# ITD105 Case Study #1

**Comparing Machine Learning Algorithms**

Name: Allan Raymart C. Paraiso

Video Link: https://drive.google.com/file/d/1h-D\_u2XaEjVgdSOuNNI10\_vDoCwXeg7-/view?usp=sharing

# CLASSIFICATION

Train the **classification dataset** using various machine learning algorithms designed for classification. Evaluate and compare these models by applying different resampling techniques and utilizing appropriate performance metrics.

# Classification Dataset

**Dataset** Name : Cirrhosis Patient Survival Prediction

**Features**: ID, N\_Days, Status, Age, Bilirubin, Cholesterol, Albumin, Copper, Alk\_Phos, SGOT, Tryglicerides, Platelets, Prothrombin, Stage

# Set A

Resampling Technique : Split into Train and Test Sets

Classification Metric : Confusion Matrix and Classification Report

|  |  |
| --- | --- |
| **ML Algorithm (Classification)** | **Confusion Matrix**  ***(Provide the matrix and classification report of each algorithm)*** |
| CART (Classification and Regression Trees) |  |
| Gaussian Naive Bayes/Naive Bayes |  |
| Gradient Boosting Machines (AdaBoost) |  |
| K-Nearest Neighbors (K-NN) |  |
| Logistic Regression |  |
| Multi-Layer Perceptron (MLP) |  |
| Perceptron |  |
| Random Forest |  |
| Support Vector Machines (SVM) |  |

**Set B** *(should use different resampling technique and classification metric)*

Resampling Technique: Repeated Random Train-Test splits

Classification Metric: Logarithmic Loss

|  |  |
| --- | --- |
| **ML Algorithm (Classification)** |  |
| CART (Classification and Regression Trees) | Mean Logarithmic Loss: 2.995 |
| Gaussian Naive Bayes/Naive Bayes | Mean Logarithmic Loss: 0.703 |
| Gradient Boosting Machines (AdaBoost) | Mean Logarithmic Loss: 0.858 |
| K-Nearest Neighbors (K-NN) | Mean Logarithmic Loss: 3.497 |
| Logistic Regression | Mean Logarithmic Loss: 0.563 |
| Multi-Layer Perceptron (MLP) | Mean Logarithmic Loss: 7.086 |
| Perceptron | Mean Logarithmic Loss: 0.700 |
| Random Forest | Mean Logarithmic Loss: 0.63732 |
| Support Vector Machines (SVM) | Mean Logarithmic Loss: 0.56745 |

**Set C** *(should use different resampling technique and classification metric)*

Resampling Technique: K-fold Cross Validation

Classification Metric: Classification Report

|  |  |
| --- | --- |
| **ML Algorithm (Classification)** | Average Precision |
| CART (Classification and Regression Trees) | 0.761 |
| Gaussian Naive Bayes/Naive Bayes | 0.710 |
| Gradient Boosting Machines (AdaBoost) | 0.759 |
| K-Nearest Neighbors (K-NN) | 0.687 |
| Logistic Regression | 0.772 |
| Multi-Layer Perceptron (MLP) | 0.634 |
| Perceptron | 0.673 |
| Random Forest | 0.759 |
| Support Vector Machines (SVM) | 0.646 |

**Results interpretation**

Overall, the results suggest that Logistic Regression, SVM, and Random Forest are the best machine learning algorithms for the task of binary classification for the reason that Logistic Regression has the lowest mean logarithmic loss, followed by Support Vector Machines (SVM) and Random Forest. Logistic Regression also has the highest average precision, followed by CART (Classification and Regression Trees) and Gradient Boosting Machines (AdaBoost)..

**Based on the results, perform algorithm/hyperparameter tuning (at least 3) of the chosen ML algorithm.**

**ML Algorithm:** Gaussian Naïve Bayes

**Sampling Technique -** Train/Test Split (80:20) **Classification Metrics –** Accuracy

**ML Algorithm:** Gradient Boosting Adaboost

**Sampling Technique -** Train/Test Split (80:20) **Classification Metrics –** Accuracy

**ML Algorithm:** K-Nearest Neighbors

**Sampling Technique -** Train/Test Split (80:20) **Classification Metrics –** Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SVM Hyperparameters** | | | |
|  | **random\_state** | **Precision Weighted avg** | **Precision macro avg** | **Accuracy** |
| Gaussian Naïve Bayes | 7 | 0.41 | 0.60 | 77.810% |
| Adaboost | 7 | 0.57 | 0.83 | 83.333% |
| KNN | 7 | 0.48 | 0.71 | 71.428% |

# Results interpretation:

Based on the result the adaboost has the most accuracy given the precision weighted average, macro average and the overall accuracy.

1. **REGRESSION**

Train the **regression dataset** using various machine learning algorithms designed for regression. Evaluate and compare these models by applying different resampling techniques and utilizing appropriate performance metrics.

# Regression Dataset

Dataset Name : insurance costs

Features: age, sex,bmi,chidren,smoker,region

# Set A

Resampling Technique : train test split 80:20

Regression Metric : Mean Absolute Error

|  |  |
| --- | --- |
| **ML Algorithm (Regression)** |  |
| CART (Classification and Regression Trees) | 2651.093 |
| Elastic Net | 7423.916 |
| Gradient Boosting Machines (AdaBoost) | 3922.616 |
| K-Nearest Neighbors (K-NN) | 7872.695 |
| Lasso Regression | 3972.271 |
| Ridge Regression | 3984.913 |
| Linear Regression | 3971.629 |
| Multi-Layer Perceptron (MLP) | 3970.469 |
| Random Forest | 2575.167 |

**Set B** *(should use different resampling technique and regression metric)*

Resampling Technique: K Fold

Regression Metric: Mean Squared Error

|  |  |
| --- | --- |
| **ML Algorithm (Regression)** |  |
| CART (Classification and Regression Trees) | 44243399.200 |
| Elastic Net | 89789075.134 |
| Gradient Boosting Machines (AdaBoost) | 25196255.396 |
| K-Nearest Neighbors (K-NN) | 121422424.326 |
| Lasso Regression | 37003521.552 |
| Ridge Regression | 37005349.740 |
| Linear Regression | 37004496.989 |
| Multi-Layer Perceptron (MLP) | 123438921.241 |
| Random Forest | 25494408.001 |

# Results interpretation (Set A and Set B):

Given the results in set a and set b, it is very evident that the dataset is seen as inaccurate given the large number of mean absolute errors and mean squared errors, but if for some reason the dataset is still used, the best choice for machine learning algorithm is random forest given that it has the lowest error of all the models tested.

**Based on the results, perform at algorithm tuning (at least 3) of the chosen ML algorithm.**

**ML Algorithm:** Random Forest

**Sampling Technique -** Train/Test Split (80:20) **Regression Metrics –** MAE

**ML Algorithm:** Support Vector Machines

**Sampling Technique –** K fold

**Regression Metrics –** MAE

**ML Algorithm:** Random Forest

**Sampling Technique –** K-FOLD

**Regression Metrics –** MAE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SVM Hyperparameters** | | | |
|  | **epsilon** | **Kernel** | **C** | **MAE** |
| Model I | 0.1 | linear | 1.0 | 6912.019 |
| Model II | 0.2 | linear | 1.25 | 6554.593 |
| Model III | 0.2 | poly | 1.0 | 7655.444 |

# Results interpretation:

# Based on the results it can be concluded that the dataset is faulty given the high numbers of mean absolute errors, across all tested models, and therefore deemed inaccurate