall, 0.87 F1-Score				



Evaluation

- Decent accuracy in all three classification tasks. May be attributed to the fact that all end nodes are pure.

- Low class 0 recall for Healthy vs Cancer classification, suggests that the predicting capability of the model is highly skewed towards class 1.

- Minimal AUC under the ROC curve (close to 0.5) suggests that the model cannot satisfactorily differentiate healthy and cancerous data entries.

Artificial Neural Network (ANN)

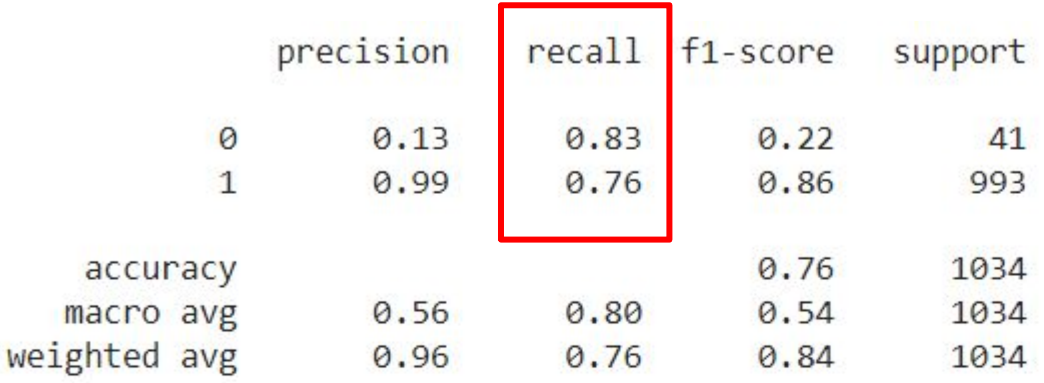
- Imbalance dataset
  - Over-sampling - results in overfitting
  - Under-sampling - reduces training dataset
  - Synthetic minority over-sampling (SMOTE)
- Feature scaling
  - Standard scaler
  - Min-max scaler
- Optimizer
  - Adaptive Moment Estimation (ADAM) - faster and more efficient than simpler optimizer SGD
- Loss function
  - Binary Cross Entropy - preferred loss function for binary classification

https://medium.com/game-of-bits/how-to-deal-with-imbalanced-data-in-classification-bd03cfc66066

https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/

Imbalance dataset

With SMOTE:

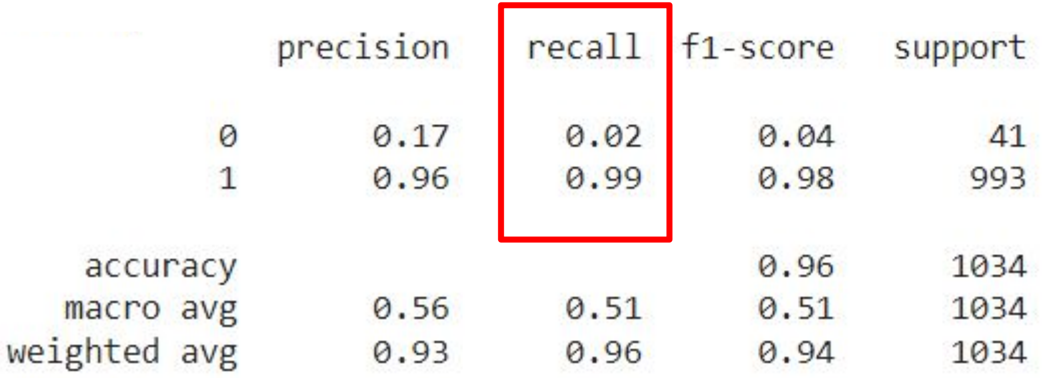


```
              precision    recall  f1-score   support

           0       0.13      0.83      0.22        41
           1       0.99      0.76      0.86       993

    accuracy                           0.76      1034
   macro avg       0.56      0.80      0.54      1034
weighted avg       0.96      0.76      0.84      1034
```

Without SMOTE:



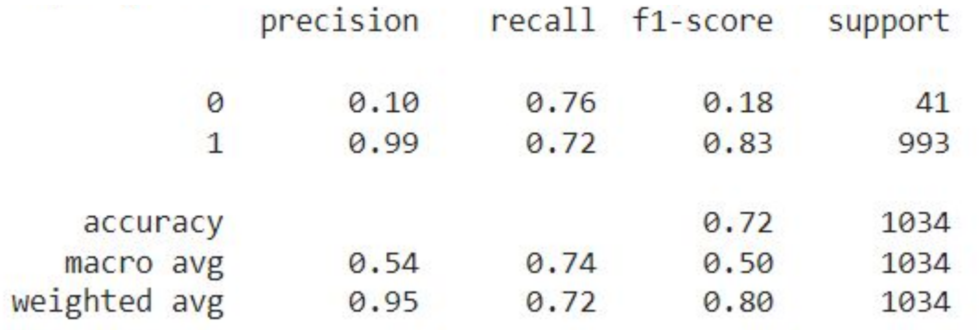
```
              precision    recall  f1-score   support

           0       0.17      0.02      0.04        41
           1       0.96      0.99      0.98       993

    accuracy                           0.96      1034
   macro avg       0.56      0.51      0.51      1034
weighted avg       0.93      0.96      0.94      1034
```

Feature scaling

Standard scaler:

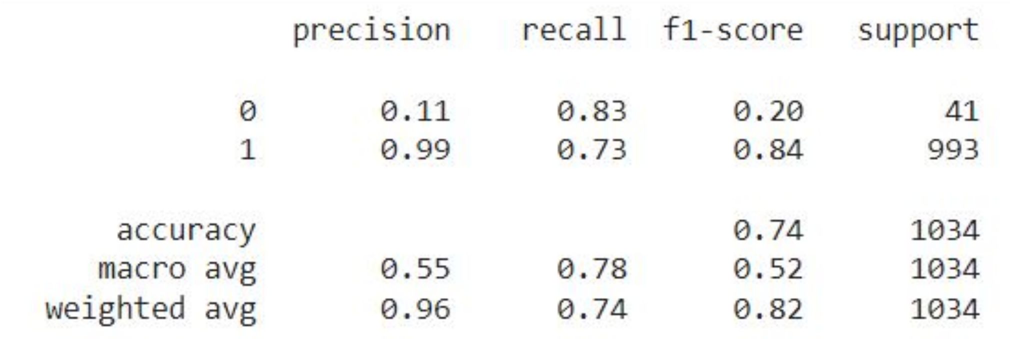


```
              precision    recall  f1-score   support

           0       0.10      0.76      0.18        41
           1       0.99      0.72      0.83       993

    accuracy                           0.72      1034
   macro avg       0.54      0.74      0.50      1034
weighted avg       0.95      0.72      0.80      1034
```

Min-max scaler:



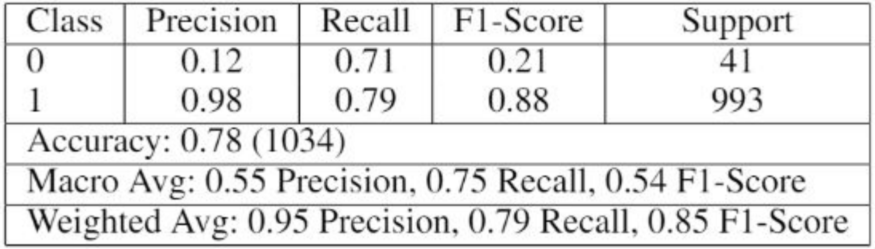
```
              precision    recall  f1-score   support

           0       0.11      0.83      0.20        41
           1       0.99      0.73      0.84       993

    accuracy                           0.74      1034
   macro avg       0.55      0.78      0.52      1034
weighted avg       0.96      0.74      0.82      1034
```

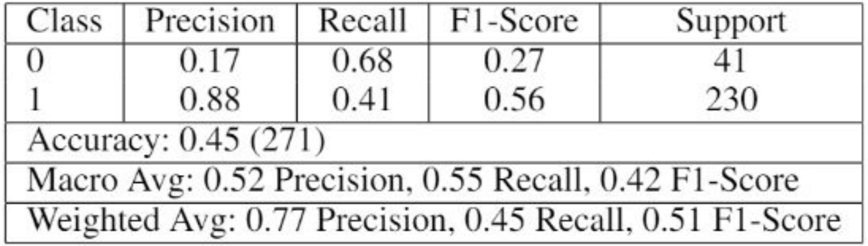
Results

Table 10: Healthy vs Cancer



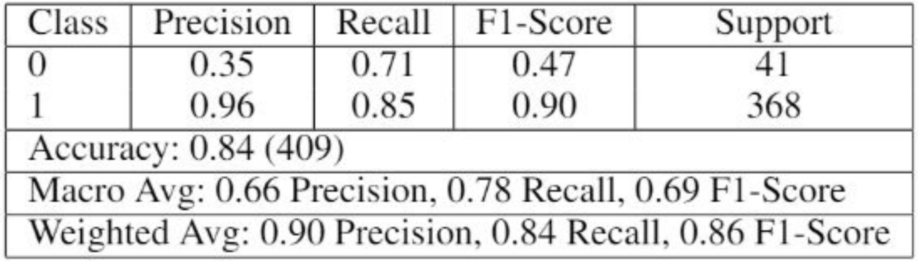
Class	Precision	Recall	F1-Score	Support
0	0.12	0.71	0.21	41
1	0.98	0.79	0.88	993
Accuracy: 0.78 (1034)				
Macro Avg: 0.55 Precision, 0.75 Recall, 0.54 F1-Score				
Weighted Avg: 0.95 Precision, 0.79 Recall, 0.85 F1-Score				

Table 11: Healthy vs Screening Stage Cancer

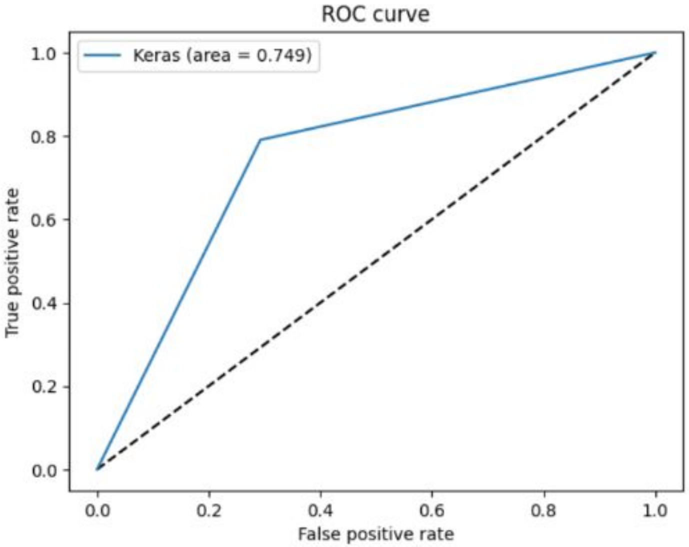


Class	Precision	Recall	F1-Score	Support
0	0.17	0.68	0.27	41
1	0.88	0.41	0.56	230
Accuracy: 0.45 (271)				
Macro Avg: 0.52 Precision, 0.55 Recall, 0.42 F1-Score				
Weighted Avg: 0.77 Precision, 0.45 Recall, 0.51 F1-Score				

Table 12: Healthy vs Early Stage Cancer



Class	Precision	Recall	F1-Score	Support
0	0.35	0.71	0.47	41
1	0.96	0.85	0.90	368
Accuracy: 0.84 (409)				
Macro Avg: 0.66 Precision, 0.78 Recall, 0.69 F1-Score				
Weighted Avg: 0.90 Precision, 0.84 Recall, 0.86 F1-Score				



Conclusion & Discussions

- Final Decision:
  - Decision Tree
- Future Works:
  - Since the dataset has 51 independent variables (regressors), Unsupervised Learning techniques such as Dimension Reduction or Principal Component Analysis (PCA) can be performed to ensure a better fit of the models.

Contribution Percentages

Jin Jiarui: 25%

Lai Zheyuan: 25%

Tan Hui En: 25%

Zhuang Yiling: 25%